Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **00**, 1–27
https://doi.org/10.1093/jrsssb/qkaf058

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Original Article

Bayesian analysis of product feature allocation models

Lorenzo Ghilotti, Federico Camerlenghi and Tommaso Rigon

Department of Economics, Management, and Statistics, University of Milano-Bicocca, Milano 20126, Italy

Address for correspondence: Federico Camerlenghi, Department of Economics, Management, and Statistics, University of Milano-Bicocca, Milano 20126, Italy. Email: federico.camerlenghi@unimib.it

Abstract

Feature allocation models are an extension of Bayesian nonparametric clustering models, where individuals can share multiple features. We study a broad class of models whose probability distribution has a product form, which includes the popular Indian buffet process. This class plays a prominent role among existing priors, and it shares structural characteristics with Gibbs-type priors in the species sampling framework. We develop a general theory for the entire class, obtaining closed form expressions for the predictive structure and the posterior law of the underlying stochastic process. Additionally, we describe the distribution for the number of features and the number of hitherto unseen features in a future sample, leading to the α-diversity for feature models. We also examine notable novel examples, such as mixtures of Indian buffet processes and beta Bernoulli models, where the latter entails a finite random number of features. This methodology finds significant applications in ecology, allowing the estimation of species richness for incidence data, as we demonstrate by analyzing plant diversity in Danish forests and trees in Barro Colorado Island.

Keywords: Bayesian nonparametrics, completely random measures, exchangeable feature probability function, Gibbs-type feature models, Indian buffet process

1 Introduction

Random feature allocations have emerged as an important area of Bayesian nonparametrics. The pioneering work of Griffiths and Ghahramani (2006) introduced the Indian buffet process (IBP), a stochastic mechanism for binary matrices, which is obtained by considering the infinite limit of a beta Bernoulli (BB) model. Unlike clustering problems, where each individual belongs to a single group, in feature allocation models every observation may possess a finite set of *features* or *attributes*. Shortly after its proposal, Thibaux and Jordan (2007) demonstrated that de Finetti's celebrated theorem could be applied to the IBP, establishing its connection to the beta process of Hjort (1990). This pivotal finding linked IBPs to the theory of completely random measures (Kingman, 1967), laying the groundwork for a new branch of Bayesian nonparametrics. These initial investigations sparked a rich stream of research, particularly within the machine learning community. IBP models have found widespread applicability across various domains, including Bayesian factor analysis and non-negative matrix factorization (Ayed & Caron, 2021; Griffiths & Ghahramani, 2006; Knowles & Ghahramani, 2011), topic modeling (Williamson et al., 2010), relational models (Miller et al., 2009; Palla et al., 2012), and object recognition (Broderick et al., 2015). We refer to Teh and Jordan (2010) and Griffiths and Ghahramani (2011) for a comprehensive review of the early contributions. The relevance of IBPs as a statistical tool is now firmly established; however, their limitations have also become apparent, as well as the need for a deeper theoretical understanding. For example, the logarithmic growth rate of the number of features in the two-parameter IBP (Griffiths & Ghahramani, 2011) spurred the proposal of a

Received: August 28, 2024. **Revised:** June 4, 2025. **Accepted:** August 24, 2025

three-parameter generalization by Teh and Gorur (2009), which exhibits a power-law behaviour; see also Broderick et al. (2012). Another major step was pursued by Broderick et al. (2013), who showed that most feature models are characterized by a combinatorial entity known as the *exchangeable feature probability function* (EFPF). More recently, James (2017) studied a general class of feature models based on completely random measures, while another extensive class, relying on random scaling of the underlying process, is investigated in Camerlenghi et al. (2024). Another recent construction is discussed in Heaukulani and Roy (2020).

In this paper, we study the broad class of feature models defined by Battiston et al. (2018), who holds the merit of characterizing all EFPFs with a product form as mixtures of the two most widely used feature models, namely the Indian buffet process (IBP, Teh & Gorur, 2009) and the beta Bernoulli (BB, Griffiths & Ghahramani, 2006). However, apart from this important representation theorem, a comprehensive statistical investigation is still lacking. We develop a general theory for this class of models, encompassing (i) the predictive structure, leading to a generalized Indian buffet metaphor, (ii) the posterior distribution of the underlying process, (iii) prior and posterior properties regarding the number of features, and (iv) the asymptotic behaviour. Our findings are available in closed form, lead to computationally efficient inferential procedures, and enjoy a transparent interpretation. Moreover, this theoretical investigation allows us to identify three novel feature allocation models that stand out for their tractability: the gamma mixture of IBPs, and the Poisson and negative binomial mixture of BBs. The latter two models entail a random but finite number of possible features, therefore being structurally different from existing IBP-like specifications that involve infinitely many features. Finally, we highlight several remarkable parallelisms between the class of Battiston et al. (2018) and Gibbs-type priors for species sampling models (De Blasi et al., 2015; Gnedin & Pitman, 2005). In light of these similarities, we will refer to this class as *Gibbs-type feature models*. In species sampling problems, Gibbs-type priors are perhaps the most natural generalization of the Dirichlet process of Ferguson (1973), owing to their balance between flexibility and analytical tractability. Notable examples are the Pitman–Yor process (Pitman & Yor, 1997), the normalized generalized gamma process (Lijoi et al., 2007b), and mixtures of Dirichlet multinomial processes (De Blasi et al., 2013; Gnedin, 2010). For similar reasons, we argue that Gibbs-type feature models are one of the most natural extensions of the IBP and the BB.

We demonstrate here the usefulness of feature models in ecological problems as a tool to measure biodiversity. There exists a rich literature about the quantification of biodiversity (Colwell, 2009; Magurran & McGill, 2011) with taxon richness, i.e. the number of different taxa present in a community, being perhaps the simplest and most natural definition. Richness estimation is, in turn, related to the notion of *taxon accumulation curves* (Gotelli & Colwell, 2001). A Bayesian nonparametric inferential framework for predicting unseen species has been laid down by Lijoi et al. (2007a) for Gibbs-type priors. See also Favaro et al. (2009) for the Pitman–Yor special case and Zito et al. (2024) for a related model-based approach. These Bayesian methods are suitable for individual-based accumulation curves, that is, when species are observed one at a time. However, species are often captured or collected in chunks, and hence, each observation takes the form of a vector of binary variables accounting for the presence or absence of a species. Feature models are well-suited for this kind of data, called *incidence data*, leading to a Bayesian analysis of sample-based accumulation curves. Despite the development of classical estimators for this setting (e.g. Chakraborty et al., 2019; Chao et al., 2014; Chiu, 2022, 2023; Chiu et al., 2014; Colwell et al., 2012), the Bayesian nonparametric literature remains much more limited, except for the recent works of Masoero et al. (2022) and Camerlenghi et al. (2024). Our theoretical investigation allows for the prediction of the number of unseen species, the modeling of accumulation curves, and the quantification of biodiversity. For instance, an important theoretical result of this paper, particularly relevant for ecological applications, is the definition of the α-diversity, a biodiversity measure that extends the notion of Pitman (2003) to sample-based designs. In the proposed Poisson and negative binomial mixture of BB models, the α-diversity coincides with the taxon richness, and its posterior distribution follows a Poisson and a negative binomial distribution, respectively. This leads to straightforward Bayesian estimators for the taxon richness whose uncertainty can be formally and easily quantified. Although this work focuses on ecological applications, the proposed methodology is broadly applicable across various domains. For instance, in biological sciences, estimating the number of unseen or rare genetic

variants in the human genome can help the understanding of human diseases or guide the design of effective clinical procedure (Gravel, 2014; Ionita-Laza et al., 2009; Zou et al., 2016). In single-cell sequencing data, predicting the number and frequency of somatic mutations at the cellular level is essential for characterizing tumor heterogeneity, which is a key factor in cancer progression and resistance to therapy. Since the expense of sequencing is nontrivial, accurate prediction is crucial to allocate limited sequencing budget (Zhang et al., 2020). Other applications include cancer biology (Chakraborty et al., 2019), precision medicine (Momozawa & Mizukami, 2021) and microbiome analysis (Sanders et al., 2019).

The paper is structured as follows. In Section 2, we review feature allocation models. In Section 3, we develop general theory for the class of Gibbs-type feature models. In Sections 4 and 5, we propose and study novel examples of Gibbs-type feature allocation models, distinguishing between models with an infinite number of features (mixtures of IBPS) and those assuming finitely many features (mixtures of BBS). Simulation studies are discussed in Section 6, while Section 7 illustrates our methodology by analyzing two real datasets. The paper ends with a discussion; proofs, additional theorems, simulation studies and additional details about the applications are collected in the supplementary material.

2 Feature allocation models

2.1 Preliminary concepts

Feature allocation models describe how features are distributed among a sample of n individuals (Broderick et al., 2013). Let $[n] = \{1, \ldots, n\}$ be a set comprising the first n natural numbers. An ordered *feature allocation* is a sequence of non-empty sets $B_{n,\ell} \subseteq [n]$, for $\ell = 1, \ldots, K_n$, such that $B_{n,\ell}$ identifies the set of individuals exhibiting the ℓth feature. To distinguish these sets, we assign them a *label*. More precisely, suppose $\mathbb{X}$ is the space of feature labels, and let $X_\ell \in \mathbb{X}$ be the label associated with the set $B_{n,\ell}$, for $\ell = 1, \ldots, K_n$. The labels X_ℓ are independent and identically distributed (i.i.d.) draws, that is $X_\ell \overset{\text{iid}}{\sim} P_0$, with P_0 being a diffuse distribution on $\mathbb{X}$, ensuring that the labels are almost surely (a.s.) distinct. The association between feature allocations and labels can be encoded using binary random variables $A_{i,\ell}$ for $i = 1, \ldots, n$ and $\ell = 1, \ldots, K_n$, so that $A_{i,\ell}$ equals 1 if the ith individual displays feature X_ℓ and 0 otherwise. In other terms, the random set $B_{n,\ell}$ may be written as $B_{n,\ell} = \{i \in [n] : A_{i,\ell} = 1\}$. A feature allocation can be represented through binary matrices, as depicted in Figure 1, where there are $n = 10$ individuals and $K_n = 18$ features. Each column of the binary matrix corresponds to a set $B_{n,\ell}$, with the ith element of the ℓth column representing the value $A_{i,\ell}$. We assume that each individual may exhibit only a finite number of features; that is, an individual belongs to a finite number of $B_{n,\ell}$'s. This implies that the total number of observed features K_n is a.s. finite for any $n \geq 1$.

A *feature allocation model* is a probability distribution for a random ordered feature allocation $F_n = (B_{n,1}, \ldots, B_{n,K_n})$. Alternatively, one may consider the probability distribution for an *unordered feature allocation* $\tilde{F}_n = \{(\tilde{B}_{n,1}, \tilde{K}_{n,1}), \ldots, (\tilde{B}_{n,H_n}, \tilde{K}_{n,H_n})\}$, where $\tilde{B}_{n,h} \subseteq [n]$ are the $H_n \leq K_n$ distinct sets among the $B_{n,1}, \ldots, B_{n,K_n}$, with $\tilde{K}_{n,h}$ being the number of sets equal to $\tilde{B}_{n,h}$. The unordered $\tilde{F}_n$ is sometimes used to define feature allocation models (see Broderick et al., 2013), but for our purposes, it is more convenient to deal with the ordered F_n. In any event, we assume that the probability of $\tilde{F}_n$ being equal to any unordered feature allocation $\tilde{f}_n$ is equally split between the $K_n!/\prod_{h=1}^{H_n} \tilde{K}_{n,h}!$ possible orders of the sets $B_{n,1}, \ldots, B_{n,K_n}$. Hence, the following relationship holds

$$\mathbb{P}(\tilde{F}_n = \tilde{f}_n) = \frac{K_n!}{\prod_{h=1}^{H_n} \tilde{K}_{n,h}!} \mathbb{P}(F_n = f_n),$$

where f_n is one of the possible orderings of $\tilde{f}_n$.

In this paper, we focus on feature allocation models admitting an *exchangeable feature probability function* (EFPF). This is a probabilistic object whose role is analogous to the exchangeable partition probability function (EPPF) in the species sampling framework (Pitman, 1996), as carefully discussed in Broderick et al. (2013) . Specifically, let $(M_{n,1}, \ldots, M_{n,K_n})$ be the random vector of

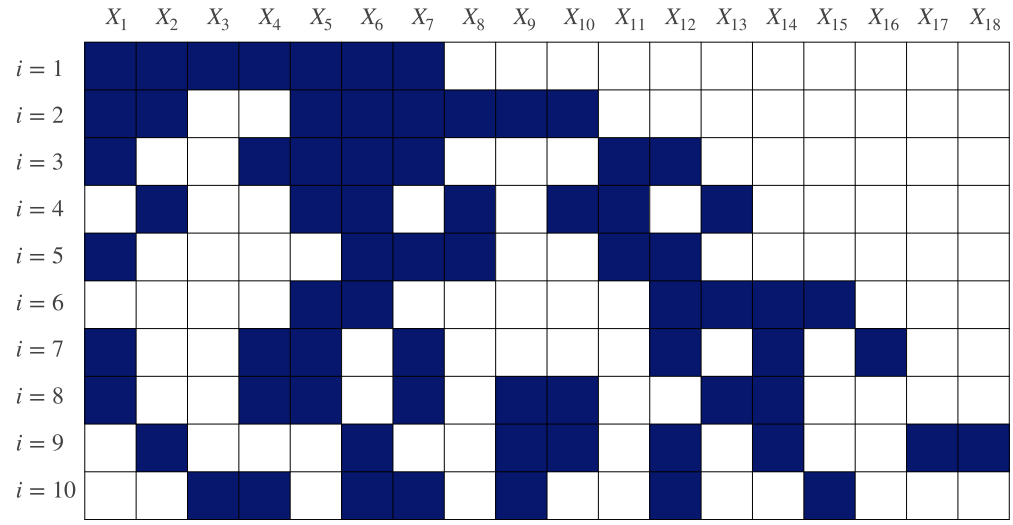

Figure 1. Matrix representation of a feature allocation, with $n = 10$ individuals and $K_n = 18$ observed features. Features are in *order of appearance* (Broderick et al., 2013). The blue squares correspond to $A_{i,\ell} = 1$ (the ith individual displays feature X_ℓ), the white squares correspond to $A_{i,\ell} = 0$ (the ith does not express feature X_ℓ).

feature frequencies, where $M_{n,\ell} = \#B_{n,\ell} = \sum_{i=1}^{n} A_{i,\ell}$, for $\ell = 1, \ldots, K_n$. We assume that the probability distribution of F_n depends on the sample solely through the vector $\boldsymbol{M}^{(n)} = (M_{n,1}, \ldots, M_{n,K_n})$, namely

$$\mathbb{P}(F_n = f_n) = \pi_n(m_1, \ldots, m_k),$$

for every f_n and for every $n \geq 1$, where π_n is a $[0, 1]$-valued symmetric function defined on $\bigcup_{k\geq 0} [n]^k$, and $(m_1, \ldots, m_k)$ are the feature frequencies for f_n. The function π_n is termed exchangeable feature probability function, and it encapsulates all the relevant properties of the model. By construction, feature allocation models admitting an EFPF are exchangeable, meaning that the distribution of F_n is invariant under any permutation of the indexes of the n individuals. It is also natural to require feature models to be *Kolmogorov consistent*, which means that the probability distribution of the feature allocation for n individuals coincides with that for $n + 1$ individuals once the last individual is integrated out. We refer to Broderick et al. (2013) for a detailed discussion.

2.2 Exchangeable Gibbs-type feature allocation models

Among the exchangeable and consistent models, the class of EFPF in product form introduced by Battiston et al. (2018) represents a special subset that is still very rich and diversified. We refer to this class as *exchangeable Gibbs-type feature allocation models*, or *Gibbs-type feature models* for brevity, for the evident similarity with exchangeable Gibbs-type random partitions (Gnedin & Pitman, 2005). We consider EFPFs of the following product form $\pi_n(m_1, \ldots, m_k) = V_{n,k} \prod_{\ell=1}^{k} W_{m_\ell} U_{n-m_\ell}$, where $V = (V_{n,k} : (n, k) \in \mathbb{N} \times \mathbb{N}_0)$ and $W = (W_j : j \in \mathbb{N})$, $U = (U_j : j \in \mathbb{N}_0)$ are two sequences of non-negative weights, with $\mathbb{N}$ denoting the set of natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Apart from some limiting cases, an important result of Battiston et al. (2018) states that Gibbs-type feature models are necessarily of the form

$$\pi_n(m_1, \ldots, m_k) = V_{n,k} \prod_{\ell=1}^{k} (1 - \alpha)_{m_\ell - 1} (\theta + \alpha)_{n - m_\ell}, \tag{1}$$

for $-\infty < \alpha < 1$ and $-\alpha < \theta < \infty$, where $(x)_m = \Gamma(x + m)/\Gamma(x)$ is the Pochhammer symbol, and $\Gamma(x)$ is the gamma function. The array V must satisfy the following recurrence relationship

$V_{n,k} = \sum_{j=0}^{\infty} (k+j)!/\{j!k!\}\{(\theta+\alpha)_n\}^j(\theta+n)^k V_{n+1,k+j}$, which guarantees the consistency of the EFPF. The limiting case $\alpha = 1$ corresponds to no feature sharing, i.e. $M_{n,\ell} = 1$ almost surely for $\ell = 1, \ldots, K_n$, whereas $\theta = -\alpha$ corresponds to complete feature sharing, that is $M_{n,\ell} = n$ almost surely, for $\ell = 1, \ldots, K_n$. These degenerate situations are uninteresting in practice.

The most popular and widely used feature allocation models are of Gibbs-type. A first noteworthy example is the three-parameter Indian buffet process (IBP), introduced in Teh and Gorur (2009), with parameters (γ, α, θ) satisfying $\gamma > 0$, $0 \leq \alpha < 1$, and $\theta > -\alpha$. The EFPF is in product form (1) and the $V_{n,k}$'s are given by

$$V_{n,k} = \frac{1}{k!}\left\{\frac{\gamma}{(\theta+1)_{n-1}}\right\}^k \exp\{-\gamma g_n(\theta,\alpha)\}, \quad \text{with} \quad g_n(\theta,\alpha) := \sum_{i=1}^{n} \frac{(\theta+\alpha)_{i-1}}{(\theta+1)_{i-1}}. \tag{2}$$

The choice $\alpha = 0$ corresponds to the two-parameter IBP, while the one-parameter model is obtained by further considering $\theta = 1$; see Griffiths and Ghahramani (2011). We stress that the distribution of K_n, in the IBP case, has unbounded support. A second notable example is the beta Bernoulli (BB), with parameters (N, α, θ) such that $N \in \mathbb{N}$, $\alpha < 0$ and $\theta > -\alpha$ (Griffiths & Ghahramani, 2011). The EFPF of a BB is also in product form (1), with the $V_{n,k}$'s given by

$$V_{n,k} = \binom{N}{k}\left\{\frac{-\alpha}{(\theta+\alpha)_n}\right\}^k \left\{\frac{(\theta+\alpha)_n}{(\theta)_n}\right\}^N \mathbb{1}_{\{0,1,\ldots,N\}}(k), \tag{3}$$

where 1_C denotes the indicator function of a set C. The BB model prescribes that the observed number of features K_n is bounded by N. Remarkably, a characterization theorem due to Battiston et al. (2018) establishes that the IBP and the BB are the building blocks of any Gibbs-type feature model. More precisely, for fixed values of (θ, α), the set of Gibbs coefficients $V_{n,k}$ satisfying the aforementioned consistency condition are necessarily mixtures of the γ and N parameters of the IBP and the BB, respectively. This is better clarified in the following result.

Proposition 1 (Theorem 1.1 of Battiston et al., 2018). For fixed values of (θ, α) such that $\theta > -\alpha$, the set of solutions of the recursions for the $V_{n,k}$'s is:

(i) for $0 \leq \alpha < 1$, mixtures over $\gamma \in \mathbb{R}^+$ of the $V_{n,k}$'s of IBPs with respect to a distribution P_γ;
(ii) for $\alpha < 0$, mixtures over $N \in \mathbb{N}$ of the $V_{n,k}$'s of BBs with respect to a distribution P_N.

Hence, any Gibbs-type feature model is obtained by considering a prior distribution for the γ and N parameters of the IBP and BB models. This draws an elegant parallelism between Gibbs-type feature models and Gibbs-type partitions of Gnedin and Pitman (2005) since, in both cases, product form distributions are obtained as mixtures of a set of simple models. These analogies will be strengthened in Section 3, where we will show that the parameter α also controls the asymptotic growth rate for the number of distinct features K_n, as in species sampling models, leading to the analogous of the α-diversity of Pitman (2003) for feature allocations.

2.3 Hierarchical representations and random measures

Gibbs-type feature models admit a hierarchical representation in terms of Bernoulli processes (BPS) and random measures. For the two-parameter IBP, such a stochastic representation was established in the illuminating contribution of Thibaux and Jordan (2007). Let us consider the sequence of all possible feature labels $(\tilde{X}_j)_{j\geq 1}$, where $\tilde{X}_j \overset{\text{iid}}{\sim} P_0$ for any $j \geq 1$. The labels $X_1, \ldots, X_{K_n}$ observed in a sample of n individuals are a subset of the complete list of labels $(\tilde{X}_j)_{j\geq 1}$. The ith individual is characterized by its expressed features, i.e. by the pairs $((\tilde{X}_j, \tilde{A}_{i,j}))_{j\geq 1}$, where each $\tilde{A}_{i,j} = 1$ if the ith

individual exhibits feature $\tilde{X}_j$, and $\tilde{A}_{i,j} = 0$ otherwise. Thus, the pairs $((\tilde{X}_j, \tilde{A}_{i,j}))_{j\geq 1}$ may be organized through a counting measure Z_i on the space $\mathbb{X}$, which is given by

$$Z_i(\cdot) = \sum_{j\geq 1} \tilde{A}_{i,j}\delta_{\tilde{X}_j}(\cdot). \tag{4}$$

The feature allocation $F_n = (B_{n,1}, \ldots, B_{n,K_n})$ and the binary variables $A_{i,\ell}$ can be then expressed through the counting measures Z_i, since we have $B_{n,\ell} = \{i \in [n] : Z_i(\{X_\ell\}) = 1\}$ and $A_{i,\ell} = Z_i(\{X_\ell\})$.

In Gibbs-type feature models, the $\tilde{A}_{i,j}$'s are conditionally independent Bernoulli random variables given a sequence of random probabilities $(\tilde{q}_j)_{j\geq 1}$ such that $\sum_{j\geq 1}\tilde{q}_j < \infty$ almost surely. In other terms, we suppose that $\tilde{A}_{i,j} \mid \tilde{q}_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\tilde{q}_j)$. We organize these probabilities using another measure $\tilde{\mu}$ on $\mathbb{X}$, namely

$$\tilde{\mu}(\cdot) := \sum_{j\geq 1} \tilde{q}_j\delta_{\tilde{X}_j}(\cdot). \tag{5}$$

We will say that, conditionally on $\tilde{\mu}$, the Z_i's are i.i.d. Bernoulli processes with base measure $\tilde{\mu}$, written $Z_i \mid \tilde{\mu} \overset{\text{iid}}{\sim} BP(\tilde{\mu})$ for any $i \geq 1$. In other words, the infinite sequence of random measures $(Z_i)_{i\geq 1}$ is exchangeable. Summarizing, the following hierarchical representation holds:

$$\begin{aligned} Z_i \mid \tilde{\mu} &\overset{\text{iid}}{\sim} BP(\tilde{\mu}), \quad i \geq 1, \\ \tilde{\mu} &\sim Q, \end{aligned} \tag{6}$$

where Q is the *prior* distribution of $\tilde{\mu}$, i.e. the *de Finetti measure*. The hierarchical generative scheme outlined in equations (4)–(6) is termed a *feature frequency model*, and it leads to a consistent and exchangeable EFFF (Broderick et al., 2013). Gibbs-type feature models always admit such a hierarchical construction under specific prior distributions for the random measure $\tilde{\mu}$. This is a well-known fact for IBP and BB models, whose random measures are denoted by $\tilde{\mu} \mid \gamma$ and $\tilde{\mu} \mid N$. Hence, thanks to Proposition 1, the law of $\tilde{\mu}$ for any Gibbs-type feature model is a mixture over γ or N of the corresponding law for the IBP or BB model.

Let us first consider the BB model with parameters (N, α, θ), in which there are N possible features $\tilde{X}_1, \ldots, \tilde{X}_N$. The hierarchical representation of the beta Bernoulli process is straightforward as we have $\tilde{\mu} \mid N = \sum_{j=1}^N \tilde{q}_j\delta_{\tilde{X}_j}$, with $\tilde{q}_j \overset{\text{iid}}{\sim} \text{Beta}(-\alpha, \theta+\alpha)$ and $\tilde{X}_j \overset{\text{iid}}{\sim} P_0$, for $j = 1, \ldots, N$, recalling that $\alpha < 0$ and $\theta > -\alpha$. See Lemma S4 in the supplementary material for a precise statement of this simple fact. The construction of the IBP, on the other hand, is more elaborate, and it involves infinitely many labels $\tilde{X}_j$ and probabilities $\tilde{q}_j$. Let us define the class of homogeneous completely random measures (CRMs, Kingman, 1967) without fixed atoms, which are characterized by a Laplace functional of the following type

$$\mathbb{E}[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}] = \exp\left\{-\int_{\mathbb{X}}\int_0^\infty \left[1 - e^{-sf(x)}\right]\rho(\mathrm{d}s)P_0(\mathrm{d}x)\right\},$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}_+$, where $\rho(\mathrm{d}s)$ is an intensity measure on $\mathbb{R}_+$, identifying the distribution of the probabilities $(\tilde{q}_j)_{j\geq 1}$, and P_0 is a diffuse distribution on $\mathbb{X}$ from which the labels $\tilde{X}_j$ are sampled. We will write $\tilde{\mu} \sim CRM(\rho; P_0)$. We refer to Daley and Vere-Jones (2008) for a mathematical treatment of CRMs and Lijoi and Prünster (2010) for a presentation of CRMs as a unifying concept in Bayesian nonparametrics. In model (6) the random measure $\tilde{\mu}$ must have jumps $\tilde{q}_j \in (0, 1)$, hence we require the intensity measure $\rho(\mathrm{d}s)$ of the CRM to be supported in $(0, 1)$. As shown in Teh and Gorur (2009), in the IBP the measure $\tilde{\mu} \mid \gamma$ is distributed as a completely random measure and, more precisely, it follows a stable-beta process, whose intensity measure $\rho(\mathrm{d}s)$ is

$$\rho(\mathrm{d}s) = \gamma\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}s. \tag{7}$$

Note that γ is sometimes called the *total mass* parameter because $\gamma = \mathbb{E}[\tilde{\mu}(\mathbb{X})] = \sum_{j\geq 1} \mathbb{E}[\tilde{q}_j]$ is the expected sum of all the probabilities. The choice $\alpha = 0$ leads to the beta process of Hjort (1990), as it was established by Thibaux and Jordan (2007). There exist several sampling strategies for the weights $\tilde{q}_j \in (0, 1)$, for example based on size-biased constructions or the inverse of the Lévy measure (Teh & Gorur, 2009). Alternative and more recent approaches include stick-breaking representations (Broderick et al., 2012), or independent finite approximations (Lee et al., 2023; Nguyen et al., 2024).

3 Predictive structure of Gibbs-type feature models

3.1 A buffet metaphor for Gibbs-type feature models

We begin our theoretical investigation of Gibbs-type feature models by presenting the predictive distribution for the $(n + 1)$th individual, given a sample of n data points. In the notation of Section 2.3, we study the conditional distribution of Z_{n+1}, given a random sample $\boldsymbol{Z}^{(n)} = (Z_1, \ldots, Z_n)$, where the latter entails $K_n = k$ observed features $X_1, \ldots, X_k$ whose presence is encoded by the binary variables $A_{i,\ell}$'s. The relevant aspects of the distribution of Z_{n+1} are conveyed by the vector of random variables $(Y_{n+1}, A_{n+1,1}, \ldots, A_{n+1,k})$ such that: (i) Y_{n+1} is the number of new features displayed by the $(n + 1)$th individual, i.e. the features hitherto unobserved in the sample $\boldsymbol{Z}^{(n)}$; (ii) each $A_{n+1,\ell}$ is a binary random variable such that $A_{n+1,\ell} = 1$ if the $(n + 1)$th individual displays feature X_ℓ and $A_{n+1,\ell} = 0$ otherwise. Our first key result provides the predictive law of Gibbs-type feature models, i.e. the probability distribution

$$p_{n+1}(y, a_1, \ldots, a_k) := \mathbb{P}((Y_{n+1}, A_{n+1,1}, \ldots, A_{n+1,k}) = (y, a_1, \ldots, a_k) \mid \boldsymbol{Z}^{(n)}).$$

We will write $\mathscr{B}(a; p) = p^a(1 - p)^{1-a}$ to denote the probability mass function of a Bernoulli random variable with parameter $p \in (0, 1)$ evaluated at $a \in \{0, 1\}$.

Theorem 1 (Predictive law). Suppose the EFPF is in product form (1), then the predictive law is

$$p_{n+1}(y, a_1, \ldots, a_k) = \binom{k+y}{k} \frac{V_{n+1,k+y}}{V_{n,k}} \{(\theta + \alpha)_n\}^y (\theta + n)^k \prod_{\ell=1}^{k} \mathscr{B}\left(a_\ell; \frac{m_\ell - \alpha}{\theta + n}\right).$$

An important remark is in order: given the sample $\boldsymbol{Z}^{(n)}$, the random variable Y_{n+1} is independent on the binary random variables $A_{n+1,1}, \ldots, A_{n+1,k}$, which are also mutually independent. This is a consequence of the product form representation (1). In the second place, the count Y_{n+1} of new features depends on the sample $\boldsymbol{Z}^{(n)}$ through the sample size n and the number of observed features $K_n = k$, but not the frequencies $m_1, \ldots, m_k$. It is also noteworthy that what distinguishes Gibbs-type feature models is only the distribution of the number of new features Y_{n+1}. In fact, the probability distribution of the variables referring to the previously observed features, $A_{n+1,1}, \ldots, A_{n+1,k}$, is common to all Gibbs-type feature models and does not depend on the chosen set of $V_{n,k}$'s. We will provide more precise comments on the distribution of Y_{n+1} when presenting specific examples in Sections 4 and 5.

As an immediate consequence of Theorem 1, one can easily determine the probability that the $(n + 1)$th individual does not exhibit new features. Such a probability may be interpreted as a sample-based version of the *sample coverage* (Good, 1953; Good & Toulmin, 1956), that is, the probability of re-observing a feature among those in the sample. However, it is worth noting that other definitions of sample coverage have been proposed in the feature setting; see, for example, Chiu (2023) and references therein.

Corollary 1 (Sample coverage). Suppose the EFPF is in product form (1), then the probability that Z_{n+1} does not show any new features, given $\boldsymbol{Z}^{(n)}$, is

$$\mathbb{P}(\text{“}Z_{n+1} \textit{ has no new features}\text{”} \mid \boldsymbol{Z}^{(n)}) = \mathbb{P}(Y_{n+1} = 0 \mid \boldsymbol{Z}^{(n)}) = \frac{V_{n+1,k}}{V_{n,k}} (\theta + n)^k.$$

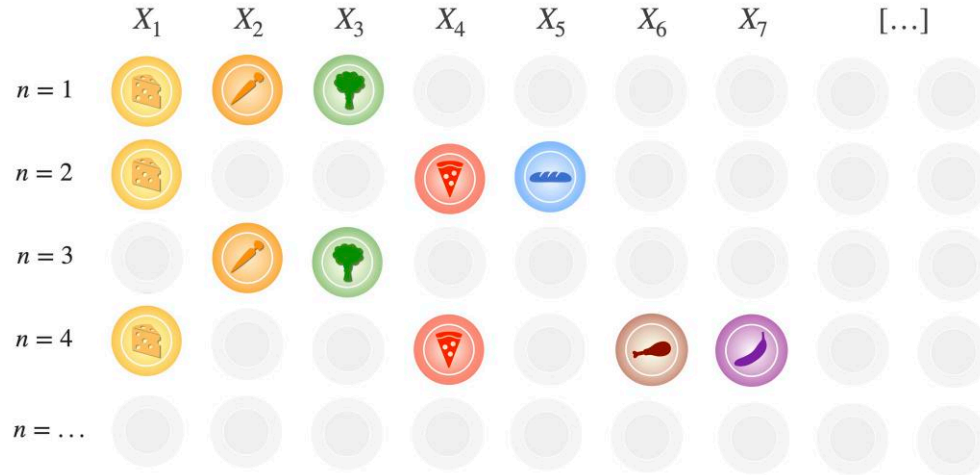

Figure 2. The buffet metaphor for Gibbs-type feature models. In this example there are $n = 4$ customers and $K_n = 7$ dishes. The observed frequencies of the dishes are $(m_1, \ldots, m_7) = (3, 2, 2, 2, 1, 1, 1)$. The numbers of new dishes picked by each customer (from top to bottom) are $Y_1 = 3$, $Y_2 = 2$, $Y_3 = 0$, and $Y_4 = 2$.

The predictive distribution presented in Theorem 1 can be likened to the Indian buffet metaphor (Griffiths & Ghahramani, 2011). Our metaphor imagines a scenario where "customers", representing individuals, sequentially enter a restaurant and select a number of "dishes", corresponding to feature labels $(\tilde{X}_j)_{j\geq 1}$, as depicted in Figure 2. Each customer has the option to choose from previously ordered dishes or select new ones. For any Gibbs-type feature model, the generative process unfolds as follows: the first customer enters the restaurant and selects Y_1 dishes according to the distribution

$$\mathbb{P}(Y_1 = y) = V_{1,y},$$

and $K_1 = Y_1$. The K_1 selected dishes are associated with labels X_ℓ, for $\ell = 1, \ldots, K_1$, provided that $K_1 > 0$. Then, the $(n+1)$th customer enters and selects dishes in two steps. First, the customer picks Y_{n+1} new dishes (not chosen by the previous n customers) according to the distribution

$$\mathbb{P}(Y_{n+1} = y \mid K_n) = \binom{k+y}{k} \frac{V_{n+1,k+y}}{V_{n,k}} \{(\theta+\alpha)_n\}^y (\theta+n)^k,$$

where $K_n = k$ denotes the number of distinct dishes chosen by the first n customers, so that $K_{n+1} = K_n + Y_{n+1}$. If $Y_{n+1} > 0$, then the new dishes are associated with labels X_ℓ, for $\ell = K_n + 1, \ldots, K_{n+1}$. Second, the $(n+1)$th customer may select some of the previously chosen dishes $X_1, \ldots, X_k$ as encoded by the binary variables $A_{n+1,1}, \ldots, A_{n+1,k}$, whose distribution is

$$A_{n+1,\ell} \mid \boldsymbol{Z}^{(n)} \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{m_\ell - \alpha}{\theta + n}\right), \tag{8}$$

where m_ℓ corresponds to the number of previous customers who selected dish X_ℓ. Higher values of m_ℓ correspond to a higher probability of dish X_ℓ being selected again.

This general buffet metaphor provides a simple sampling strategy for any Gibbs-type feature allocation model; moreover, it offers a clear interpretation for the parameters θ and α. Essentially, the posterior probability of observing the ℓth feature can be viewed as the weighted combination of two factors: one representing the observed data and the other reflecting prior beliefs, akin to a typical Bayesian updating rule. Specifically, for $\ell = 1, \ldots, k$:

$$\mathbb{P}(A_{n+1,\ell} = 1 \mid \boldsymbol{Z}^{(n)}) = \frac{m_\ell - \alpha}{\theta + n} = \frac{n}{\theta + n}\, \hat{p}_\ell + \frac{\theta}{\theta + n}\left(-\frac{\alpha}{\theta}\right),$$

where $\hat{p}_\ell = m_\ell/n$ denotes the empirical frequency of the ℓth observed feature. When $\alpha < 0$, the quantity $-\alpha/\theta \in (0, 1)$ can be conveniently interpreted as the prior frequency of a feature, while θ assumes the familiar role of prior sample size. Small values of $\theta > 0$ suggest low confidence in the prior guess, and vice versa. In cases where $\alpha \in [0, 1)$, θ continues to control the relevance of prior beliefs, while $-\alpha/\theta$ determines the extent of (potentially negative) shrinkage on the probability of re-observing an old feature.

Finally, we point out that the buffet metaphor for the IBP, also discussed in Teh and Gorur (2009), emerges as a special case. In the IBP, the predictive process is characterized by $Y_1 \mid \gamma \sim \text{Poisson}(\gamma)$ and $Y_{n+1} \mid K_n, \gamma \sim \text{Poisson}(\gamma(\theta+\alpha)_n/(\theta+1)_n)$ for $n \geq 1$. Moreover, as can be verified via Theorem 1 using $V_{n,k}$ from (3), the BB model leads to $Y_1 \mid N \sim \text{Binomial}(N, -\alpha/\theta)$ and $Y_{n+1} \mid K_n, N \sim \text{Binomial}(N-k, -\alpha/(\theta+n))$, where $\text{Binomial}(n_0, p)$ denotes a binomial random variable with parameters $n_0 \in \mathbb{N}$ and $p \in (0, 1)$. Consequently, in a BB model, no new features can be displayed once $K_n = N$ features are observed.

3.2 Distribution of the number of features

We now investigate distributional properties of K_n, the total number of distinct features observed in a sample of size n. Note that we can also express K_n as $Y_1 + Y_2 + \cdots + Y_n$, which is the summation of newly discovered features at each step of the buffet metaphor. In ecological applications, the expectations $\mathbb{E}(K_1), \ldots, \mathbb{E}(K_n)$ should be regarded as a model-based *rarefaction* curve (Gotelli & Colwell, 2001; Zito et al., 2023), with the fundamental difference, compared to species sampling models, that our approach is appropriate for sample-based accumulation curves, and not individual-based. While these two frameworks are not comparable, as they refer to different sampling designs, there are several analogies between our work and Lijoi et al. (2007a). For example, the following theorem describes the distribution of K_n for any Gibbs-type feature model, which is the direct equivalent of a result of Gnedin and Pitman (2005) for species sampling models.

Theorem 2 (Number of features K_n). Suppose the EFPF is in product form (1) and let $\alpha < 0$. Then, the number of distinct features observed in a sample of n individuals, K_n, satisfies, for any $k \geq 0$,

$$\mathbb{P}(K_n = k) = V_{n,k} \left\{ \frac{(\theta+\alpha)_n - (\theta)_n}{\alpha} \right\}^k.$$

Alternatively, let $g_n(\theta, \alpha)$ be defined as in (2), if $\alpha \in [0, 1)$ then for any $k \geq 0$,

$$\mathbb{P}(K_n = k) = V_{n,k} \left\{ g_n(\theta, \alpha)(\theta+1)_{n-1} \right\}^k.$$

A preliminary version of this result is also presented in Battiston et al. (2018), up to some further simplifications. In Sections 4 and 5, we will specialize the formulas of Theorem 2 for specific choices of $V_{n,k}$'s, recovering well-known distributions. In Proposition S1 of the supplementary material, we also provide distributional insights for the statistic $K_{n,r}$ for $r \in \{1, \ldots, n\}$, denoting the number of features appearing exactly r times among n individuals, thereby defining $K_n = K_{n,1} + \cdots + K_{n,n}$. Note that this is helpful to determine the law of the number of rare features, i.e. features expressed only in a single individual, corresponding to $K_{n,1}$.

Theorem 2 and Proposition S1 of the supplementary material pertain to *a priori* properties of K_n and $K_{n,r}$. We now consider the much more compelling problem of predicting the number of hitherto unseen features $K_m^{(n)}$ that would be observed in an additional sample of size m, conditioned on the available data $\boldsymbol{Z}^{(n)}$, i.e. *a posteriori*. In ecological contexts, this leads to the *extrapolation* of the accumulation curve (Zito et al., 2023), namely the expectations $\mathbb{E}(K_{n+m} \mid \boldsymbol{Z}^{(n)}) = k + \mathbb{E}(K_m^{(n)} \mid \boldsymbol{Z}^{(n)})$ for any $m \geq 1$. What follows is a key result of this paper since we provide the distribution of $K_m^{(n)}$ for any Gibbs-type feature model, analogous to the main finding of (Proposition 1, Lijoi et al., 2007a) for Gibbs-type species sampling models.

Theorem 3 (Number of hitherto unobserved features). Suppose the EFPF is in product form (1). Then, the distribution of $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}$ is such that, for any $y \geq 0$,

$$\mathbb{P}(K_m^{(n)} = y \mid \boldsymbol{Z}^{(n)}) = \binom{k+y}{k} \frac{V_{n+m,k+y}}{V_{n,k}} \{(\theta+n)_m\}^{k+y} \{c_{n,m}(\theta,\alpha)\}^y,$$

where, if $\alpha < 0$,

$$c_{n,m}(\theta,\alpha) = \frac{(\theta+\alpha)_n}{\alpha} \left(\frac{(\theta+\alpha+n)_m}{(\theta+n)_m} - 1 \right),$$

and if $\alpha \in [0, 1)$,

$$c_{n,m}(\theta,\alpha) = (\theta+1)_{n-1} \{g_{n+m}(\theta,\alpha) - g_n(\theta,\alpha)\}.$$

The predictive distribution for Y_{n+1} corresponds to the special case $m = 1$ in the above theorem, as it can be easily checked. Moreover, from Theorem 3, it is evident that the distribution of $K_m^{(n)}$ depends on the data $\boldsymbol{Z}^{(n)}$ only through the sample size n and the number of distinct features K_n, but not on the feature counts m_ℓ. In other words, the number of observed features $K_n = k$ is sufficient for predicting $K_m^{(n)}$. This remarkable property is also a characteristic of Gibbs-type priors (Lijoi et al., 2007a). We refer once again to Sections 4 and 5 for tractable special cases of Theorem 3.

3.3 Asymptotic behaviour and α-diversity

The α parameter of a Gibbs-type feature model plays a key role, as hinted by Proposition 1. We show that α identifies the asymptotic behaviour of K_n, precisely as in Gibbs-type species sampling models (De Blasi et al., 2015). In particular, as $n \to \infty$, the number K_n converges to a finite random variable whenever $\alpha < 0$, and it diverges when $\alpha \in [0, 1)$. Moreover, K_n grows at a logarithmic rate when $\alpha = 0$ and at a polynomial rate when $\alpha \in (0, 1)$. This behaviour is well known for BB and IBP models (e.g. Griffiths & Ghahramani, 2011; Teh & Gorur, 2009), but it is in fact a general property of Gibbs-type feature models, as the following proposition illustrates.

Proposition 2 (α-diversity). Suppose the EFPF is in product form (1) and let K_n be the number of features observed in a sample of n individuals, as $n \to \infty$

(i) if $\alpha < 0$, then $K_n \xrightarrow{\text{d}} N$,

(ii) if $\alpha = 0$, then $K_n / \log(n) \xrightarrow{\text{d}} \gamma\theta$,

(iii) if $0 < \alpha < 1$, then $K_n / n^\alpha \xrightarrow{\text{d}} \gamma\Gamma(\theta+1)/\{\alpha\Gamma(\theta+\alpha)\}$,
where the random variables N and γ have distributions P_γ and P_N as in the mixture representation of the weights $V_{n,k}$'s described in Proposition 1.

Summarizing, let $c_\alpha(n)$ be a function such that $c_\alpha(n) = 1$ if $\alpha < 0$, $c_\alpha(n) = \log(n)$ if $\alpha = 0$, and $c_\alpha(n) = n^\alpha$ if $\alpha \in (0, 1)$, then, in general, $K_n / c_\alpha(n) \xrightarrow{\text{d}} S_\alpha$, as $n \to \infty$. We call random variable S_α the α-*diversity* of a feature allocation model, analogous to the α-diversity introduced by Pitman (2003). The random variable S_α is often of direct interest in ecological problems, as it represents a synthetic biodiversity measurement. Naturally, comparing α-diversities across different datasets makes sense only if they are based on the same growth rate. Note that, for fixed values of α and θ, in the BB and IBP models the α-diversity is deterministic. In practice, the α-diversity is unknown, and it may be estimated employing a prior distribution for N or γ, which results in a Gibbs-type feature model thanks to Proposition 1. Moreover, the posterior law of γ and N may be obtained by

combining the prior density $p_\gamma(\mathrm{d}\gamma)$ associated to P_γ, or the prior probability distribution $p_N(y)$ associated to P_N, with the EFPF of equations (2) and (3), giving respectively

$$p_\gamma(\mathrm{d}\gamma \mid \boldsymbol{Z}^{(n)}) \propto p_\gamma(\mathrm{d}\gamma)\gamma^k \exp\{-\gamma g_n(\theta, \alpha)\}, \quad p_N(y \mid \boldsymbol{Z}^{(n)}) \propto p_N(y) \frac{y!}{(y-k)!} \left(\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^y, \tag{9}$$

for $y = k, k+1, \ldots$. We note that under suitable choices for $p_\gamma(\mathrm{d}\gamma)$ and $p_N(y)$, the posterior law corresponds to well-known distributions. Such a posterior distribution of S_α also has an elegant connection with accumulation curves, as shown in the following proposition.

Proposition 3 (Posterior law of the α-diversity). Suppose the EFPF is in product form (1). Let $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}$ be the number of hitherto unseen features and let S_α be the α-diversity. Then, as $m \to \infty$

$$\frac{K_m^{(n)} + k}{c_\alpha(m)} \mid \boldsymbol{Z}^{(n)} \xrightarrow{\mathrm{d}} S'_\alpha, \qquad S'_\alpha \overset{\mathrm{d}}{=} S_\alpha \mid \boldsymbol{Z}^{(n)}.$$

Thus, the posterior law of S_α coincides with the α-diversity associated with the extrapolation of the accumulation curve $K_m^{(n)} + k \mid \boldsymbol{Z}^{(n)}$, providing an insightful alternative perspective.

3.4 Posterior characterization

We conclude our theoretical investigation with another pivotal result: the determination of the posterior distribution arising from model (6) of $\tilde{\mu} = \sum_{j\geq 1} \tilde{q}_j \delta_{\tilde{X}_j}$, given $\boldsymbol{Z}^{(n)}$. This posterior characterization of $\tilde{\mu}$ not only elucidates the learning mechanism underpinning Gibbs-type feature models but also enables the simulation of arbitrary functionals of interest associated with $\tilde{\mu}$. Its availability also proves advantageous for Markov Chain Monte Carlo algorithms when Gibbs-type feature models are employed as latent building blocks of more complex models.

Posterior characterizations have already been studied for specific models. For the two-parameter IBP, namely the beta process, Thibaux and Jordan (2007) used the conjugacy result of Hjort (1990) to obtain the posterior distribution. For the IBP with $\alpha \in (0, 1)$, namely the stable-beta process, the posterior can be deduced by carefully reading Teh and Gorur (2009), which, in turn, is based on Kim (1999). Finally, for the BB model, the posterior derivation is straightforward thanks to the independence among the $\tilde{q}_j$'s and the beta-binomial conjugacy. A systematic discussion for the broad class of CRMs priors for $\tilde{\mu}$ is provided in James (2017). Refer also to Camerlenghi et al. (2024) for related findings. The following theorem applies to any Gibbs-type feature model, albeit with notable simplifications forthcoming in Sections 4 and 5, where we discuss specific choices for the priors of γ and N.

Theorem 4 (Posterior distribution of $\tilde{\mu}$). Suppose $\boldsymbol{Z}^{(n)} = (Z_1, \ldots, Z_n)$ follows model (6), then the posterior distribution of $\tilde{\mu}$, given $\boldsymbol{Z}^{(n)}$, satisfies the following decomposition

$$\tilde{\mu} \mid \boldsymbol{Z}^{(n)} \overset{\mathrm{d}}{=} \tilde{\mu}' + \tilde{\mu}^*, \tag{10}$$

where $\tilde{\mu}^* \overset{\mathrm{d}}{=} \sum_{\ell=1}^k q_\ell \delta_{X_\ell}$ is a random measure such that $q_\ell \overset{\mathrm{iid}}{\sim} \mathrm{Beta}(m_\ell - \alpha, \alpha + \theta + n - m_\ell)$, for $\ell = 1, \ldots, k$, and $X_1, \ldots, X_k$ denote the observed features. Moreover, the random measure $\tilde{\mu}'$ in (10) is independent of $\tilde{\mu}^*$, and its distribution depends on the value of α, as specified below.

(i) If $\alpha < 0$, then the random measure $\tilde{\mu}' \mid N' \overset{\mathrm{d}}{=} \sum_{j=1}^{N'} q'_j \delta_{\tilde{X}_j}$, where $q'_j \overset{\mathrm{iid}}{\sim} \mathrm{Beta}(-\alpha, \theta + \alpha + n)$ and $\tilde{X}_j \overset{\mathrm{iid}}{\sim} P_0$, for $j = 1, \ldots, N'$. Moreover, $N' + k \overset{\mathrm{d}}{=} N \mid \boldsymbol{Z}^{(n)}$ as in (9).

(ii) If $\alpha \in [0, 1)$, then the random measure $\tilde{\mu}' \mid \gamma' \sim CRM(\rho'; P_0)$ with updated intensity $\rho'(\mathrm{d}s) = \gamma'\Gamma(1+\theta)/\{\Gamma(1-\alpha)\Gamma(\theta+\alpha)\}s^{-\alpha-1}(1-s)^{\theta+\alpha+n-1}\mathrm{d}s$. Moreover, $\gamma' \overset{\mathrm{d}}{=} \gamma \mid \boldsymbol{Z}^{(n)}$ as in (9).

The previous distributional equality (10) decomposes the posterior distribution of $\tilde{\mu}$ into two parts: $\tilde{\mu}'$ accounts for the newly observed features, while $\tilde{\mu}^*$ deals with those previously observed. Regarding the latter, the $(n+1)$th individual exhibits an existing feature X_ℓ if $A_{n+1,\ell} = 1$, where each $A_{n+1,\ell} \mid q_\ell \overset{\text{iid}}{\sim} \text{Bernoulli}(q_\ell)$ and $q_\ell \overset{\text{iid}}{\sim} \text{Beta}(m_\ell - \alpha, \alpha + \theta + n - m_\ell)$. By integrating out the q_ℓ's, we obtain $A_{n+1,\ell} \mid \boldsymbol{Z}^{(n)} \overset{\text{iid}}{\sim} \text{Bernoulli}((m_\ell - \alpha)/(\theta + n))$, consistent with the predictive structure (8). It is worth highlighting that the distribution of $\tilde{\mu}^*$ remains the same across all Gibbs-type feature models. However, $\tilde{\mu}'$ exhibits structural differences depending on the specific choices for the $V_{n,k}$'s. In the case of $\alpha \in [0, 1)$, we observe that $\tilde{\mu}' \mid \gamma' \sim CRM(\rho'; P_0)$ benefits from a conjugacy property, as the intensity measure $\rho'(\mathrm{d}s)$ characterizes a stable-beta process with updated parameters $(\gamma'(\alpha+\theta)_n/(\theta+1)_n, \alpha, \theta+n)$, a point noted by Teh and Gorur (2009) in the IBP case.

Simulating posterior samples for $\tilde{\mu}$ is straightforward. Initially, one needs to draw values from $N \mid \boldsymbol{Z}^{(n)}$ or $\gamma \mid \boldsymbol{Z}^{(n)}$, which typically follow highly tractable distributions. Then, conditioned on N or γ, both random measures $\tilde{\mu}'$ and $\tilde{\mu}^*$ are simple to sample from: $\tilde{\mu}^*$ involves a finite number of beta random variables, as does $\tilde{\mu}'$ when $\alpha < 0$. Even simulating $\tilde{\mu}'$ when $\alpha \in [0, 1)$ is straightforward, as due to conjugacy, $\tilde{\mu}'$ follows a stable-beta process, for which efficient sampling algorithms exist; see, for example, Teh and Gorur (2009).

4 Gamma mixture of IBPs

4.1 Predictive structure, number of features, and α-diversity

In this section, we discuss relevant special cases of Gibbs-type feature models within the $\alpha = 0$ and $\alpha \in (0, 1)$ regimes, where there are infinitely many possible features $\tilde{X}_j$. We define a new Gibbs-type feature model termed *gamma mixture of* IBPs by employing a gamma prior for γ. Upon examining the posterior distribution in (9), it becomes evident that a gamma prior is *conjugate*, as its posterior remains gamma with updated parameters. If $\gamma \sim \text{Gamma}(a, b)$, then the associated EFPF, obtained by integrating (2) with respect to the prior density, follows a product form (1) and the weights are

$$V_{n,k} = \frac{1}{k!}\frac{b^a(a)_k}{\{(\theta+1)_{n-1}\}^k\{b + g_n(\theta, \alpha)\}^{a+k}}. \tag{11}$$

This Gibbs-type feature model has connections with the stable-beta scaled process of Camerlenghi et al. (2024) , which is, in fact, a special case of (11). In particular, a stable-beta scaled process with parameters (α, c, d) can be represented as a gamma mixture of IBPs under the constraint $\theta = 1 - \alpha$ and prior distribution $\gamma \sim \text{Gamma}(c+1, d(1-\alpha)/\alpha)$. Such a hierarchical representation is not discussed in Camerlenghi et al. (2024), but it can be proved by directly inspecting the EFPFs.

We now compare the IBP of Teh and Gorur (2009) with the feature model (11), utilizing the general findings from Section 3. The predictive laws of the IBP and the gamma mixture of IBPs follow by specializing Theorem 1, substituting the $V_{n,k}$'s of equations (2) and (11), respectively, into the general formulas. As discussed in Section 3.1, the predictive distributions of Gibbs-type feature models differ only in the law governing the number of new features, whereas the law of the binary variables $A_{n+1,1}, \ldots, A_{n+1,k}$, already described in (8), is the same. Thus, for the sake of brevity, here we concentrate on the distribution of the number of new features Y_{n+1}. Let $Y \sim \text{NegBinomial}(n_0, \mu_0)$ denote a negative binomial random variable with mean parameter $\mu_0 > 0$ and dispersion parameter $n_0 > 0$, where its probability mass function $N(y; n_0, \mu_0) \propto p^{n_0}(1-p)^y$ is defined for any $y \in \mathbb{N}_0$, with $p = n_0/(\mu_0 + n_0)$, so that $\mathbb{E}(Y) = \mu_0$ and $\text{Var}(Y) = \mu_0 + \mu_0^2/n_0$. Moreover, let $Y_{n+1} \mid K_n, \gamma$ be the number of new features for the $(n+1)$th individual in the IBP

case, and let $Y_{n+1} \mid K_n$ be the same quantity for the gamma mixture model. Then, simple calculus yields

$$Y_{n+1} \mid K_n, \gamma \sim \text{Poisson}\left(\gamma \frac{(\theta+\alpha)_n}{(\theta+1)_n}\right), \quad Y_{n+1} \mid K_n \sim \text{NegBinomial}\left(a+k, \frac{a+k}{b+g_n(\theta,\alpha)} \frac{(\theta+\alpha)_n}{(\theta+1)_n}\right).$$

Additional distributional properties can be derived by specializing the results of Section 3 for the IBP and gamma mixture of IBPs. We summarize some of these properties in the following proposition and refer to the supplementary material for additional findings and simplifications, such as the distribution of the number of shared features $K_{n,r}$ or the sample coverage.

Proposition 4 Suppose the EFPF is in product form (1) with $\alpha \in [0, 1)$. Let $K_n \mid \gamma$ and K_n be the number of features observed in a sample of n individuals for the IBP and the gamma mixture in (11). Then, we have

$$\begin{aligned} K_n \mid \gamma &\sim \text{Poisson}\big(\gamma g_n(\theta,\alpha)\big), \\ K_n &\sim \text{NegBinomial}\big(a, (a/b) g_n(\theta,\alpha)\big). \end{aligned} \tag{12}$$

Moreover, let $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}$ be the number of hitherto unseen features in an additional sample of size $m \geq 1$, then for the IBP

$$K_m^{(n)} \mid \boldsymbol{Z}^{(n)}, \gamma \sim \text{Poisson}\big(\gamma\big(g_{n+m}(\theta,\alpha) - g_n(\theta,\alpha)\big)\big), \tag{13}$$

whereas for the gamma mixture

$$K_m^{(n)} \mid \boldsymbol{Z}^{(n)} \sim \text{NegBinomial}\left(a+k, \frac{a+k}{b+g_n(\theta,\alpha)}\big(g_{n+m}(\theta,\alpha) - g_n(\theta,\alpha)\big)\right). \tag{14}$$

The results regarding the IBP have been deduced from the general theory outlined in Theorems 2 and 3, albeit these results were already known. Indeed, the distribution of $K_n \mid \gamma$ has been determined by Teh and Gorur (2009), while the distribution of $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}, \gamma$ has been unveiled by Masoero et al. (2022). Conversely, the results concerning the gamma mixture are novel. The expected values of the number of features, also known as rarefaction, are $\mathbb{E}(K_n \mid \gamma) = \gamma g_n(\theta,\alpha)$ and $\mathbb{E}(K_n) = (a/b) g_n(\theta,\alpha)$. The function $g_n(\theta,\alpha)$ has an interesting interpretation, being the expected value of the number of clusters in a sample of size n from a Pitman–Yor process (Pitman & Yor, 1997) with parameters (α,θ). The $\alpha = 0$ case corresponds to the Dirichlet process (Ferguson, 1973), reducing to $g_n(\theta, 0) = \sum_{i=1}^{n} \theta/(\theta+i-1)$. This fact underscores once more the close relationship between Gibbs-type feature models and Gibbs-type species sampling models.

One notable advantage of the negative binomial distribution derived from the gamma mixture model is its ability to introduce overdispersion in K_n. Furthermore, the posterior distribution of $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}, \gamma$, corresponding to the IBP case, remains independent of the number of observed features $K_n = k$, a characteristic that some may find unappealing. In contrast, the negative binomial posterior for $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}$ considers k, influencing both the mean and variance of the distribution. Higher values of k result in overdispersion, which is often desirable. The Bayesian estimators for the number of previously unseen features, crucial for extrapolating the accumulation curve, are as follows:

$$\mathbb{E}(K_m^{(n)} \mid \boldsymbol{Z}^{(n)}, \gamma) = \gamma\big(g_{n+m}(\theta,\alpha) - g_n(\theta,\alpha)\big), \quad \mathbb{E}(K_m^{(n)} \mid \boldsymbol{Z}^{(n)}) = \frac{a+k}{b+g_n(\theta,\alpha)}\big(g_{n+m}(\theta,\alpha) - g_n(\theta,\alpha)\big),$$

for the IBP and the gamma mixture of IBPs, respectively.

Recall that, as stated in Proposition 2, the parameter α controls the growth rate of K_n so that when $\alpha = 0$, then $K_n/\log(n) \xrightarrow{\mathrm{d}} \gamma\theta$, whereas when $0 < \alpha < 1$, then $K_n/n^\alpha \xrightarrow{\mathrm{d}} \gamma\Gamma(\theta+1)/\{\alpha\Gamma(\theta+\alpha)\}$. In the IBP, the parameter γ, and hence the α-diversity, is deterministic (Masoero et al., 2022). Conversely, by definition, γ follows a priori a Gamma(a, b) in the gamma mixture model (11), which allows the Bayesian learning of the α-diversity through its posterior.

Proposition 5 Suppose the EFPF is in product form (1) with $\alpha \in [0, 1)$ and the $V_{n,k}$'s defined as in (11), so that a priori $\gamma \sim \text{Gamma}(a, b)$. Then, the posterior is $\gamma \mid \boldsymbol{Z}^{(n)} \sim \text{Gamma}(a+k, b+g_n(\theta,\alpha))$ and therefore the α-diversity, for $\alpha = 0$, is given by

$$\frac{K_m^{(n)}}{\log(m)} \mid \boldsymbol{Z}^{(n)} \xrightarrow{\mathrm{d}} S'_\alpha, \qquad S'_\alpha \overset{\mathrm{d}}{=} S_\alpha \mid \boldsymbol{Z}^{(n)} \sim \text{Gamma}\left(a+k, \frac{b+g_n(\theta,0)}{\theta}\right),$$

as $m \to \infty$, whereas for $\alpha \in (0, 1)$

$$\frac{K_m^{(n)}}{m^\alpha} \mid \boldsymbol{Z}^{(n)} \xrightarrow{\mathrm{d}} S'_\alpha,$$
$$S'_\alpha \overset{\mathrm{d}}{=} S_\alpha \mid \boldsymbol{Z}^{(n)} \sim \text{Gamma}\left(a+k, \{b+g_n(\theta,\alpha)\}\frac{\Gamma(\theta+\alpha)\alpha}{\Gamma(\theta+1)}\right).$$

A consequence of the deterministic α-diversity in the IBP is that the width of the credible intervals for $K_m^{(n)}$ grows at a rate slower than m^α. In contrast, the mixtures of IBPs yield larger credible intervals, whose widths grow proportionally to m^α, as highlighted by the previous proposition. This difference can be observed in simulation study B of the supplementary material: Figure S8 suggests that the IBP underestimates the uncertainty in predicting the number of unseen features. Proposition 5 presents some of the first results concerning the posterior distribution of the α-diversity for feature allocation models, with an early contribution for the stable-beta scaled process being available in Camerlenghi et al. (2024). Analogous findings for the Pitman–Yor species sampling model are given in Favaro et al. (2009) when $\alpha \in (0, 1)$, while for the Dirichlet process ($\alpha = 0$) an interpretable and tractable prior is proposed in Zito et al. (2024).

4.2 Posterior characterizations and negative binomial processes

We specialize here the posterior characterization of Theorem 4 for the gamma mixture of IBPs, which can be conveniently described in terms of *negative binomial processes* (Gregoire, 1984), whose use in Bayesian nonparametrics is still much unexplored. Building upon (Gregoire, 1984), we say that $\tilde{\mu}$ is a *negative binomial random measure* with parameter $a > 0$, intensity measure $\rho(\mathrm{d}s)$ on $\mathbb{R}_+$ and diffuse base measure P_0 on $\mathbb{X}$, if $\tilde{\mu}$ has Laplace functional

$$\mathbb{E}[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}] = \left\{1 + \int_{\mathbb{X}}\int_0^\infty \left[1 - e^{-sf(x)}\right]\rho(\mathrm{d}s)P_0(\mathrm{d}x)\right\}^{-a}, \tag{15}$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}_+$. We will write $\tilde{\mu} \sim NB(a, \rho; P_0)$ and we assume the intensity measure $\rho(\mathrm{d}s)$ is supported in $(0, 1)$ as before. A negative binomial random measure may arise by considering a CRM with random intensity measure $\tilde{c}\rho(\mathrm{d}s)$, where $\tilde{c}$ is distributed as a gamma random variable with parameters $(a, 1)$. Hence, the hierarchical formulation for the gamma mixture of IBPs becomes

$$\begin{aligned} Z_i \mid \tilde{\mu} &\overset{\text{iid}}{\sim} BP(\tilde{\mu}), \quad i \geq 1, \\ \tilde{\mu} &\sim NB(a, \rho; P_0), \end{aligned} \tag{16}$$

where the intensity measure is $\rho(\mathrm{d}s) = (1/b)\Gamma(1+\theta)/\{\Gamma(1-\alpha)\Gamma(\theta+\alpha)\}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}s$. The proof is shown in Section S2 of the supplementary material, but it is merely a consequence of mixing the intensity measure of a completely random measure with respect to a gamma distribution. The following corollary of Theorem 4 characterizes the posterior distribution of the process $\tilde{\mu}$, given $\boldsymbol{Z}^{(n)}$, in terms of negative binomial random measures.

Corollary 2 Suppose $\boldsymbol{Z}^{(n)} = (Z_1, \ldots, Z_n)$ follows model (16), then the posterior distribution of $\tilde{\mu}$, given $\boldsymbol{Z}^{(n)}$, satisfies the decomposition $\tilde{\mu} \mid \boldsymbol{Z}^{(n)} \stackrel{\mathrm{d}}{=} \tilde{\mu}' + \tilde{\mu}^*$ in (10), where $\tilde{\mu}'$ and $\tilde{\mu}^*$ are independent random measures such that $\tilde{\mu}^*$ is distributed as in Theorem 4, whereas $\tilde{\mu}' \sim NB(a+k, \rho'; P_0)$, with updated intensity $\rho'(\mathrm{d}s) = 1/\{b + g_n(\theta, \alpha)\}\Gamma(1+\theta)/\{\Gamma(1-\alpha)\Gamma(\theta+\alpha)\}s^{-\alpha-1}(1-s)^{n+\theta+\alpha-1}\mathrm{d}s$.

5 Gibbs-type feature models with finitely many features

5.1 Predictive structure, number of features, and richness estimation

In species sampling models, mixtures of Dirichlet multinomial processes are an important subclass of Gibbs-type priors (see, e.g. De Blasi et al., 2015), which include, for instance, the models of Gnedin (2010) and De Blasi et al. (2013). In feature allocation models, a similar role is played by mixtures of BB models with N features, which corresponds to the $\alpha < 0$ case. These feature allocation models assume a finite number of features N, representing the *richness* in ecological problems. The standard BB model assumes that N is known in advance, although this is a critical parameter and object of inference. In the following, we concentrate on two novel and tractable specifications: (i) N is a Poisson random variable with parameter $\lambda > 0$, referred to as *Poisson mixture of* BBS; (ii) N is a negative binomial random variable with parameters (n_0, μ_0), referred to as *negative binomial mixture of* BBS. These random variables serve as the prior distribution for the richness, enabling its Bayesian learning. We begin by providing the expressions for the corresponding EFPFS.

Proposition 6 If $N \sim \text{Poisson}(\lambda)$ in the mixture representation of Proposition 1, then the model has EFPF in product form (1) and the $V_{n,k}$'s are given by

$$V_{n,k} = \frac{1}{k!}\exp\left\{-\lambda\left(1 - \frac{(\theta+\alpha)_n}{(\theta)_n}\right)\right\}\left\{\frac{-\lambda\alpha}{(\theta)_n}\right\}^k. \tag{17}$$

If instead $N \sim \text{NegBinomial}(n_0, \mu_0)$, then the $V_{n,k}$'s are given by

$$V_{n,k} = \binom{k+n_0-1}{k} \times \left\{\frac{-\alpha}{(\theta)_n}\frac{\mu_0}{\mu_0+n_0}\right\}^k \left(1 - \frac{\mu_0}{\mu_0+n_0}\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^{-n_0-k}\left(\frac{n_0}{\mu_0+n_0}\right)^{n_0}. \tag{18}$$

Clearly, the negative binomial mixture of BBS allows for a higher degree of prior uncertainty regarding the total number of features N compared to the Poisson case, as the negative binomial induces overdispersion. It is worth noting that the negative binomial mixture of BBS may be obtained by choosing a gamma prior for the λ parameter of the Poisson mixture of BBS. Specifically, assuming $N \mid \lambda \sim \text{Poisson}(\lambda)$, and $\lambda \sim \text{Gamma}(a, b)$ is equivalent to having $N \sim \text{NegBinomial}(n_0, \mu_0)$, with $n_0 = a$ and $\mu_0 = a/b$. This provides a hierarchical justification for the negative binomial mixture of BBS: one may initially consider the Poisson mixture model, but if there is uncertainty about λ, then one could learn it by employing a gamma prior, resulting in a negative binomial mixture of BBS.

We now apply the general results of Section 3 to the aforementioned mixtures of BB models. For brevity, we focus solely on the number of features K_n, representing the rarefaction, and the number

of hitherto unseen features $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}$, leading to the extrapolation of the accumulation curves. It is worth reiterating that the predictive distribution $Y_{n+1} \mid \boldsymbol{Z}^{(n)}$ involved in the buffet metaphor can be derived as a special case, setting $m = 1$ in the distribution of $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}$, which is a trivial task in the following formulas.

Proposition 7 Suppose the EFPF is in product form (1) with $\alpha < 0$ and let $p_n(\theta, \alpha) = 1 - (\theta + \alpha)_n/(\theta)_n$. If N is fixed, corresponding to the BB model (3), then

$$K_n \mid N \sim \text{Binomial}\big(N, p_n(\theta, \alpha)\big),$$
$$K_m^{(n)} \mid \boldsymbol{Z}^{(n)}, N \sim \text{Binomial}\big(N - k, p_m(\theta + n, \alpha)\big).$$

If instead $N \sim \text{Poisson}(\lambda)$, corresponding to model (17), then

$$K_n \sim \text{Poisson}\big(\lambda p_n(\theta, \alpha)\big),$$
$$K_m^{(n)} \mid \boldsymbol{Z}^{(n)} \sim \text{Poisson}\big(\lambda p_m(\theta + n, \alpha)\{1 - p_n(\theta, \alpha)\}\big).$$

Finally, if $N \sim \text{NegBinomial}(n_0, \mu_0)$, corresponding to model (18), then

$$K_n \sim \text{NegBinomial}\big(n_0, \mu_0 p_n(\theta, \alpha)\big),$$
$$K_m^{(n)} \mid \boldsymbol{Z}^{(n)} \sim \text{NegBinomial}\left(n_0 + k, \frac{n_0 + k}{n_0/\mu_0 + p_n(\theta, \alpha)} p_m(\theta + n, \alpha)\{1 - p_n(\theta, \alpha)\}\right).$$

Proposition 7 is the first result of this kind for feature allocation models with finitely many features. Moreover, it underscores the high degree of interpretability and transparency in Gibbs-type feature allocation models. Specifically, when N is deterministic the prior expectation for $\mathbb{E}(K_n \mid N) = N p_n(\theta, \alpha)$ depends on $p_n(\theta, \alpha) = 1 - (\theta + \alpha)_n/(\theta)_n$. This probability may be understood as the expected fraction of features observed in a sample of size n, out of a pool of N possible features. This elegant interpretation holds true also for the Poisson and negative binomial cases, as $\mathbb{E}(K_n) = \lambda p_n(\theta, \alpha)$ and $\mathbb{E}(K_n) = \mu_0 p_n(\theta, \alpha)$, respectively, so that $p_n(\theta, \alpha)$ can be read as the expected fraction of observed features out of the expected total number of features λ and μ_0. The conditional distributions for $K_m^{(n)}$ may also be expressed in terms of the probability $1 - p_n(\theta, \alpha)$, accounting for the old features, and the "updated" probability $p_m(\theta + n, \alpha)$, representing the future sample. If N is deterministic, the Bayesian estimator for the number of unseen features is $\mathbb{E}(K_m^{(n)} \mid \boldsymbol{Z}^{(n)}, N) = (N - k) p_m(\theta + n, \alpha)$, in which $p_m(\theta + n, \alpha)$ is the expected fraction of features in a future sample of size m, out of the remaining $N - k$ features. In the Poisson and negative binomial cases, where the total number of features N is learned from the data, the predictive mechanism is more sophisticated. The Bayesian estimators for the number of hitherto unseen features, under a Poisson and negative binomial prior for N, are

$$\mathbb{E}(K_m^{(n)} \mid \boldsymbol{Z}^{(n)}) = \begin{cases} \lambda p_m(\theta + n, \alpha)\{1 - p_n(\theta, \alpha)\}, & \text{(Poisson mixture)}, \\ \dfrac{n_0 + k}{n_0/\mu_0 + p_n(\theta, \alpha)} p_m(\theta + n, \alpha)\{1 - p_n(\theta, \alpha)\}, & \text{(negative binomial mixture)}. \end{cases}$$

Note that $\mathbb{E}(\lambda \mid \boldsymbol{Z}^{(n)}) = (n_0 + k)/\{n_0/\mu_0 + p_n(\theta, \alpha)\}$ is the posterior expectation of λ in the Poisson-gamma representation of the negative binomial, where μ_0/n_0 denotes the overdispersion. It is important to emphasize how the sampling information $\boldsymbol{Z}^{(n)}$ affects the distribution of the statistic $K_m^{(n)}$. In the Poisson mixture, the distribution of $K_m^{(n)}$ depends on the initial sample $\boldsymbol{Z}^{(n)}$ only through the sample size n. Conversely, under the negative binomial mixture, $K_m^{(n)}$ also depends on the number of distinct features $K_n = k$ observed in the initial sample. Finally, Proposition 7 offers a key motivation for adopting the proposed mixtures of BBS over the standard BB model. In the latter, the uncertainty around $K_m^{(n)}$ monotonically decreases for large values of m, and in the limit

$K_m^{(n)}$ degenerates to a point mass at $N-k$. This eventual shrinkage of uncertainty is a very undesirable behaviour. In contrast, under both the Poisson and negative binomial mixtures of BBS, the variance of $K_m^{(n)}$ monotonically increases with m, yielding a more realistic representation of uncertainty.

Finally, we study the total number of features N, i.e. the richness, which coincides with the α-diversity, since $K_n \xrightarrow{\text{d}} N$ as $n \to \infty$. The posterior distribution of the richness is one of the main quantities of interest in ecology and, as outlined in Proposition 3, it may be equivalently obtained by extrapolating the accumulation curve, specifically by considering $\lim_{m\to\infty} K_m^{(n)} + k \mid \boldsymbol{Z}^{(n)}$. For the BB model, the posterior distribution of N is deterministic, since N is known a priori. This yields critical issues as highlighted in simulation study A of the supplementary material (Figures S4 and S5). For the proposed mixtures of BBS, the following result can be easily proved using Proposition 7 and noting that $\lim_{m\to\infty} p_m(\theta+n,\alpha)=1$.

Proposition 8 Suppose the EFPF is in product form (1) with $\alpha<0$. Then $K_m^{(n)} \mid \boldsymbol{Z}^{(n)} \xrightarrow{\text{d}} N'$ with $N'+k \overset{\text{d}}{=} N \mid \boldsymbol{Z}^{(n)}$, as $m\to\infty$. Let $p_n(\theta,\alpha)=1-(\theta+\alpha)_n/(\theta)_n$, if $N\sim \text{Poisson}(\lambda)$, then

$$N' \sim \text{Poisson}\big(\lambda(1-p_n(\theta,\alpha))\big),$$

whereas if $N\sim \text{NegBinomial}(n_0,\mu_0)$, then

$$N' \sim \text{NegBinomial}\left(n_0+k,\ \frac{n_0+k}{n_0/\mu_0+p_n(\theta,\alpha)}\{1-p_n(\theta,\alpha)\}\right). \tag{19}$$

Note that, as before, the results have an appealing interpretation in terms of the probability $1-p_n(\theta,\alpha)$. The Bayesian estimators for the richness, under a Poisson and negative binomial prior for N, are the posterior expectations

$$\mathbb{E}(N \mid \boldsymbol{Z}^{(n)}) = \begin{cases} k+\lambda\{1-p_n(\theta,\alpha)\}, & \text{(Poisson mixture)},\\ k+\dfrac{n_0+k}{n_0/\mu_0+p_n(\theta,\alpha)}\{1-p_n(\theta,\alpha)\}, & \text{(negative binomial mixture)}. \end{cases}$$

To make these formulas operative, one needs to specify values for λ, θ, and α and possibly for n_0 and μ_0. We remark that all these quantities have a very transparent interpretation. Therefore, their elicitation may be based on prior information in many applied contexts. In our numerical studies, we will propose an empirical Bayes approach to set the hyperparameters. Moreover, in the supplementary material we investigate a fully Bayesian procedure in which we specify suitable priors for the hyperparameters.

5.2 Hierarchical formulation for mixtures of beta Bernoulli models

We now specialize the general posterior characterization in Theorem 4 for the Poisson and negative binomial mixtures of BB models. Recall that when $\alpha<0$ the underlying random measure $\tilde{\mu}$ for any Gibbs-type feature model in the hierarchical representation (6) can be described as $\tilde{\mu} \mid N = \sum_{j=1}^{N} \tilde{q}_j \delta_{\tilde{X}_j}$, with $\tilde{q}_j \overset{\text{iid}}{\sim} \text{Beta}(-\alpha,\theta+\alpha)$ and $\tilde{X}_j \overset{\text{iid}}{\sim} P_0$, for some prior distribution $N\sim P_N$. When N follows a Poisson distribution, this can be compactly expressed in terms of completely random measures. Specifically, in the the Poisson mixture of BBS, the statistical model which induces the EFPF in equation (17) may be written as

$$\begin{aligned} Z_i \mid \tilde{\mu} &\overset{\text{iid}}{\sim} BP(\tilde{\mu}), \quad i\geq 1\\ \tilde{\mu} &\sim CRM(\rho; P_0), \end{aligned} \tag{20}$$

where the intensity measure $\rho(\text{d}s)$ is finite since $\alpha<0$ and is proportional to the density of a $\text{Beta}(-\alpha,\theta+\alpha)$ distribution $\rho(\text{d}s)=\lambda\Gamma(\theta)/\{\Gamma(-\alpha)\Gamma(\theta+\alpha)\}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\text{d}s$. This result

follows by a construction of completely random measures with finitely many jumps $\tilde{q}_j$, whose number has a Poisson distribution (Daley & Vere-Jones, 2008). As for the negative binomial mixture, the EFPF (18) is associated with the following statistical model involving negative binomial processes:

$$\begin{aligned} Z_i \mid \tilde{\mu} &\overset{\text{iid}}{\sim} BP(\tilde{\mu}), \quad i \geq 1, \\ \tilde{\mu} &\sim NB(n_0, \rho; P_0), \end{aligned} \tag{21}$$

where $\rho(\mathrm{d}s) = (\mu_0/n_0)\Gamma(\theta)/\{\Gamma(-\alpha)\Gamma(\theta+\alpha)\}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}s$. The proof of these results is straightforward, but it is provided in Section S3 of the supplementary material for completeness. The following corollary is a consequence of Theorem 4, and it characterizes the posterior distribution of $\tilde{\mu}$ in both cases.

Corollary 3 Suppose $\boldsymbol{Z}^{(n)} = (Z_1, \ldots, Z_n)$ follows models (20) or (21), then the posterior distribution of $\tilde{\mu}$, given $\boldsymbol{Z}^{(n)}$, satisfies the decomposition $\tilde{\mu} \mid \boldsymbol{Z}^{(n)} \overset{\mathrm{d}}{=} \tilde{\mu}' + \tilde{\mu}^*$ in (10), where $\tilde{\mu}'$ and $\tilde{\mu}^*$ are independent random measures such that $\tilde{\mu}^*$ is distributed as in Theorem 4. Under model (20) $\tilde{\mu}' \sim CRM(\rho'; P_0)$, with updated intensity $\rho'(\mathrm{d}s) = \lambda\Gamma(\theta)/\{\Gamma(-\alpha)\Gamma(\theta+\alpha)\}s^{-\alpha-1}(1-s)^{\theta+n+\alpha-1}\mathrm{d}s$, whereas under model (21) $\tilde{\mu}' \sim NB(n_0+k, \rho'; P_0)$, with updated intensity $\rho'(\mathrm{d}s) = 1/\{n_0/\mu_0 + p_n(\theta,\alpha)\}\Gamma(\theta)/\{\Gamma(-\alpha)\Gamma(\theta+\alpha)\}s^{-\alpha-1}(1-s)^{\theta+n+\alpha-1}\mathrm{d}s$.

The relevance of this corollary is mostly theoretical. In practice, to simulate a posterior value for $\tilde{\mu}$ one relies on the hierarchical representation of Theorem 4, given the availability of the posterior for N given in Proposition 8. However, this posterior result unveils the central role of abstract constructions, such as completely random measures and negative binomial processes, even for allocation models with finitely many features. For example, it reveals that the lack of dependence on $K_n = k$ in the predictive structure of $K_m^{(n)}$, in the Poisson mixture model, follows because $\tilde{\mu}$ is distributed as CRMS, as explained in James (2017) and Camerlenghi et al. (2024).

6 Model fitting and simulation studies

6.1 Elicitation of the hyperparameters

We describe here an empirical Bayes approach to selecting the hyperparameters, albeit other Bayesian strategies could be considered. As for the two mixtures of BBS, we suggest to maximize the EFPF (1) of a BB model, obtained with $V_{n,k}$'s as in (3). This procedure provides us with an estimate $\hat{\alpha}$ and $\hat{\theta}$ of the values of α and θ, together with an estimate $\hat{N}$ for the total number of features N. Then, in the Poisson mixture we set $\lambda = \mathbb{E}(N) = \hat{N}$, whereas we set $\mu_0 = \mathbb{E}(N) = \hat{N}$ in the negative binomial mixture of BBS. We remark that, differently from the Poisson mixture, the negative binomial mixture also allows us to specify the variance of the prior distribution on N. We argue that such variance should just reflect the practitioner's degree of uncertainty around the prior guess $\mathbb{E}(N)$, possibly performing a sensitivity analysis for it. In our simulated scenarios, available in the supplementary material, we will consider additional choices of the prior expectation $\mathbb{E}(N)$ for the mixtures of BBS, instead of specifying it through the data-driven approach just described. In such cases, we estimate the parameters α and θ by maximizing the model-specific EFPF, that is (17) for the Poisson mixture, with $\lambda = \mathbb{E}(N)$, and (18) for the negative binomial mixture, with $\mu_0 = \mathbb{E}(N)$ and n_0 such that the desired prior variance is obtained.

Along a similar argument, for the mixtures of IBPS we first propose to maximize the EFPF (1) with $V_{n,k}$'s as in (2) to find an estimate $\hat{\alpha}$ and $\hat{\theta}$ for the parameters α, θ, together with an estimate $\hat{\gamma}$ for the total mass γ. Second, we choose the parameters of the prior for γ by enforcing the condition $\mathbb{E}(\gamma) = \hat{\gamma}$. In particular, for the gamma mixture of IBPS, we assume $\gamma \sim \text{Gamma}(a, b)$ and we set $\hat{\gamma} = \mathbb{E}(\gamma) = a/b$. Similarly to the negative binomial mixture of BBS, the prior variance of γ can be specified according to the user's preferences, possibly exploring different values for robustness checks.

Finally, we remark that a fully Bayesian approach might be adopted, instead of the proposed empirical Bayes one. This consists of assuming prior distributions for parameters α and θ for all

the mixtures. We exploit such a fully Bayesian approach in the real data analysis of Section S5.2 of the supplementary material, where we also show that posterior inferences obtained with the two procedures are coherent.

6.2 Model-checking

A preliminary step of our simulation studies and applications consists in the choice of the best model: either mixtures of IBPs or mixtures of BBs. The decision between the two classes pertains to the analyst. We propose two approaches to guide the selection of the best class of models: (i) a pair of visual procedures; (ii) a quantitative criterion for establishing which class of mixtures best fits the data. As for (i), the first check relies on comparing the observed values $K_1, \ldots, K_n$ with the expected values $\mathbb{E}(K_1), \ldots, \mathbb{E}(K_n)$ under different models. Since the observed values $K_1, \ldots, K_n$ refer to a particular ordering of the observations, in place of $K_1, \ldots, K_n$, we will consider the in-sample accumulation curve $K'_1, \ldots, K'_n$, obtained by averaging the number of distinct features over all possible orderings of the data. The second informal model check is based on the statistic $K_{n,r}$, i.e. the number of features observed with prevalence $r \geq 1$ in a sample of size n. To assess the model performance, we compare the observed values $K_{n,1}, \ldots, K_{n,\bar{r}}$ and the expected values $\mathbb{E}(K_{n,1}), \ldots, \mathbb{E}(K_{n,\bar{r}})$, until a certain $\bar{r} \leq n$, under different models' choices. While the empirical curves are always obtained from the data, the expected values depend on α, θ and the prior mean of N (resp. γ) if a mixture of BBs (resp. IBPs) is selected. As a consequence, if we adopt the empirical Bayes approach described in Section 6.1 for parameters elicitation, then all the mixtures of BBs (resp. IBPs) have the same rarefaction curve and the same curve $\mathbb{E}(K_{n,r})$, for $r = 1, \ldots, n$. By visual inspections, the previous model checks provide an indication of whether the mixtures of BBs or the mixtures of IBPs may be appropriate for the problem at the hand, which is the ultimate goal of our model selection.

For a quantitative comparison of the goodness-of-fit between the two classes of mixtures, we rely on the (minimum) *deviances* of the BB model and the IBP model, as representatives within the two classes. Given an observed dataset $\boldsymbol{Z}^{(n)}$ and a model described by parameters $\boldsymbol{\theta}$ (referred to as hyperparameters in our Bayesian setting), the (minimum) deviance is defined as $D(\hat{\boldsymbol{\theta}}) = -2 \log \mathcal{L}(\boldsymbol{Z}^{(n)} \mid \hat{\boldsymbol{\theta}}) = -2 \log \pi_n(m_1, \ldots, m_k \mid \hat{\boldsymbol{\theta}})$, where $\mathcal{L}$ denotes the likelihood of the model, i.e. the EFPF in our case, and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$. We suggest to compute the (minimum) deviances of the BB and the IBP model, and the one yielding the smaller deviance is selected. Notably, since both models involve the same number of hyperparameters, comparing deviances yields the same conclusions as comparisons based on standard model selection criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).

In general, we expect the conclusions drawn from the visual inspections of the curves and the quantitative information criteria to be consistent. We will show that the two criteria consistently support the same model selection decisions across all our simulation studies and real data analyses.

6.3 Overview of simulation studies

The goal of the simulation studies is to showcase the prediction abilities of our models under different experiments. We stress that the class of Gibbs-type feature models with finitely many features also allows to perform inference on the total number of features N, corresponding to the α-diversity. Specifically, in ecological applications, such quantity is referred to as taxon richness, which is a natural measure of biodiversity, as we will highlight in applications.

In Section S4 of the supplementary material, we extensively discuss three main simulation studies (A, B and C) to test the performance of our models. For each simulated dataset, we fix the parameters of the models via the empirical Bayes approach described in Section 6.1. As a second step, we apply the model-checking approaches we presented in Section 6.2 to conclude that either the mixtures of BBs or the mixtures of IBPs may be assumed to be correctly specified for the data at the hand. Experiments A and C correspond to situations where our model-checking clearly indicates that mixtures of BBs can be assumed to be correctly specified. Consequently, the assumption of finite species richness is plausible. Conversely, in experiment B, mixtures of IBPs best fit the data according to our model-checking procedures. In all the cases, we focus on the prediction of the number of unseen features in an additional sample of size m. We present the predictions obtained via the selected models and compare them with a variation of the Good–Toulmin estimators (GT)

from Chakraborty et al. (2019). Additionally, in experiments A and C, we also compare with the well-known frequentist estimator from Chao et al. (2014), that is specifically designed under the assumption of finite species richness. Our simulation studies indicate that the estimator of Chakraborty et al. (2019) exhibits poor predictive performance and stability issues as m grows. In general, our models show good predictive abilities, often outperforming the competitors. In experiments A and C, we also address the estimation of the species richness N and its uncertainty quantification. In this regard, our experiments highlight that the negative binomial mixture of BBS is usually more robust than the Poisson mixture under bad prior guesses on N.

7 Assessing diversity in ecological applications

The quantification of biological diversity is a central aspect of many ecological studies and an active research focus of ecology, due to its importance in many conservation strategies, in monitoring and management projects (Chao et al., 2014). The most commonly employed and basic metric for biodiversity in a community is undoubtedly the species richness, namely the total number of species in the assemblage. Besides, another insightful characterization of the biological diversity of the assemblage may be provided by the asymptotic growth rate of the extrapolation curve $K_m^{(n)} + k \mid \mathbf{Z}^{(n)}$, for $m = 1, 2, \ldots$, and the associated α-diversity (refer to Proposition 3 of Section 3.3). In addition, these kinds of extrapolation problems are commonly faced in order to assess whether it is worth investing additional resources in looking for possibly new species. Specifically, ecologists may be interested in how many new species they are going to observe if they sample a number m of additional plots. Based on such estimation, they might decide not to further analyze additional plots in the region if they expect to record a number of new species that are not worth the additional resources they are required to invest. Such information is naturally and straightforwardly available within our Bayesian framework, described by the posterior distribution of the statistic $K_m^{(n)}$, for $m \geq 1$.

Here, we illustrate how we address the aforementioned ecological research questions for two real-world datasets, which present different structural characteristics. We discuss the adequacy of the Poisson and negative binomial mixtures of BBS and the gamma mixture of IBPS, where the parameters are estimated via the empirical Bayes approach described in Section 6.1. Prediction and inference are then faced using the most appropriate model, selected through the model-checking described in Section 6.2. For a more exhaustive assessment of the models' predictive ability, we perform a data-holdout experiment in Section S5.1 of the supplementary material, where all the models are trained on half of the observed data, and predictions on the withheld data are compared. These analyses further support the decisions on model selection obtained through the two proposed procedures. In Section S5.2 of the supplementary material, we also report posterior inferences obtained when we adopt a fully Bayesian approach for parameters' elicitation, showing that it leads to similar results obtained via the empirical Bayes procedure described in Section 6.1.

7.1 Vascular plants in danish forest

We consider the data collected in Mazziotta et al. (2016a) concerning the forest of Lille Vildmose nature reserve in Denmark. Here, for each of the 102 forest plots object of the 2013 monitoring campaign, the species incidence (presence-absence) for four organism groups, i.e. epiphytic bryophytes, epiphytic lichens, vascular plants, and wood-inhabiting fungi, are measured. For the purpose of illustration, we focus on vascular plants, also analyzed in Mazziotta et al. (2016b), where $k = 215$ distinct species are recorded on the $n = 102$ plots. In Figure S14 of the supplementary material, we also report the taxon accumulation curve, which has clearly not yet reached convergence, thus the richness is certainly expected to be larger than the 215 observed species.

In order to assess whether mixtures of BBS (finite species richness) or mixtures of IBPS (infinite species richness) are more appropriate in this context, we rely on both the visual inspections of the model-checking tools we introduced in Section 6.2 and the quantitative assessment through the comparison of deviances. From the plots of Figure 3, we argue that the mixtures of IBPS, which assume infinitely many features, are plausible models for such data, and are definitely more suitable than mixtures of BBS. This claim is further supported by the comparison of deviances, with the BB model yielding $D(\hat{\boldsymbol{\theta}}) = 10320.1$ compared to $D(\hat{\boldsymbol{\theta}}) = 10312.4$ for the IBP model. Therefore we focus

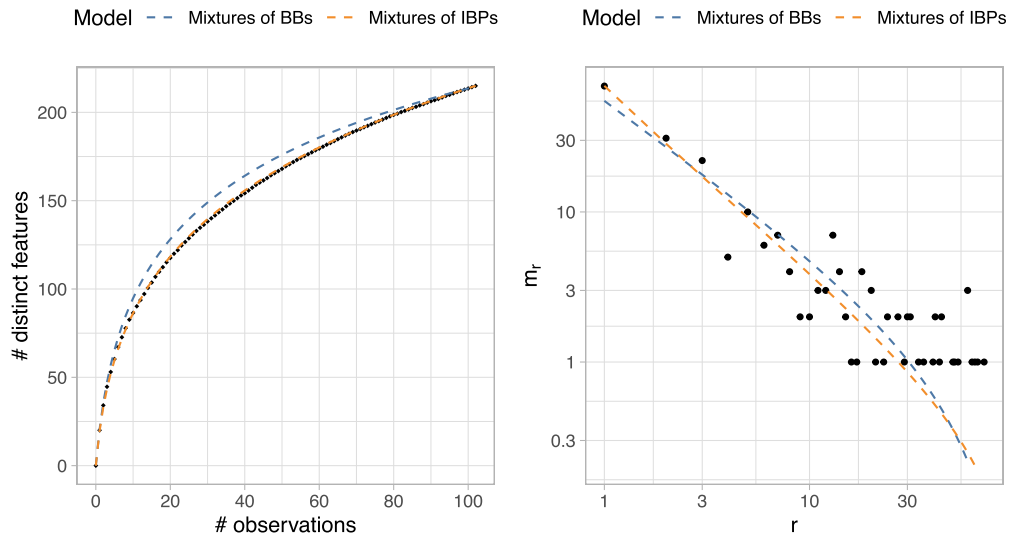

Figure 3. Left panel: the empirical accumulation curve (black dots) and the rarefaction curve of the models (blue and orange dashed lines). Right panel: the observed values of $K_{n,r}$ (black dots) compared with the expected curve $\mathbb{E}(K_{n,r})$ of the models (blue and orange dashed lines). The right plot is in log-log scale; the orange (resp. blue) curves are identical for all the mixtures of IBPs (resp. BBs).

on the gamma mixture of IBPs, and we consider two possible prior variances for γ, i.e. $\text{Var}(\gamma) \in \{1, 100\}$. From Proposition 5, the asymptotic growth rate of the curve $K_m^{(n)} + k \mid \boldsymbol{Z}^{(n)}$, for $m = 1, 2, \ldots$, is of order m^{α}, with an estimated rate of $\hat{\alpha} = 0.17$, and the posterior α-diversity S'_{α} is gamma distributed; the expected value of S'_{α} equals 186.48 for both the choices of the prior variance of γ, but we get $\text{Var}(S'_{\alpha}) = 58.3$ if $\text{Var}(\gamma) = 1$, and $\text{Var}(S'_{\alpha}) = 158.9$ if $\text{Var}(\gamma) = 100$.

The extrapolation problem is addressed in Figure 4, where we report the expected values and the 95% credible intervals for the total number of species that might be observed in m additional plots, given the observed collection of n plots, i.e. $K_m^{(n)} + k \mid \boldsymbol{Z}^{(n)}$, with $k = 215$. The posterior point estimates are similar for the two selected prior variances, while the variability increases as the prior variance of γ increases.

To provide a more quantitative answer to the extrapolation problem ecologists might be interested in, we report in Table 1 the expected values and the credible intervals for the number of new species that are going to be observed if a number m of additional plots is examined, for some values of m. Both the choices for the prior variance of γ in the gamma mixture of IBPs lead to the same point-wise estimates: if ecologists are considering whether to analyze additional plots in the region, they should expect to find 6.50 new species if they sample additional $m = 10$ plots. Differences between the two gamma mixtures are visible for $m \in \{100, 1{,}000\}$ in terms of their credible intervals.

7.2 Trees in Barro Colorado Island

As a second illustration, we analyze the presence-absence dataset of tree species in $n = 50$ plots of one hectare in Barro Colorado Island, for a total of $k = 225$ observed species. The data are publicly available in the VEGAN package in R. In terms of richness estimation, exploring the taxon accumulation curve, reported in Figure S17 of the supplementary material, we may argue that it has not reached convergence yet, though the growth is rather slow. Overall, the species richness is expected to be larger than the 225 observed species.

In order to select which model best fits the observed data, we perform the usual model-checking we introduced in Section 6.2. The visual inspection of Figure 5 suggests that the mixtures of BBs can be considered correctly specified for such data, while the mixtures of IBPs are not. This preference for the mixtures of BBs is further supported by the comparison of deviances: $D(\hat{\boldsymbol{\theta}}) = 10245.6$ for the

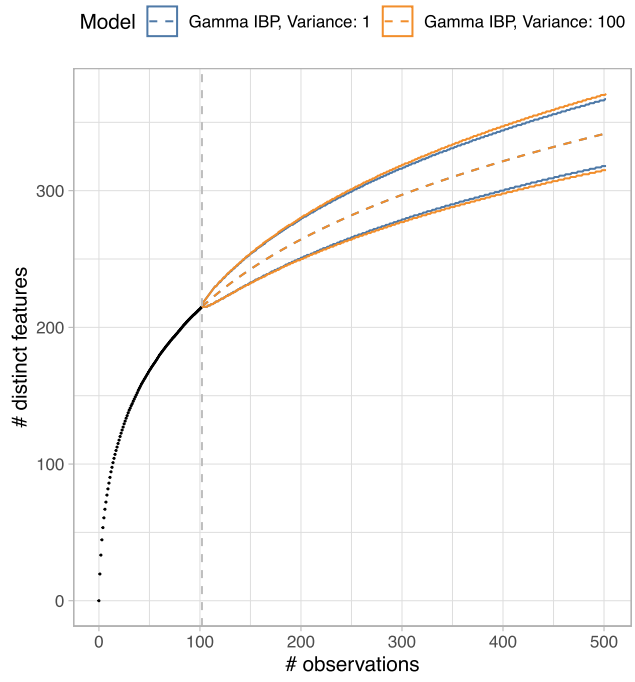

Figure 4. Expected values and the 95% credible intervals for $K_m^{(n)} + k \mid \boldsymbol{Z}^{(n)}$, with $k = 215$, for the gamma mixture of IBPS. The extrapolation horizon is $m = 1, \ldots, 400$.

Table 1. Expected values $\mathbb{E}(K_m^{(n)} \mid \boldsymbol{Z}^{(n)})$ and 95% credible intervals (in brackets) for the statistic $K_m^{(n)} \mid \boldsymbol{Z}^{(n)}$, for $m \in \{1, 10, 100, 1{,}000\}$, for the vascular plants in Mazziotta et al. (2016a)

Gamma mixture of IBPS ($n = 102$, $k = 215$)	$m = 1$	$m = 10$	$m = 100$	$m = 1{,}000$
Prior variance $\mathrm{Var}(\gamma) = 1$	0.673 [0, 3]	6.50 [2, 12]	50.1 [36, 65]	204 [172, 237]
Prior variance $\mathrm{Var}(\gamma) = 100$	0.673 [0, 3]	6.50 [2, 12]	50.1 [35, 66]	204 [166, 244]

BB model, compared to $D(\hat{\boldsymbol{\theta}}) = 10266.9$ for the IBP model. Differently from the vascular plant data analyzed in the previous section, we thus claim that it is reasonable to assume that the species richness is finite. Hence we focus on the Poisson and negative binomial mixtures of BBS, as for the latter, we analyze two choices for the prior variance of N, i.e. $\mathrm{Var}(N) = \mu_0 \times c$, for $c \in \{10, 100\}$. In such contexts, the species richness represents the most natural measure of biodiversity of the assemblage, therefore there is interest in estimating it. In the left panel of Figure 6, we report the posterior distribution of the species richness N, for the different mixtures of BBS. Specifically, the expected species richness is equal to 296.13 for the Poisson mixture of BBS, with a credible interval equal to [280, 313]. For both the negative binomial mixtures, we get an expected species richness of 296.17, with credible intervals [278, 316].

As far as the extrapolation problem is concerned, Figure 6 (right panel) reports the expected values and the 95% credible intervals for the posteriors $K_m^{(n)} + k \mid \boldsymbol{Z}^{(n)}$, for $m = 1, \ldots, 400$, where $k = 225$. It can be noted that the expected number of new features grows rather slowly with the size of the additional sample m; moreover, from Proposition 3, we remark that such a sequence $K_m^{(n)} + k \mid \boldsymbol{Z}^{(n)}$, $m = 1, 2, \ldots$, converges to the posterior distribution of the species richness N. As we have just discussed, all the mixtures of BBS that we have fitted provide an expected richness of around 296.

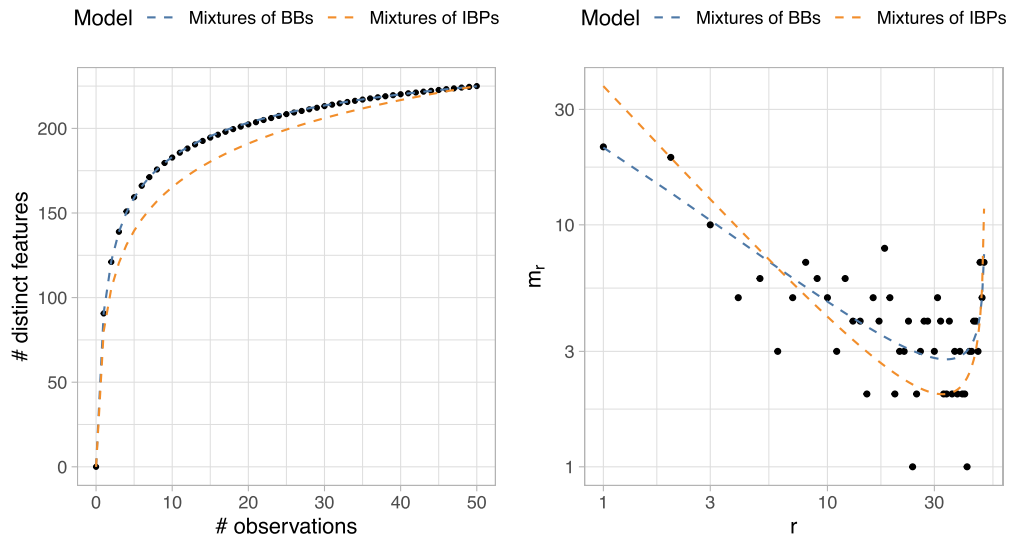

Figure 5. Left panel: the empirical accumulation curve (black dots) and the rarefaction curve of the models (blue and orange dashed lines). Right panel: the observed values of $K_{n,r}$ (black dots) compared with the expected curve $\mathbb{E}(K_{n,r})$ of the models (blue and orange dashed lines). The right plot is in log-log scale; the orange (resp. blue) curves are identical for all the mixtures of IBPs (resp. BBs).

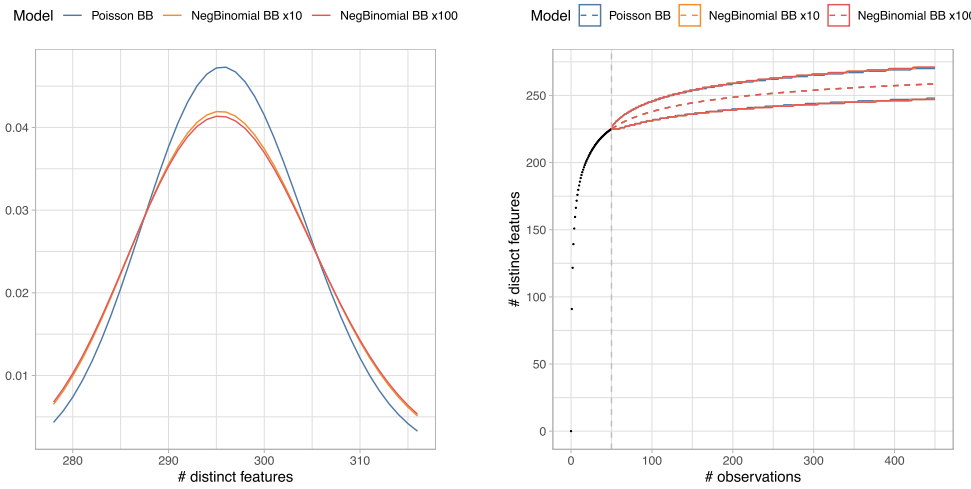

Figure 6. Left panel: posterior distributions of the species richness N for different mixtures of BBs. Right panel: expected values and 95% credible intervals for $K_m^{(n)} + k \mid \boldsymbol{Z}^{(n)}$, for different mixtures of BBs.

We finally report the numerical values of expected values and the credible intervals for the number of new species that are going to be observed if a number m of additional plots are sampled, for some values of m. All the mixtures of BBs fitted here provide the same point-wise estimates as well as the same credible intervals for $m \in \{1, 10, 100\}$: if ecologists are considering whether to analyze additional plots in the region, they should expect to find 0.41 new species if they sample additional $m = 1$ plots, 3.66 new species for $m = 10$ and 19.4 new species for $m = 100$. Moreover, the credible intervals are [0, 2] for $m = 1$, [0, 8] for $m = 10$ and [11, 29] for $m = 100$. We can spot a difference

among the models for $m = 1{,}000$: the expected number of new species is 41.8, with credible intervals equal to [30, 55] for the Poisson mixture and [29, 56] for the negative binomial mixtures.

8 Discussion

In the present paper, we analyzed Gibbs-type feature models, namely those models exhibiting an EFPF in product form (1). We argue that this class stands out among feature allocation models, similar to Gibbs-type priors, which play a fundamental role within species sampling models. We provided a comprehensive distribution theory for this class of models and a plethora of results in closed form. Additionally, we discussed two noteworthy examples: mixtures of IBPs and mixtures of BBs. While the first class assumes an infinite number of features in the population, the latter can be adopted when the total number of features is supposed to be finite. We also proposed coherent methods for parameters' elicitation and model selection. Finally, we have emphasized the importance of our findings in addressing ecological problems, such as estimating biodiversity and quantifying species richness. The code is available online at the link: https://github.com/LGhilotti/ProductFormFA.

Our contribution could pave the way for several future research directions. First, recall that we introduced a class of Gibbs-type feature models for exchangeable observations. However, in some applied problems, data are divided into different, though related, groups and the assumption of partial exchangeability would be more appropriate. Hence, an interesting direction for future research involves defining and investigating feature allocation models in the presence of multi-sample data. Although numerous models are available for partially exchangeable data within the species setting (Quintana et al., 2022), work is still ongoing in the framework of feature allocation models; see, e.g. Teh and Jordan (2010) for a few early examples. The availability of tractable classes of priors for grouped incidence data would enable the prediction for the number of shared species as well as the quantification of the so-called β-biodiversity, namely the heterogeneity among different ecological communities. Along these lines, the recent paper by Stolf and Dunson (2025) extends the IBP in defining a multivariate probit IBP.

Second, Gibbs-type are natural tools for modeling biodiversity when focussing on a single level of the Linnean taxonomy, such as *family*, *genus* or *species*. However, in modern sampling designs, each statistical unit often comprises a collection of L different taxa, which are organized in a nested fashion; see, e.g. Zito et al. (2023). As a crude exemplification, one might consider using a separate Gibbs-type feature model for each layer of the Linnean taxonomy. However, this approach would overlook the rich and informative nested structure of the data. Bayesian nonparametric models for such data are underdeveloped even for species sampling models, and there is ample room for new ideas and theoretical developments.

Third, a potentially impactful ramification of our results pertains to the usage of Gibbs-type feature models as a building block of more complex hierarchical models, i.e. when employed as a latent component. We refer to Griffiths and Ghahramani (2011) for a general discussion. Among the potential applications of Gibbs-type feature models, it is worth mentioning their role in Bayesian factor analysis, in which N, in our notation, would represent the number of factors. The IBP has been successfully used in this context to incorporate sparsity for instance to model gene expression data (Knowles & Ghahramani, 2011). A further example is given by Ayed and Caron (2021), who explored suitable extensions of the IBP to discover latent communities in network data. Our paper provides several alternatives to the IBP for Bayesian factor models and related applications. It is also worth mentioning that the mixtures of BBs, unlike the traditional IBP or the BB, enable the incorporation of prior opinions on the number of latent factors in Bayesian modeling. Work on these problems is deemed to be future research.

Acknowledgments

The authors are extremely grateful to an Associate Editor and three anonymous referees for their constructive comments and valuable suggestions, which led to a substantial improvement of the paper.

Conflicts of interest: None declared.

Funding

The authors acknowledge the support from the Italian Ministry of University and Research (MUR), "Dipartimenti di Eccellenza" grant 2023–2027. FC and LG were supported by the European Union – Next Generation EU funds, component M4C2, investment 1.1., PRIN-PNRR 2022 (P2022H5WZ9). TR gratefully acknowledges the support from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506).

Data availability

The data and code underlying the article are available at: https://github.com/LGhilotti/ProductFormFA.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Ayed F., & Caron F. (2021). Nonnegative Bayesian nonparametric factor models with completely random measures. *Statistics and Computing*, *31*(5), 31–63. https://doi.org/10.1007/s11222-021-10037-3

Battiston M., Favaro S., Roy D. M., & Teh Y. W. (2018). A characterization of product-form exchangeable feature probability functions. *The Annals of Applied Probability*, *28*(3), 1423–1448. https://doi.org/10.1214/17-AAP1333

Broderick T., Jordan M. I., & Pitman J. (2012). Beta processes, stick-breaking and power laws. *Bayesian Analysis*, *7*(2), 439–476. https://doi.org/10.1214/12-BA715

Broderick T., Mackey L., Paisley J., & Jordan M. I. (2015). Combinatorial clustering and the beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 290–306. https://doi.org/10.1109/TPAMI.2014.2318721

Broderick T., Pitman J., & Jordan M. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, *8*(4), 801–836. https://doi.org/10.1214/13-BA823

Camerlenghi F., Favaro S., Masoero L., & Broderick T. (2024). Scaled process priors for Bayesian nonparametric estimation of the unseen genetic variation. *Journal of the American Statistical Association*, *119*(545), 320–331. https://doi.org/10.1080/01621459.2022.2115918

Chakraborty S., Arora A., Begg C. B., & Shen R. (2019). Using somatic variant richness to mine signals from rare variants in the cancer genome. *Nature Communications*, *10*(1), 5506. https://doi.org/10.1038/s41467-019-13402-z

Chao A., Gotelli N. J., Hsieh T. C., Sander E. L., Ma K. H., Colwell R. K., & Ellison A. M. (2014). Rarefaction and extrapolation with hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45–67. https://doi.org/10.1890/13-0133.1

Chiu C.-H. (2022). Incidence-data-based species richness estimation via a beta-binomial model. *Methods in Ecology and Evolution*, *13*(11), 2546–2558. https://doi.org/10.1111/mee3.v13.11

Chiu C.-H. (2023). Sample coverage estimation, rarefaction, and extrapolation based on sample-based abundance data. *Ecology*, *104*(8), 1–15. https://doi.org/10.1002/ecy.v104.8

Chiu C.-H., Wang Y.-T., Walther B. A., & Chao A. (2014). An improved nonparametric lower bound of species richness via a modified good-turing frequency formula. *Biometrics*, *70*(3), 671–682. https://doi.org/10.1111/biom.v70.3

Colwell R. (2009). *Biodiversity: Concepts, patterns, and measurement* (pp. 257–263). Princeton University Press.

Colwell R. K., Chao A., Gotelli N. J., Lin S.-Y., Mao C. X., Chazdon R. L., & Longino J. T. (2012, March). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, *5*(1), 3–21. https://doi.org/10.1093/jpe/rtr044

Daley D. J., & Vere-Jones D. (2008). *An introduction to the theory of point processes*. Probability and its Applications. (Vol. *II*, 2nd ed.). Springer. General theory and structure.

De Blasi P., Favaro S., Lijoi A., Mena R. H., Prunster 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. https://doi.org/10.1109/TPAMI.2013.217

De Blasi P., Lijoi A., & Prünster I. (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica*, *23*(3), 1299–1321. https://doi.org/10.5705/ss.2012.047

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 (Methodological)*, *71*(5), 993–1008. https://doi.org/10.1111/j.1467-9868.2009.00717.x

Ferguson T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209–230. https://doi.org/10.1214/aos/1176342360
Gnedin A. (2010). A species sampling model with finitely many types. *Electronic Communications in Probability*, *15*(none), 79–88. https://doi.org/10.1214/ECP.v15-1532
Gnedin A., & Pitman J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov, POMI*, *325*, 83–102. https://doi.org/10.1007/s10958-006-0335-z
Good I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3–4), 237–264. https://doi.org/10.1093/biomet/40.3-4.237
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. https://doi.org/10.1093/biomet/43.1-2.45
Gotelli N. J., & Colwell R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, *4*(4), 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x
Gravel S. (2014). Predicting discovery rates of genomic features. *Genetics*, *197*(2), 601–610. https://doi.org/10.1534/genetics.114.162149
Gregoire G. (1984). Negative binomial distributions for point processes. *Stochastic Processes and their Applications*, *16*(2), 179–188. https://doi.org/10.1016/0304-4149(84)90018-8
Griffiths T. L., & Ghahramani Z. (2006). Infinite latent feature models and the Indian buffet process. In Y. Weiss, B. Schölkopt, & J. Platt (Eds.), *Advances in neural information processing systems* (Vol. *18*, pp. 475–482). MIT Press.
Griffiths T. L., & Ghahramani Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224. http://jmlr.org/papers/v12/griffiths11a.html
Heaukulani C., & Roy D. M. (2020). Gibbs-type Indian buffet processes. *Bayesian Analysis*, *15*(3), 683–710. https://doi.org/10.1214/19-BA1166
Hjort N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, *18*(3), 1259–1294. https://doi.org/10.1214/aos/1176347749
Ionita-Laza I., Lange C., & Laird N. M. (2009). Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, *106*(13), 5008–5013. https://doi.org/10.1073/pnas.0807815106
James L. F. (2017). Bayesian poisson calculus for latent feature modeling via generalized Indian buffet process priors. *Annals of Statistics*, *45*(5), 2016–2045. https://doi.org/10.1214/16-AOS1517
Kim Y. (1999). Nonparametric Bayesian estimators for counting processes. *Annals of Statistics*, *27*(2), 562–588. https://doi.org/10.1214/aos/1018031207
Kingman J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, *21*(1), 59–78. https://doi.org/10.2140/pjm
Knowles D., & Ghahramani Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, *5*(2B), 1534–1552. https://doi.org/10.1214/10-AOAS435
Lee J., Miscouridou X., & Caron F. (2023). A unified construction for series representations and finite approximations of completely random measures. *Bernoulli*, *29*(3), 2142–2166. https://doi.org/10.3150/22-BEJ1536
Lijoi A., Mena R. H., & Prünster I. (2007a). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, *94*(4), 769–786. https://doi.org/10.1093/biomet/asm061
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 (Methodological)*, *69*(4), 715–740. https://doi.org/10.1111/j.1467-9868.2007.00609.x
Lijoi A., & Prünster I. (2010). Models beyond the Dirichlet process. In *Bayesian nonparametrics, Volume 28 of Camb. Ser. Stat. Probab. Math* (pp. 80–136) Cambridge Univ. Press.
Magurran A. E., & McGill B. J. (2011). *Biological diversity: Frontiers in measurement and assessment*. Oxford Biology.
Masoero L., Camerlenghi F., Favaro S., & Broderick T. (2022). More for less: Predicting and maximizing genomic variant discovery via Bayesian nonparametrics. *Biometrika*, *109*(1), 17–32. https://doi.org/10.1093/biomet/asab012
Mazziotta A., Heilmann-Clausen J., Bruun H. H., Fritz O., Aude E., & Tøttrup A. P. (2016a). Dataset on species incidence, species richness and forest characteristics in a Danish protected area. *Data in Brief*, *9*, 895–897. https://doi.org/10.1016/j.dib.2016.10.021
Mazziotta A., Heilmann-Clausen J., Bruun H. H., Fritz O., Aude E., & Tøttrup A. P. (2016b). Restoring hydrology and old-growth structures in a former production forest: Modelling the long-term effects on biodiversity. *Forest Ecology and Management*, *381*, 125–133. https://doi.org/10.1016/j.foreco.2016.09.028
Miller K. T., Griffiths T. L., & Jordan M. I. (2009). Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems 22 (NIPS 2009)* (pp. 719–726). Curran Associates Inc.

Momozawa Y., & Mizukami K. (2021). Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of Human Genetics*, *66*(1), 11–23. https://doi.org/10.1038/s10038-020-00845-2

Nguyen T. D., Huggins J., Masoero L., Mackey L., & Broderick T. (2024). Independent finite approximations for Bayesian nonparametric inference. *Bayesian Analysis*, *19*(4), 1187–1224. https://doi.org/10.1214/23-BA1385

Palla K., Knowles D. A., & Ghahramani Z. (2012). An infinite latent attribute model for network data. In *Proceedings of the 29th International Coference on International Conference on Machine Learning* (pp. 395–402). Omnipress.

Pitman J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, probability and game theory, Volume 30 of IMS Lecture Notes Monogr. Ser* (pp. 245–267). Inst. Math. Statist., Hayward, CA.

Pitman J. (2003). Poisson-Kingman partitions. *Lecture Notes-Monograph Series*, *40*, 1–34. https://doi.org/10.1214/lnms/1215091133

Pitman J., & Yor M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, *25*(2), 855–900. https://doi.org/10.1214/aop/1024404422

Quintana F. A., Müller P., Jara A., & MacEachern S. N. (2022). The dependent Dirichlet process and related models. *Statistical Science*, *37*(1), 24–41. https://doi.org/10.1214/20-STS819

Sanders J. G., Nurk S., Salido R. A., Minich J., Xu Z. Z., Zhu Q., Martino C., Fedarko M., Arthur T. D., & Chen F. (2019). Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biology*, *20*(1), 1–14. https://doi.org/10.1186/s13059-019-1834-9

Stolf F., & Dunson D. B. (2025). Infinite joint species distribution models. *Biometrika*. https://doi.org/10.1093/biomet/asaf055

Teh Y. W., & Gorur D. (2009). Indian buffet processes with power-law behavior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. *22*). MIT Press.

Teh Y. W., & Jordan M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian nonparametrics, Volume 28 of Camb. Ser. Stat. Probab. Math* (pp. 158–207). Cambridge Univ. Press.

Thibaux R., & Jordan M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (Vol. *2*, pp. 564–571). PMLR.

Williamson S., Wang C., Heller K. A., & Blei D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, Madison, WI, USA* (pp. 1151–1158). Omnipress.

Zhang M. J., Ntranos V., & Tse D. (2020). Determining sequencing depth in a single-cell rna-seq experiment. *Nature Communications*, *11*, article number 774. https://doi.org/10.1038/s41467-020-14482-y

Zito A., Rigon T., & Dunson D. B. (2023). Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa. *Methods in Ecology and Evolution*, *14*(2), 529–542. https://doi.org/10.1111/mee3.v14.2

Zito A., Rigon T., & Dunson D. B. (2024). Bayesian nonparametric modeling of latent partitions via stirling-gamma priors. *Bayesian Analysis*, 1–28. https://doi.org/10.1214/24-BA1463

Zito A., Rigon T., Ovaskainen O., & Dunson D. B. (2023). Bayesian modeling of sequential discoveries. *Journal of the American Statistical Association*, *188*(544), 2521–2532. https://doi.org/10.1080/01621459.2022.2060835

Zou J., Valiant G., Valiant P., Karczewski K., Chan S. O., Samocha K., Lek M., Sunyaev S., Daly M., & MacArthur D. G. (2016). Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, *7*(1), Article 13293. https://doi.org/10.1038/ncomms13293