SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of
**Economics, Management and Statistics**

PhD program: **Economics, Statistics and Data Science** Cycle: **XXXVII**
Curriculum: **Statistics**

# Feature allocation models in Bayesian Statistics

Surname: **Ghilotti**
Name: **Lorenzo**
Registration number: **884011**

Supervisor: Prof. **Federico Camerlenghi**

Coordinator: Prof. **Matteo Manera**

**Academic Year: 2024/2025**

# Abstract

Feature allocation models have emerged as a central topic in modern Bayesian statistics, attracting increasing attention across diverse scientific fields. This thesis offers a comprehensive Bayesian investigation of the feature allocation framework, in which each observation is associated with a finite collection of features. This paradigm generalizes the classical species sampling framework, where observations belong to a single species, by allowing them to exhibit multiple features simultaneously. This distinction gives rise to new theoretical and inferential challenges concerning feature sharing and the appearance of previously unseen features as the sample size grows. The feature allocation framework is an active area of research, particularly within the machine learning community, with broad applications in diverse fields such as ecology, microbiome analysis, topic modeling, image segmentation.

After revisiting the well-established theory of species sampling, the thesis develops a parallel and extensive account of the feature allocation framework. We emphasize similarities and differences between the species and feature settings, and we also highlight several open and urgent questions within the feature framework. As a first contribution, we introduce and study *Gibbs-type feature models*, a broad class that plays for feature allocations the same role Gibbs-type priors play in the species setting, achieving a balance between flexibility and tractability. We develop a complete Bayesian analysis of this class, and we illustrate its methodological relevance through applications to biodiversity assessment in ecology. In our second contribution, we propose a unified Bayesian framework for *extended feature allocation models*, capable of capturing dependencies such as attraction or repulsion among features. Several examples are presented, extending beyond the standard feature allocation setting. Within this framework, we derive new predictive characterizations, establishing feature-based analogues of the classical *sufficientness postulates* from the species sampling literature. As a third contribution, we introduce a general class of priors for *trait allocation models* under partial exchangeability. The trait setting naturally generalizes the feature framework by associating quantitative measurements with the presence of features in each observation. The proposed prior leads to tractable posterior inference and forms the basis for a novel mixture model that enables clustering of trait allocations. The practical relevance of this approach is demonstrated through an application in the context of criminal network data. Finally, the thesis develops a novel probabilistic result on the *Palm distributions of superposed point processes*, with important statistical applications to extended feature models and beyond.

# Contents

# Summary

This thesis offers a comprehensive investigation of *feature allocation models* and their generalizations from a Bayesian perspective. Within this framework, each observation is associated with a finite collection of *features* or *characteristics*, and the primary modeling goal is to understand the underlying mechanisms that govern how features are shared among individuals and how previously unseen features appear as the sample size increases. Feature allocation models can be regarded as a natural generalization of the classical *species sampling* framework: while in species sampling models each observation is assigned to a single category (or species), in the feature allocation setting an observation may simultaneously exhibit multiple features. This distinction gives rise to a different class of models that pose new theoretical and inferential challenges, while retaining fundamental goals such as predicting the appearance of new features and describing the structure of feature sharing across observations. Research on feature allocation has grown rapidly in recent years, particularly within the machine learning community. Feature-based models now find applications across a broad range of domains, including Bayesian factor analysis and nonnegative matrix factorization, image segmentation, text mining, topic modeling, network analysis, ecology, and microbiome studies.

As with any stochastic mechanism, the modeling framework for feature allocations must reflect the assumptions underlying the data collection process. In this thesis, we explore two fundamental scenarios: data collected under *identical experimental conditions* and data collected under *different yet related experimental conditions*. The Bayesian paradigm naturally arises in both settings as a direct consequence of de Finetti representation theorems (de Finetti, 1937; Hewitt and Savage, 1955; de Finetti, 1938) for exchangeable and partially exchangeable models. In particular, when observations $Z_1, Z_2, \ldots$ are collected under identical experimental conditions, it is natural to assume that their order is irrelevant to statistical inference. In probabilistic terms, this property is formalized through *exchangeability*: a sequence of random variables $(Z_i)_{i\geq 1}$ is exchangeable if, for any $n \geq 1$ and any permutation $\sigma$ of $\{1, \ldots, n\}$,

$$(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}) \stackrel{\mathrm{d}}{=} (Z_1, \ldots, Z_n),$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution between the corresponding probability laws. The celebrated de Finetti theorem (de Finetti, 1937; Hewitt and Savage, 1955) guarantees that exchangeability is equivalent to the existence of a random probability measure $\tilde{p}$ with distribution $Q$ such that

$$Z_i \,|\, \tilde{p} \stackrel{\text{iid}}{\sim} \tilde{p}, \qquad \tilde{p} \sim Q. \tag{1}$$

This fundamental characterization theorem is regarded as one of the main motivations for the Bayesian approach, as it establishes the equivalence between exchangeability and the

existence of a *likelihood* $\tilde{p}$ and a *prior* distribution $Q$, also called the *de Finetti measure*. In general, (1) defines a Bayesian nonparametric model, where $\tilde{p} \sim Q$ is a *random probability measure*. Parametric models arise as special cases when the likelihood $\tilde{p}$ has some parametric form $f_\theta$ and the prior $Q$ is transferred to the prior on the parameter $\theta \sim \pi$.

In many applications, however, data are collected under *different yet related experimental conditions*. For instance, consider measurements $Z_{iq}$ obtained from patients in a control group ($q = 1$) and a treatment group ($q = 2$). Such settings exhibit heterogeneity while offering opportunities for information sharing, that can be exploited through the notion of *partial exchangeability* (de Finetti, 1938). This assumes that the order of the observations within each group is irrelevant, while group membership itself remains essential. Formally, two sequences of random variables $(Z_{i1})_{i\geq 1}$ and $(Z_{i2})_{i\geq 1}$ are said to be partially exchangeable if, for any $n, m \geq 1$ and any permutations $\sigma_1, \sigma_2$, we have

$$((Z_{i1})_{i=1}^n, (Z_{i2})_{i=1}^m) \overset{\mathrm{d}}{=} ((Z_{\sigma_1(i)1})_{i=1}^n, (Z_{\sigma_2(i)2})_{i=1}^m).$$

A generalized version of de Finetti's theorem (de Finetti, 1938) then ensures that partial exchangeability is equivalent to the existence of a probability distribution $Q$ such that, for any $i \geq 1$,

$$(Z_{i1}, Z_{i2}) \,|\, \tilde{p}_1, \tilde{p}_2 \overset{\text{iid}}{\sim} \tilde{p}_1 \times \tilde{p}_2, \qquad (\tilde{p}_1, \tilde{p}_2) \sim Q. \tag{2}$$

Analogously to the exchangeable case and its representation in (1), partial exchangeability admits a natural Bayesian interpretation: conditionally on the random measures $(\tilde{p}_1, \tilde{p}_2)$, the observations are independent, and identically distributed within each group. The distribution $Q$ acts as a joint prior on $(\tilde{p}_1, \tilde{p}_2)$, thereby capturing the dependence structure across groups. Exchangeability is recovered as a special case when $\tilde{p}_1 = \tilde{p}_2$ almost surely. The notion of partial exchangeability extends naturally to $d$ groups, leading to a generalized version of de Finetti theorem in (2), where the representation involves a vector of $d$ random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_d)$.

The thesis is organized into six chapters, summarized below.

Chapter 1 provides a thorough review of the *species sampling* framework, a natural and conceptually related counterpart to the framework of *feature allocations*. As one of the most extensively studied areas in Bayesian nonparametrics, the species sampling setting offers a clear perspective on key inferential goals and the methodological strategies developed to address them. This framework later serves as a foundation for understanding how related problems emerge and are reformulated within the feature allocation setting, a fundamental yet more recent area of research, where several open questions are addressed and resolved in this thesis.

Chapter 2 presents a comprehensive review of the central theme of this thesis: the *feature allocation* framework and its generalizations. The chapter is structured in parallel with Chapter 1 to highlight both the connections and the differences between the two settings. While primarily introductory, this chapter establishes the theoretical and methodological foundations for the original contributions developed in the remainder of the thesis, and highlights important gaps and open challenges in the feature allocation literature.

Chapter 3 develops a general investigation for the class of *Gibbs-type feature models* (Battiston et al., 2018), which play for feature allocations the same role Gibbs-type priors play in species sampling models, offering an optimal balance between flexibility and tractability. We derive closed form expressions for the marginal, posterior, and predictive distributions that hold for any Gibbs-type feature model, enabling efficient inference and interpretability. In addition, we introduce three new and analytically tractable feature allocation models. An ecological application illustrates their practical relevance for biodiversity assessment. This chapter is based on the published paper: Lorenzo Ghilotti, Federico Camerlenghi, and Tommaso Rigon (2025). "Bayesian analysis of product feature allocation models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Chapter 4 introduces a unified Bayesian framework for *extended feature allocation models*, capable of capturing interactions such as repulsion or attraction among features and their associated probabilities. We provide a full Bayesian analysis of these models, specialize our general theory to noteworthy classes of priors, and characterize priors satisfying predictive sufficientness postulates analogous to those studied in the species sampling literature since W.E. Johnson's work on the Dirichlet distribution. This chapter is based on the preprint: Mario Beraha, Federico Camerlenghi, and Lorenzo Ghilotti (2025+). "Bayesian calculus and predictive characterizations of extended feature allocation models". In: *arXiv:2502.10257*.

Chapter 5 concerns *trait allocation models*, a natural generalization of feature allocation models. We propose a general and tractable class of Bayesian nonparametric priors for partially exchangeable trait allocations based on completely random vectors. We provide a comprehensive Bayesian analysis, deriving marginal, posterior, and predictive distributions in closed form, which enable fast and efficient posterior inference. Building on these results, we develop a mixture model for clustering trait allocations. An application to the 'Ndrangheta criminal network (*Operazione Infinito*) illustrates and motivates the methodology. This chapter is based on the preprint: Lorenzo Ghilotti, Federico Camerlenghi, Tommaso Rigon, and Michele Guindani (2025+). "Bayesian nonparametric modeling of multivariate count data with an unknown number of traits". In: *arXiv:2510.24526v2*.

Chapter 6 develops a novel probabilistic result in point process theory, characterizing the Palm distributions of superposed independent point processes. The statistical implications of this result are explored through several applications, including one that connects to feature allocation models. In this example, the result enables inference for a specific instance of an *extended feature model* within the class introduced in Chapter 4. Unlike the preceding chapters, the primary focus of Chapter 6 is the novel result in the point process literature, rather than the feature allocation models themselves. This chapter builds on the preprint: Mario Beraha, Federico Camerlenghi, and Lorenzo Ghilotti (2025+). "Palm distributions of superposed point processes for statistical inference". In: *arXiv:2508.20924*.

# 1. The Species Sampling Framework

This chapter reviews the *species sampling* framework, a cornerstone of Bayesian nonparametrics. Although this is not the main focus of the thesis, it provides a natural and instructive parallel to the framework of *feature allocations*, which constitutes the central topic of our investigation. The species framework has been extensively studied and represents one of the most established areas in Bayesian nonparametrics. By first reviewing this setting, we highlight both the key inferential goals it addresses and the methodological solutions developed to overcome them. This framework will later serve as a foundation for understanding how analogous problems arise and are reformulated in the feature allocation setting, a more recent and less explored domain, where several key questions remain open and are resolved in this thesis.

In the species sampling setting, each individual in a population is assigned to a single species within a (potentially infinite) collection. The associated inferential tasks, collectively referred to as *species sampling problems*, first appeared in ecology for the estimation of the species richness or diversity of ecological populations. The classical problem of estimating the number of unobserved species dates back to the seminal work of Fisher et al. (1943). Over time, these ideas have found applications well beyond ecology, spanning the biological and physical sciences, statistical machine learning, electrical engineering, theoretical computer science, information theory, and forensic statistics. In particular, large-scale genomic data have provided a fertile ground for species sampling problems, as testified by the work of Deng et al. (2019), highlighting the continued relevance of these models in modern data analysis.

A Bayesian nonparametric approach to these problems was formalized by Lijoi et al. (2007), providing a flexible framework for inference. Beyond their original ecological motivation, species sampling models form one of the fundamental building blocks of Bayesian nonparametric statistics, owing to their intimate connection with mixture models. Since their introduction by Ferguson (1983) and Lo (1984), and their popularization through Escobar and West (1995), species-based mixture models have become central tools for model-based clustering and density estimation.

The remainder of this chapter is organized as follows. In Section 1.1, we introduce the species sampling framework and presents the three fundamental perspectives used to describe these models. Section 1.2 examines the class of Gibbs-type priors, the most prominent family of species sampling models, which achieve a practical balance between flexibility and analytical tractability. In Section 1.3, we discuss predictive characterizations, also known as *sufficientness postulates*, which play a key role in Bayesian inference for species models. Section 1.4 illustrates the use of species sampling priors as latent structures in mixture models, one of their most influential and widely applied extensions. Finally,

Section 1.5 extends the discussion to partially exchangeable species models, relevant when observations are naturally divided into multiple groups.

## 1.1 Species sampling: foundations and modeling perspectives

In the context of *species sampling*, consider a population of subjects of various *species* and a random sample $(Z_1, \dots, Z_n)$ is drawn from the population, with $Z_i \in \mathbb{X}$ representing the species of the $i$th subject sampled. The space $\mathbb{X}$, assumed Polish for later rigorous treatment, should in this setting be thought of as an arbitrary set of tags used to label the various species. Fundamental consideration assumes that the subjects are collected under the same experimental condition, so that their order is irrelevant for statistical analysis, therefore it calls for *exchangeability* among the $Z_i$'s. Given a sample of observations $(Z_1, \dots, Z_n)$, there are some major and popular questions of interest that arise, referred to as *species sampling problems*. The most popular problem is the one called *unseen species* problem, which refers to the estimation of

$$K_m^{(n)} := |\{Z_{n+1}, \dots, Z_{n+m}\} \setminus \{Z_1, \dots, Z_n\}|, \tag{1.1}$$

namely the number of hitherto unseen (distinct) species that would be observed if $m \geq 1$ additional subjects $(Z_{n+1}, \dots, Z_{n+m})$ were collected from the same distribution. The estimation of the number of unseen species is a classical problem in statistics, dating back to the seminal work of Fisher et al. (1943). The celebrated Good-Turing estimator (Good, 1953) provides an estimate of the probability of discovering at the $(n+1)$th draw a species not observed in the sample $(Z_1, \dots, Z_n)$. The Good-Toulmin estimator (Good and Toulmin, 1956) represents a $m$-steps ahead generalization for the probability of discovering a new species. The Bayesian estimator for the unseen species problem, under the class of Gibbs-type priors (Section 1.2), is derived in Lijoi et al. (2007). The unseen species problem has a long history and several applications; see, for example, Good and Toulmin (1956); Orlitsky et al. (2016) for frequentist contributions and Favaro et al. (2009, 2012, 2016) for Bayesian contributions. Other problems relate, for example, to (i) the expected population frequency of a species with frequency $r \geq 1$ in the sample and (ii) the number of species with frequency $r \geq 1$ in the sample that will be observed in additional samples. See Balocchi et al. (2024) for a complete and thorough review on species sampling problems.

In defining the probabilistic model for an exchangeable sequence $(Z_i)_{i\geq 1}$ in the *species sampling* setting, de Finetti theorem implies that such a model must take the form (1). In the implied Bayesian formulation of (1), it is natural to employ a discrete random probability measure $\tilde{p}$ to account for the ties among the $Z_i$'s. The Bayesian nonparametric literature, in particular, has focused on the broad class of (proper) *species sampling models* (Pitman, 1996), defined next.

**Definition 1** (Species sampling model)**.** *A (proper) species sampling model (SSM) is a pair of a sequence of random variables $(Z_i)_{i\geq 1}$ and a random probability measure $\tilde{p}$ such that*

$$Z_i \,|\, \tilde{p} \overset{iid}{\sim} \tilde{p}, \qquad \tilde{p} = \sum_{j\geq 1} W_j \delta_{\tilde{X}_j}, \tag{1.2}$$

*for* $\tilde{X}_j \overset{iid}{\sim} G_0$*, where* $G_0$ *is a diffuse probability distribution on* $\mathbb{X}$*, and an independent random probability sequence* $(W_j)_{j\geq 1}$*, i.e.,* $\sum_{j\geq 1} W_j = 1$*. The random probability measure* $\tilde{p}$ *in a* SSM *is called a species sampling process (*SSP*).*

For SSMs in (1.2), $W_j$ is interpreted as the relative frequency of the $j$th species, identified by label $\tilde{X}_j$, in the population. It is worth remarking that, in the context of species sampling problems, the assumption that the species labels $\tilde{X}_j$'s are i.i.d. from $G_0$ is not restrictive, since they just serve to tag the various species, but the values themselves are not relevant, given that they are almost surely distinct.

Most of the species sampling problems deal with *prediction*. For example, in the most popular unseen species problem, a sample $(Z_1, \ldots, Z_n)$ is observed, and the estimation of $K_m^{(n)}$ in (1.1) relates to the *prediction* of a future sample $(Z_{n+1}, \ldots, Z_{n+m})$, for $m \geq 1$. Prediction problems ultimately boil down to the prediction of the species associated to next subject $Z_{n+1}$ given $(Z_1, \ldots, Z_n)$, for any $n$. One advantage of the class of SSMs in (1.2) is that the predictive distributions take a simple form. For a sample $(Z_1, \ldots, Z_n)$, let $K_n = k$ denote the number of distinct observed species, with labels $X_\ell$, as $\ell = 1, \ldots, k$. Moreover, let $\boldsymbol{n} := (n_1, \ldots, n_k)$, where $n_\ell$ is the multiplicity of $X_\ell$ in $(Z_1, \ldots, Z_n)$, as $\ell = 1, \ldots, k$. Then, the predictive distributions of a SSM in (1.2) take the form $Z_1 \sim G_0$ and, for $n \geq 1$,

$$Z_{n+1} \,|\, Z_1, \ldots, Z_n \sim \sum_{\ell=1}^{k} p_\ell(\boldsymbol{n})\delta_{X_\ell} + p_{k+1}(\boldsymbol{n})G_0, \tag{1.3}$$

for some collection of functions $p_\ell : \cup_{n\geq 1}\mathcal{C}_n \to [0,1]$, with $\mathcal{C}_n$ denoting the set of all compositions of $n$. Lemma 1.2 will characterize the collection of functions $p_\ell$'s.

The predictive rule in (1.3) is easily interpretable: the species for the next subject will be *new*, i.e., hitherto unseen, with probability $p_{k+1}(\boldsymbol{n})$, or it will coincide with the $\ell$th *old* species, i.e., species $X_\ell$ which has been already observed in the sample, with probability $p_\ell(\boldsymbol{n})$. In general, prediction in SSMs depends on the observed sample through the number of distinct observed species $k$ and the vector of multiplicities $\boldsymbol{n}$, as a consequence of exchangeability. Clearly, the specific dependence of $p_\ell$ on $k$ and $\boldsymbol{n}$ is determined by the prior on $\tilde{p}$ in (1.2). Remarkably, classes of priors in SSMs can be characterized by the specific properties of such a predictive rule in (1.3): we devote Section 1.3 to this discussion.

As it should be clear from the previous introduction on SSMs, the species sampling setting intrinsically aims at the collection of species under a sequential procedure. After sampling $n$ subjects, the information in the sample can be described by the number $K_n = k$ of distinct observed species and the subgroups of subjects belonging to the same species. That is, the essential information in the sample is the partition of the (first) $n$ observed subjects. As evident, the use of species labels $Z_i$'s in SSMs, see (1.2), represents a convenient modeling trick for describing the species collection. Indeed, the presence of ties in the sample $(Z_1, \ldots, Z_n)$ induces the partition of the $n$ subjects $\mathcal{P}_n = \{C_1, \ldots, C_k\}$, where $C_\ell = \{i : Z_i = X_\ell\}$, and the distinct species labels $X_\ell$'s associated to the groups of the partition. In this alternative representation through partitions, $K_n = k$ corresponds to the number of clusters in the partition, and the belonging of a subject to a species translates into the belonging of a subject to a cluster. Strictly related to the species sampling context,

the labels $X_\ell$'s are of no interest, while key targeted quantity is the partition $\mathcal{P}_n$, with special emphasis on the number of clusters $K_n$. However, when SSMs are used outside the species sampling context, for example in clustering problems, then the labels $X_\ell$'s become of interest. We defer the discussion of this scenario to Section 1.4.

In terms of probability distribution induced on $\mathcal{P}_n = \{C_1, \ldots, C_k\}$, $n \geq 1$, by a SSM, exchangeability of the $Z_i$'s yields a law that depends on $\{n_1, \ldots, n_k\}$, where $n_\ell = \#C_\ell$ is the cardinality of $C_\ell$, i.e., the multiplicity of the $\ell$th species. In particular, for any SSM, there exists a symmetric function $\pi : \cup_{n\geq 1}\mathcal{C}_n \to [0,1]$ such that, for every $n$ and every partition $\{C_1, \ldots, C_k\}$ of $\{1, \ldots, n\}$,

$$\mathbb{P}(\mathcal{P}_n = \{C_1, \ldots, C_k\}) = \pi(n_1, \ldots, n_k). \tag{1.4}$$

The function $\pi$ is called the *exchangeable partition probability function* (EPPF). EPPFs are related to *infinite exchangeable random partition*s, which are now defined for completeness.

**Definition 2** (Exchangeable random partition)**.** *A random partition $\mathcal{P}_n$ of $\{1, \ldots, n\}$ is said to be exchangeable if, for any permutation $\sigma$ of $\{1, \ldots, n\}$ and any partition $\{C_1, \ldots, C_k\}$, it holds that $\mathbb{P}(\mathcal{P}_n = \{C_1, \ldots, C_k\}) = \mathbb{P}(\mathcal{P}_n = \{\sigma(C_1), \ldots, \sigma(C_k)\})$. This is equivalent to saying that there exists a symmetric function $\pi_n : \mathcal{C}_n \to [0,1]$ such that, for every partition $\{C_1, \ldots, C_k\}$ of $\{1, \ldots, n\}$,*

$$\mathbb{P}(\mathcal{P}_n = \{C_1, \ldots, C_k\}) = \pi_n(n_1, \ldots, n_k).$$

*The function $\pi_n$ is called the EPPF of $\mathcal{P}_n$.*

*An infinite exchangeable random partition is a sequence $(\mathcal{P}_n)_{n\geq 1}$ of exchangeable random partitions of $\{1, \ldots, n\}$ which is consistent. Consistent means that $\mathcal{P}_{n-1}$ is equal to the partition obtained from $\mathcal{P}_n$ by removing the element $n$, almost surely, for every $n$. In this case, the function $\pi : \cup_{n\geq 1}\mathcal{C}_n \to [0,1]$ such that the restriction of $\pi$ to $\mathcal{C}_n$ is equal to $\pi_n$, the EPPF of $\mathcal{P}_n$, is called the EPPF of the infinite exchangeable random partition.*

A pivotal notion for infinite exchangeable random partitions is the *predictive probability function*. Let $(\mathcal{P}_n)_{n\geq 1}$ be an *infinite exchangeable random partition*, whose EPPF is indicated with $\pi$. Define the functions $p_\ell : \cup_{n\geq 1}\mathcal{C}_n \to [0,1]$ as

$$p_\ell(\boldsymbol{n}) = \frac{\pi(\boldsymbol{n}^{\ell+})}{\pi(\boldsymbol{n})}, \quad \ell = 1, \ldots, k+1, \tag{1.5}$$

for $\boldsymbol{n} = (n_1, \ldots, n_k)$, where $\boldsymbol{n}^{\ell+} = (n_1, \ldots, n_{\ell-1}, n_\ell + 1, n_{\ell+1}, \ldots, n_k)$, $\ell = 1, \ldots, k$, and $\boldsymbol{n}^{(k+1)+} = (n_1, \ldots, n_k, 1)$. The collection of functions $p_\ell$'s in (1.5) is called the *predictive probability function* (PPF) of the infinite exchangeable random partition. In particular, $p_\ell(\boldsymbol{n})$, such that $\sum_{\ell=1}^k n_\ell = n$, represents the probability that the $(n+1)$th subject is added to the $\ell$th group of the partition $\mathcal{P}_n$, for $\ell = 1, \ldots, k$, and $p_{k+1}(\boldsymbol{n})$ is the probability that the $(n+1)$th subject forms a new group. Some characterizing conditions for the collection of functions $p_\ell$'s to be the PPF of an infinite exchangeable random partition are available (Ghosal and van der Vaart, 2017, Lemma 14.8). We report them next.

**Lemma 1.1** (Conditions for the PPF of an infinite exchangeable random partition)**.** *A collection of functions $p_\ell$'s is the PPF of an infinite exchangeable random partition if and only if:*

*(i)* *for every composition* $\boldsymbol{n} = (n_1, \dots, n_k)$, $(p_1(\boldsymbol{n}), \dots, p_{k+1}(\boldsymbol{n}))$ *is a probability vector;*

*(ii)* *for all* $i, j \geq 1$, *it holds* $p_i(\boldsymbol{n}) p_j(\boldsymbol{n}^{i+}) = p_j(\boldsymbol{n}) p_i(\boldsymbol{n}^{j+})$;

*(iii)* *for every permutation* $\sigma$ *of the first* $k$ *natural numbers and* $i = 1, \dots, k$, *it holds* $p_i(n_1, \dots, n_k) = p_{\sigma^{-1}(i)}(n_{\sigma(1)}, \dots, n_{\sigma(k)})$.

After having introduced exchangeable random partitions and PPFs, we can now discuss the connection with SSMs. First, all SSMs induce probability distributions on $(\mathcal{P}_n)_{n\geq 1}$ which are *infinite exchangeable random partitions*, by (1.4). Interestingly, any *infinite exchangeable random partition* $(\mathcal{P}_n)_{n\geq 1}$, whose EPPF is indicated with $\pi$, can be induced by a SSM in (1.2). This is shown in (Ghosal and van der Vaart, 2017, Lemma 14.11), which provides a characterization between the predictive distributions of a SSM in (1.3) and the PPF of the induced infinite exchangeable random partition.

**Lemma 1.2** (Predictive distributions of a SSM)**.** *The predictive distributions of a SSM in* (1.2) *take the form* $Z_1 \sim G_0$ *and, for* $n \geq 1$,

$$Z_{n+1} \,|\, Z_1, \dots, Z_n \sim \sum_{\ell=1}^{k} p_\ell(\boldsymbol{n}) \delta_{X_\ell} + p_{k+1}(\boldsymbol{n}) G_0,$$

*where the collection of functions* $p_\ell : \cup_{n\geq 1} \mathcal{C}_n \to [0, 1]$ *are the PPF of an infinite exchangeable random partition. Notably, by assigning the predictive distribution as in* (1.3)*, such that the* $p_\ell$*'s are a PPF and* $G_0$ *is a diffuse probability measure, then the sequence* $(Z_i)_{i\geq 1}$ *follows a SSM in* (1.2) *and the induced exchangeable random partition has PPF equal to* $p_\ell$*'s.*

While the first part of Lemma 1.2 offers a more detailed description of the predictive distributions (1.3) of a SSM, the second part guarantees that a valid SSM can indeed be defined through such predictive distributions, so that it induces an *infinite exchangeable random partition* with EPPF equal to $\pi$.

To conclude this general introduction to the species sampling framework, we summarize the preceding discussion with the following remark.

**Remark 1.1.** *Any SSM can be equivalently characterized through any of the following three pairs of objects:*

*(i)* *the prior on* $\tilde{p}$ *in* (1.2)*, i.e.,* $G_0$ *and the distribution of* $(W_j)_{j\geq 1}$;

*(ii)* *the predictive distributions in Lemma 1.2, i.e.,* $G_0$ *and the PPF* $p_\ell$*'s;*

*(iii)* *the infinite exchangeable random partition* $(\mathcal{P}_n)_{n\geq 1}$ *and the law of the distinct labels* $X_\ell$*'s, i.e.,* $G_0$ *and the EPPF* $\pi$.

*For a comprehensive review and details on the proofs, refer to (Ghosal and van der Vaart, 2017, Section 14.2). It is worth emphasizing that each of the three perspectives highlights different aspects of the models and can be particularly convenient for describing them. Furthermore, depending on the context, one perspective may provide a more natural or straightforward route to defining valid models.*

The class of SSMs has been defined in Pitman (1996), but the statistical (Bayesian nonparametric) approach to species sampling problems through SSMs has been pioneered by Lijoi et al. (2007). In particular, Lijoi et al. (2007) investigate prediction and inference for SSMs under the class of Gibbs-type priors (Gnedin and Pitman, 2005; De Blasi et al., 2015) for $\tilde{p}$, which encompasses notable examples such as the Dirichlet process (Ferguson, 1973) and the two parameters Poisson-Dirichlet model (Pitman, 1995), also known as Pitman-Yor process. In the next section, we provide a brief account on Gibbs-type priors and notable examples.

## 1.2 Gibbs-type species sampling models

Within the broad family of models for the species sampling framework, the Gibbs-type class stands out as the most widely studied, offering a convenient balance between analytical tractability and modeling flexibility. The most natural way to introduce this class is through perspective (iii) of Remark 1.1, namely through infinite exchangeable random partitions whose EPPF depends multiplicatively on the sizes of the partition blocks, as

$$\pi(n_1, \ldots, n_k) = V_{n,k} \prod_{\ell=1}^{k} \rho(n_\ell),$$

for any $n$, $(n_1, \ldots, n_k) \in \mathcal{C}_n$, where $V_{n,k}$ is an array of constants and $\rho$ is a strictly positive, non-constant function. These are called *product partition* models, where the multiplicative form of the EPPF is attractive for its simplicity. By (Ghosal and van der Vaart, 2017, Lemma 14.21), the array $V_{n,k}$ and the function $\rho$ need to satisfy specific requirements (in order not to induce trivial partition models), yielding the class of *product partition* models to coincide with the class of *Gibbs-type partition* model, defined next.

**Definition 3** (Gibbs-type model)**.** *A Gibbs-type random partition (or Gibbs-type partition model) of parameter $\sigma \in (-\infty, 1)$ is an infinite exchangeable random partition with EPPF of the form*

$$\pi(n_1, \ldots, n_k) = V_{n,k} \prod_{\ell=1}^{k} (1-\sigma)_{n_\ell - 1},$$

*where $(x)_n := \Gamma(x+n)/\Gamma(x)$ is the Pochammer symbol and $(V_{n,k})_{n,k}$, $n, k \geq 1$, is an array of non-negative numbers satisfying the recursive relation $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}$, with $V_{1,1} = 1$.*

For a given parameter $\sigma \in (-\infty, 1)$, multiple arrays $V_{n,k}$ satisfy the recursive equation, and describe different Gibbs-type random partitions. By pairing any Gibbs-type random partition with a diffuse probability measure $G_0$, one obtains a SSM, as discussed in point (iii) of Remark 1.1.

**Definition 4** (Gibbs-type species sampling model)**.** *A Gibbs-type SSM is a pair of a Gibbs-type random partition and a diffuse probability measure $G_0$. In the formulation (1.2), the random probability measure $\tilde{p}$ is said be a Gibbs-type process.*

In presenting the most notable examples within the Gibbs-type class, it is common to adopt the de Finetti's representation of SSMs, corresponding to point (i) of Remark

1.1, even if the diffuse measure $G_0$ is not of primary interest. From this perspective, we now recall two prominent examples of Gibbs-type processes, which correspond to popular Bayesian nonparametric priors for discrete random probability measures $\tilde{p}$.

**Example 1.1** (Dirichlet process (Ferguson, 1973))**.** A Dirichlet process of parameter $M > 0$, compactly written as $\text{DP}(M)$, is a Gibbs-type process obtained from a Gibbs-type random partition of parameter $\sigma = 0$ and

$$V_{n,k} = \frac{M^k}{(M)_n}.$$

In particular, the Gibbs-type random partition is called *Chinese restaurant* process, also well-known as *Ewens* model.

**Example 1.2** (Pitman-Yor process (Pitman and Yor, 1997))**.** A Pitman-Yor process (or two parameter Poisson-Dirichlet distribution) of parameters $(\sigma, M)$, compactly written as $\text{PY}(\sigma, M)$, is a Gibbs-type process obtained from a Gibbs-type random partition of parameter $\sigma$ and

$$V_{n,k} = \frac{\prod_{\ell=1}^{k-1}(M + \ell\sigma)}{(M+1)_{n-1}}.$$

The parameters $(\sigma, M)$ are such that either (i) $\sigma < 0$ and $M \in \{-2\sigma, -3\sigma, \ldots\}$, or (ii) $\sigma \in [0, 1)$ and $M > -\sigma$. Special cases are included.

(i) For $\sigma = 0$, the Dirichlet process is recovered, i.e., $\text{PY}(0, M) \overset{\text{d}}{=} \text{DP}(M)$.

(ii) For $\sigma < 0$, the process is known as the *Fisher model*. For $M = -m\sigma$, the prior on $\tilde{p}$ is such that $\tilde{p}$ has a finite number of atoms, equal to $m$, and $(W_1, \ldots, W_m)$ follows a Dirichlet distribution of parameters $(-\sigma, \ldots, -\sigma)$.

### 1.2.1 Predictive distributions of Gibbs-type models

In light of point (ii) of Remark 1.1, the predictive distributions of a SSM are characterized by the PPF of the associated partition model. In particular, the predictive distributions for a *Gibbs-type species sampling model* of parameter $\sigma \in (-\infty, 1)$ are in the form of (1.3),

$$Z_{n+1} \,|\, Z_1, \ldots, Z_n \sim \sum_{\ell=1}^{k} p_\ell(\boldsymbol{n})\delta_{X_\ell} + p_{k+1}(\boldsymbol{n})G_0,$$

with

$$p_\ell(n_1, \ldots, n_k) = \begin{cases} \dfrac{V_{n+1,k}}{V_{n,k}}(n_\ell - \sigma), & \ell = 1, \ldots, k; \\ \dfrac{V_{n+1,k+1}}{V_{n,k}}, & \ell = k+1. \end{cases} \tag{1.6}$$

This class of predictive rules shows some notable features: after having sampled $n$ subjects, where $k$ distinct species have been recorded, the probability to record a *new* species for the next subject corresponds to $V_{n+1,k+1}/V_{n,k}$. Therefore, in general, the probability of discovering a new species depends on the observed sample through the sample size $n$ and the number of distinct species $k$, but not on the multiplicities $(n_1, \ldots, n_k)$. Moreover, the probability that the next subject belongs to the $\ell$th *old* species depends on the observed sample through $n$, $k$ and also the multiplicity $n_\ell$ of the $\ell$th species. In particular, this is proportional to the multiplicity of the species, adjusted by $\sigma$, which highlights the reinforcement mechanism typical of Gibbs-type processes.

**Example 1.3** (Predictive rules for Dirichlet and Pitman-Yor processes)**.** For the Dirichlet process of parameter $M$, the predictive rule determined by (1.6) specializes as

$$p_\ell(n_1,\dots,n_k) = \begin{cases} n_\ell/(M+n), & \ell = 1,\dots,k; \\ M/(M+n), & \ell = k+1. \end{cases} \tag{1.7}$$

According to this predictive rule, the probability of discovering a new species depends on the observed sample only through the sample size $n$. Thus, the Dirichlet process displays a very poor predictive rule for the discoveries, which basically ignores the information in the initial sample, except for the sample size.

For the Pitman-Yor process of parameter $(\sigma, M)$, the predictive rule specializes as

$$p_\ell(n_1,\dots,n_k) = \begin{cases} (n_\ell - \sigma)/(M+n), & \ell = 1,\dots,k; \\ (M+k\sigma)/(M+n), & \ell = k+1. \end{cases} \tag{1.8}$$

According to this predictive rule, the probability of discovering a new species depends on the observed sample through both the sample size $n$ and the number of distinct species $k$. Given the role of $\sigma$ in the first line of (1.8), it is also referred to as the *discount* parameter of the Pitman-Yor process.

### 1.2.2 Unseen species problems via Gibbs-type models

In the study of SSMs, particular interest is given to the number of distinct species observed, as this serves as a natural measure of biodiversity in ecological applications. For a Gibbs-type SSM, the distribution of the number $K_n$ of distinct species in a sample of size $n$ is provided by Gnedin and Pitman (2005); De Blasi et al. (2015),

$$\mathbb{P}(K_n = k) = \frac{V_{n,k}}{\sigma^k}\,\mathscr{C}(n,k;\sigma), \tag{1.9}$$

for $k = 0,\dots,n$, where $\mathscr{C}(n,k;\sigma)$ is the generalized factorial coefficient, which can be represented as

$$\mathscr{C}(n,k;\sigma) = \frac{1}{k!}\sum_{\ell=1}^{k}(-1)^\ell \binom{k}{\ell}(-\ell\sigma)_n,$$

see Charalambides and Singh (1988); Charalambides (2005). The distribution in (1.9) represents the prior distribution on the number of distinct species that are to be observed in a sample of size $n$.

The much more compelling *unseen species* problem wants to target a similar quantity but in a predictive perspective. In particular, the goal is the estimation of $K_m^{(n)}$ defined in (1.1). Specifically, suppose to collect a sample $(Z_1,\dots,Z_n)$ of species, where $K_n = k$ distinct species have been recorded. The *unseen species* problem poses the question of predicting the number $K_m^{(n)}$ of *new* species which will be observed in an additional sample $(Z_{n+1},\dots,Z_{n+m})$ of size $m$, for $m \geq 1$. Lijoi et al. (2007) characterize the distribution of $K_m^{(n)}$ for any Gibbs-type SSM, as

$$\mathbb{P}(K_m^{(n)} = y \,|\, Z_1,\dots,Z_n) = \frac{V_{n+m,k+y}}{V_{n,k}}\,\frac{1}{\sigma^y}\,\mathscr{C}(m,y;\sigma,-n+k\sigma), \tag{1.10}$$

for $y = 0, \ldots, m$, where $\mathscr{C}(m, y; \sigma, -n + k\sigma)$ is the non-central generalized factorial coefficient, which can be written as

$$\mathscr{C}(m, y; \sigma, \gamma) = \frac{1}{y!} \sum_{\ell=1}^{y} (-1)^{\ell} \binom{y}{\ell} (-\ell\sigma - \gamma)_m,$$

see Charalambides (2005).

**Example 1.4** (Distribution of number of species for Dirichlet process)**.** For the Dirichlet process of parameter $M$, the prior distribution of the number $K_n$ of distinct species in a sample of size $n$ has been derived in Ewens (1972) and Antoniak (1974). Specifically, such law is given by

$$\mathbb{P}(K_n = k) = \frac{M^k}{(M)_n} |s(n, k)|, \qquad k = 0, \ldots, n,$$

which is a version of the celebrated *Ewens sampling formula*. Here, $|s(n, k)|$ denotes for the sign-less Stirling number of the first kind, which is related to the generalized factorial coefficients by

$$|s(n, k)| = \lim_{\sigma \to 0} \frac{\mathscr{C}(n, k; \sigma)}{\sigma^k}.$$

For the unseen species problem, the result of Lijoi et al. (2007) in (1.10) specializes as

$$\mathbb{P}(K_m^{(n)} = y \,|\, Z_1, \ldots, Z_n) = \frac{M^y (M)_n}{(M)_{n+m}} \sum_{\ell=y}^{m} \binom{m}{\ell} |s(\ell, y)| (n)_{m-\ell}, \qquad y = 0, \ldots, m.$$

As expected from the prediction rule in (1.7), the probability of discovering a certain number of new species does not depend on the number $k$ of species recorded in the initial sample. This feature of the Dirichlet process is undesirable from an inferential point of view since inference about the number of distinct species in a future sample would not depend on the number of distinct species present in the initial sample.

**Example 1.5** (Distribution of number of species for Pitman-Yor process)**.** For the Pitman-Yor process of parameter $(\sigma, M)$, the prior distribution of the number $K_n$ of distinct species in the sample of size $n$ is given by

$$\mathbb{P}(K_n = k) = \frac{\prod_{\ell=1}^{k-1} (M + \ell\sigma)}{\sigma^k (M + 1)_{n-1}} \mathscr{C}(n, k; \sigma), \qquad k = 0, \ldots, n.$$

For the unseen species problem, by applying (1.10), we obtain, for $y = 0, \ldots, m$,

$$\mathbb{P}(K_m^{(n)} = y \,|\, Z_1, \ldots, Z_n) = \frac{(M + 1)_{n-1}}{(M + 1)_{n+m-1}} \frac{\prod_{\ell=k}^{k+y-1} (M + \ell\sigma)}{\sigma^k} \mathscr{C}(m, y; \sigma, -n + k\sigma).$$

As for the prediction rule in (1.8), the probability of discovering a certain number of new species does depend also on the number $k$ of distinct species recorded in the initial sample, other than the sample size $n$. This feature of the Pitman-Yor process enriches the undesirable predictive structure of the Dirichlet process.

## 1.3 Sufficientness postulates for species sampling models

One of the three possible perspectives on a SSM is through its predictive distributions, as described in point (ii) of Remark 1.1. This viewpoint captures the sequential sampling process from the population, providing insight into how the discovery of new species evolves. Consequently, the quantities appearing in the predictive distributions admit a direct interpretation, which makes it easier to compare and select models when interpretability is the criterion. Recall that the predictive distributions of any SSM can be written as

$$Z_{n+1} \,|\, Z_1, \ldots, Z_n \sim \sum_{\ell=1}^{k} p_\ell(\boldsymbol{n})\delta_{X_\ell} + p_{k+1}(\boldsymbol{n})G_0, \tag{1.11}$$

where the functions $p_\ell$'s constitute the PPF of an infinite exchangeable random partition. In generality, the probability $p_{k+1}(\boldsymbol{n})$ of discovering a new species and the probabilities of recording already seen species, i.e., $p_\ell(\boldsymbol{n})$, for $\ell = 1, \ldots, k$, depend on the initial sample $(Z_1, \ldots, Z_n)$ through the whole information in the vector of multiplicities $\boldsymbol{n} = (n_1, \ldots, n_k)$, that is the sample size $n$, the number of distinct observed species $k$ and the multiplicities of the species themselves $n_\ell$, for $\ell = 1, \ldots, k$. Specific choices of the PPF yield specific dependencies on the initial sample for the predictive rule. In this context, such predictive characterizations are referred to as *sufficientness postulates*, a notion originally introduced by the English philosopher W.E. Johnson in the 1920s in relation to the symmetric Dirichlet distribution (see Zabell (1982) for a historical account). In the next discussion, instead of referring to the PPF, we equivalently refer to the corresponding SSP as prior for $\tilde{p}$ in (1.2).

In Example 1.3, we have recalled that if $\tilde{p}$ follows a Dirichlet process, then $p_{k+1}(\boldsymbol{n})$ depends on the initial sample only through $n$ and $p_\ell(\boldsymbol{n})$, for $\ell = 1, \ldots, k$, depends on $n$ and the multiplicity $n_\ell$. See the prediction rule in (1.7). Interestingly, this dependence structure on the initial sample characterizes the Dirichlet process within the class of SSMs. In particular, Regazzini (1978) and Lo (1991) prove the following.

**Proposition 1.1** (Predictive characterization of the Dirichlet process)**.** *Let $(Z_i)_{i\geq 1}$ be distributed as a SSM in* (1.2)*, whose predictive distributions write as in* (1.11)*. Then, the following are equivalent:*

*(i) as function of the initial sample, $p_{k+1}(\boldsymbol{n})$ depends only on $n$, and $p_\ell(\boldsymbol{n})$ depends only on $n$ and $n_\ell$;*

*(ii) the random probability measure $\tilde{p}$ is a Dirichlet process.*

In discussing the predictive rule (1.8) in Example 1.3, we noted that the Pitman-Yor process extends the predictive structure of the Dirichlet process by making $p_{k+1}(\boldsymbol{n})$ depend not only on the sample size $n$ but also on the number $k$ of distinct observed species. At the same time, similar to the Dirichlet process, the Pitman-Yor process assigns predictive probabilities $p_\ell(\boldsymbol{n})$ that depend on $n$ and the corresponding multiplicity $n_\ell$. A characterization of the Pitman-Yor process, paralleling Proposition 1.1, is provided in Zabell (2005).

**Proposition 1.2** (Predictive characterization of the Pitman-Yor process)**.** *Let $(Z_i)_{i\geq 1}$ be distributed as a SSM in* (1.2)*, whose predictive distributions write as in* (1.11)*. Then, the following are equivalent:*

*(i) as function of the initial sample, $p_{k+1}(\boldsymbol{n})$ depends only on $n$ and $k$, and $p_\ell(\boldsymbol{n})$ depends only on $n$ and $n_\ell$;*

*(ii) the random probability measure $\tilde{p}$ is a Pitman-Yor process.*

Finally, the class of Gibbs-type priors exhibits an even richer dependence on the initial sample in its predictive distributions, as seen in the general expression (1.6). In particular, $p_{k+1}(\boldsymbol{n})$ depends on both the sample size $n$ and the number $k$ of distinct observed species, as in the Pitman-Yor process. The key enrichment over the Pitman-Yor process lies in the probabilities $p_\ell(\boldsymbol{n})$ of observing an existing species, which depend not only on $n$ and the multiplicity $n_\ell$ but also on $k$. A characterization result for Gibbs-type priors is also available; see De Blasi et al. (2015) and Bacallado et al. (2017).

**Proposition 1.3** (Predictive characterization of the Gibbs-type processes)**.** *Let $(Z_i)_{i\geq 1}$ be distributed as a* SSM *in* (1.2)*, whose predictive distributions write as in* (1.11)*. Then, the following are equivalent:*

*(i) as function of the initial sample, $p_{k+1}(\boldsymbol{n})$ depends only on $n$ and $k$, and $p_\ell(\boldsymbol{n})$ depends only on $n$, $n_\ell$ and $k$;*

*(ii) the random probability measure $\tilde{p}$ is a Gibbs-type process.*

In conclusion, achieving predictive dependencies more complex than those described in point (i) of Proposition 1.3 requires moving beyond the class of Gibbs-type SSMs. In practice, this is typically done by placing hyperpriors on the parameters of the Dirichlet or Pitman-Yor processes, resulting in SSMs that no longer belong to the Gibbs-type class. While this approach increases modeling flexibility, it comes at the cost of tractability, requiring computational strategies, e.g., Markov chain Monte Carlo algorithms, for fitting the model to data.

Moreover, the preceding characterization results hold within the class of SSMs defined in (1.2), whose predictive distributions take the form in (1.11). In this setting, the most general dependence of the predictive distributions on the initial sample is through the vector of multiplicities $\boldsymbol{n} = (n_1, \ldots, n_k)$, with no dependence on the specific values of the labels $X_\ell$ for $\ell = 1, \ldots, k$. This property follows naturally from the i.i.d. assumption on the $\tilde{X}_j$'s in (1.2), and it is entirely consistent with the species sampling framework, where the labels themselves are irrelevant.

However, when SSMs are used as latent structures for clustering continuous observations, the atoms $\tilde{X}_j$ of $\tilde{p}$ in (1.2) become important. In such cases, extensions of SSMs allowing the $\tilde{X}_j$'s to have a more complex joint distribution have been considered in the literature. In these scenarios, the predictive distributions may depend both on $\boldsymbol{n}$ and on the values of the $X_\ell$'s. We discuss this in more detail in the next section.

## 1.4 Species sampling models as latent structures in mixture models

From the three perspectives summarized in Remark 1.1, it is clear that SSMs are naturally suited for modeling discrete observations. When the data are continuous, SSMs are typically combined with appropriate kernels to define mixture models, which provide a

flexible framework for density estimation and model-based clustering. In this context, we denote realizations of the SSM by $(\theta_i)_{i\geq 1}$ rather than $(Z_i)_{i\geq 1}$, as these now represent latent parameters in the mixture model. Specifically, a common formulation of a mixture model for continuous observations $(Y_i)_{i\geq 1}$ is, for $i \geq 1$,

$$\begin{aligned} Y_i \,|\, \theta_i &\overset{\text{ind}}{\sim} f(\cdot;\theta_i) \\ (\theta_i)_{i\geq 1} &\sim \text{SSM}, \end{aligned} \tag{1.12}$$

where $\{f(\cdot;\theta)\}_{\theta\in\Theta}$ is a parametric family of densities, commonly referred to as the kernel of the mixture model, with the parameter space $\Theta$ depending on the application. For real-valued observations $(Y_i)_{i\geq 1}$, Gaussian kernel mixture models represent the most popular example: in this case, $\theta = (\mu, \sigma)$, $\Theta = \mathbb{R} \times \mathbb{R}_+$, and $f(\cdot;\theta) = \mathcal{N}(\cdot;\mu,\sigma^2)$, where, by an abuse of notation, we use $\mathcal{N}(\cdot;\mu,\sigma^2)$ to denote the probability density function of a Gaussian random variable with mean $\mu$ and variance $\sigma^2$. The formulation in (1.12) defines the general class of nonparametric mixtures, which have been first introduced in Ferguson (1983) and Lo (1984), and became popular due to Escobar and West (1995).

For the purpose of density estimation, the mixture model in (1.12) provides a flexible class of distributions for the observations, as it can be expressed as

$$Y_i \,|\, (W_j)_{j\geq 1}, (\tilde{\theta}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \sum_{j\geq 1} W_j f(\cdot;\tilde{\theta}_j), \qquad (W_j)_{j\geq 1} \sim \mathcal{W}, \;\; \tilde{\theta}_j \overset{\text{iid}}{\sim} G_0, \; j \geq 1,$$

for some distribution $\mathcal{W}$ such that $\sum_{j\geq 1} W_j = 1$. Here, the index $j$ runs over the mixture components, with $f(\cdot;\tilde{\theta}_j)$ and $W_j$ representing the density and weight of component $j$, respectively. The model's flexibility is particularly evident simply by considering the Dirichlet process as a special case of a SSM: in this case, the induced distribution on $(Y_i)_{i\geq 1}$ has full support under the weak topology on the space of probability measures (Ferguson, 1973). In other words, the law of $\tilde{p}$ assigns positive probability to neighborhoods of any target distribution $P_0$, provided that $P_0$ is absolutely continuous with respect to the base measure $G_0$.

As far as model-based clustering is concerned, the mixture models in (1.12) represent the most popular approach. To best understand why SSMs are particularly well-suited for inferring clusters among continuous observations $Y_1, \ldots, Y_n$, it is helpful to adopt the perspective described in point (iii) of Remark 1.1. From this viewpoint, the sample $\theta_1, \ldots, \theta_n$ drawn from a SSM can be equivalently characterized in terms of the partition of the $n$ subjects $\mathcal{P}_n = \{C_1, \ldots, C_k\}$ and the distinct labels $\theta^*_\ell$ associated with each cluster, so that $C_\ell = \{i : \theta_i = \theta^*_\ell\}$. The mixture model in (1.12) can be formulated as

$$\begin{aligned} Y_i : i \in C_\ell \,|\, \mathcal{P}_n, \theta^*_1, \ldots, \theta^*_k &\overset{\text{iid}}{\sim} f(\cdot;\theta^*_\ell), \\ \mathcal{P}_n = \{C_1, \ldots, C_k\} &\sim \pi(n_1, \ldots, n_k), \end{aligned}$$

where $\pi$ is the EPPF of the exchangeable random partition. In this representation, the random partition $\mathcal{P}_n$ directly determines the clustering of the $n$ observations, while each cluster $C_\ell$ is associated with a cluster-specific parameter $\theta^*_\ell$, which governs the distribution of the observations in that cluster. In other words, all observations in cluster $C_\ell$ are drawn from $f(\cdot;\theta^*_\ell)$. This makes it clear that the values of $\theta^*_\ell$, and correspondingly the atoms $\tilde{\theta}_j$ of $\tilde{p} = \sum_{j\geq 1} W_j \delta_{\tilde{\theta}_j}$ in (1.2), are crucial in the clustering context, as they define the cluster-specific parameters of the likelihood. At the same time, the i.i.d. assumption for the atoms

$\tilde{\theta}_j$ tends to favor density estimation over clustering. Illustrative examples can be found in Beraha et al. (2022). To reverse the trade-off in favor of clustering, a substantial body of work (Petralia et al., 2012; Xu et al., 2016; Fúquene et al., 2019; Quinlan et al., 2021; Bianchini et al., 2020; Xie and Xu, 2019) has focused on enforcing separation among the component densities. This is typically achieved by placing a (repulsive) point process prior on the component parameters $\tilde{\theta}_j$, which encourages distinct and well-separated cluster-specific distributions.

## 1.5 Extending the species framework to multiple populations

The classical framework for species sampling models relies on the assumption of full exchangeability among observations, implying a homogeneous structure across all subjects. This assumption has underlain the discussion throughout Chapter 1. However, in many applied contexts, full exchangeability is overly restrictive. The seminal works of MacEachern (1999, 2000) paved the way for a broad line of research in statistics and machine learning aimed at relaxing this assumption to accommodate heterogeneity across data sources or experimental conditions. In particular, when observations are collected from multiple distinct populations, such as in multi-center studies, the assumption of exchangeability across all samples becomes unrealistic, as it neglects between-group differences. Conversely, assuming complete independence between populations precludes any sharing of information, which is often desirable in multi-sample analyses. A natural compromise is partial exchangeability, which assumes exchangeability within, but not across, populations while still permitting dependence among them. Analogously to the fully exchangeable case in (1.2), partial exchangeability implies the existence of a vector of (dependent) random probability measures $(\tilde{p}_1, \dots, \tilde{p}_d)$ such that

$$Z_{iq} \,|\, \tilde{p}_1, \dots, \tilde{p}_d \overset{\text{ind}}{\sim} \tilde{p}_q, \qquad i \geq 1, q = 1, \dots, d,$$

where $Z_{iq}$ denotes observation $i$ in group $q$, and $d$ is the number of groups. From a Bayesian viewpoint, selecting a partially exchangeable model corresponds to specifying a prior distribution on the vector of dependent random probability measures. Countless approaches have been proposed in the literature for this partially exchangeable setting, collectively known as *dependent nonparametric priors*. These encompass hierarchical Dirichlet processes (Teh et al., 2006), hierarchical normalized completely random measures (Camerlenghi et al., 2019), hierarchical species sampling models (Bassetti et al., 2020), nested constructions (Rodriguez et al., 2008; Camerlenghi et al., 2019), additive constructions (Müller et al., 2004; Lijoi et al., 2014). Other notable contributions include normalized compound random measures (Griffin and Leisen, 2017), normalized completely random vectors (Catalano et al., 2021), single-atoms dependent processes (MacEachern, 1999, 2000; Quintana et al., 2022), vectors of normalized independent finite point processes (Colombi et al., 2025). Recent developments combine some of these mechanisms (Beraha et al., 2021; Lijoi et al., 2023; Denti et al., 2023). Comprehensive overviews of this literature can be found in Quintana et al. (2022), while Franzolini et al. (2025) introduces the class of *multivariate species sampling models*, offering a unified theoretical framework that encompasses many of the above specifications.

# 2. The Feature Allocation Framework

In this chapter, we review and discuss the central theme of the thesis: the *feature allocation* framework. Unlike the species setting reviewed in Chapter 1, where each individual is assigned to a single species, the feature allocation framework allows each observation to be associated with a finite collection of *features* or *characteristics*. Since the introduction of the celebrated Indian buffet process (IBP) by Griffiths and Ghahramani (2005), feature allocation models (Broderick et al., 2013) have gained significant attention across statistics, machine learning, and related fields. The seminal paper by Thibaux and Jordan (2007) established a connection between the original construction of the IBP as a distribution over binary matrices (Griffiths and Ghahramani, 2005) and its Bayesian nonparametric formulation under a beta process prior (Hjort, 1990). This finding not only clarified the probabilistic structure of the IBP, but also linked it to the broader theory of completely random measures (Kingman, 1967), thereby laying the foundation for a new branch of Bayesian nonparametrics. Subsequent research has extended the IBP by modifying both prior distributions and likelihoods. A comprehensive analysis of such generalizations of the IBP when the prior is a completely random measure (CRM) can be found in Broderick et al. (2018) and James (2017), while non-exchangeable constructions inspired by CRMs have been proposed in Benedetto et al. (2020). These developments sparked a rich stream of research, particularly within the machine learning community. Applications of feature models now span diverse areas, such as Bayesian factor analysis and nonnegative matrix factorization (Griffiths and Ghahramani, 2005; Knowles and Ghahramani, 2011; Ayed and Caron, 2021; Zhou et al., 2012, 2016), image segmentation (Titsias, 2007; Griffiths and Ghahramani, 2011; Hu et al., 2012; Broderick et al., 2015), text mining (Thibaux and Jordan, 2007; Teh and Gorur, 2009), topic modeling (Williamson et al., 2010), network analysis (Miller et al., 2009; Palla et al., 2012), ecology (Stolf and Dunson, 2025), and microbiome studies (James et al., 2025). For comprehensive surveys of early contributions, see Teh and Jordan (2010) and Griffiths and Ghahramani (2011).

The structure of this chapter is designed to parallel that of Chapter 1, thereby emphasizing the similarities and differences between the two frameworks. While this chapter primarily serves as an introduction, its purpose is to establish the foundations necessary for the novel contributions presented later in the thesis. The mirroring structure is intentional: it not only facilitates direct comparison with the species sampling setting but also highlights important gaps in the current feature allocation literature. By drawing these connections, the present chapter shows how our work advances the feature allocation framework.

The chapter is organized as follows. In Section 2.1, we introduce the feature allocation framework from the basics and discuss the three fundamental perspectives for describing

feature models. Section 2.2 then focuses on the class of Gibbs-type feature models, so named for their close analogy with Gibbs-type species sampling models. We review representative examples, which coincide with the most important models commonly employed in the feature setting, and illustrate how they can be used to address key inferential problems. In Section 2.3, we turn to predictive characterizations for feature models, which play a central role in inference. Section 2.4 reviews the important generalization from feature allocation models to trait allocation models. Finally, Section 2.5 considers the case of partially exchangeable trait allocations, which extend beyond the standard exchangeable setting and are particularly relevant when the data are naturally partitioned into groups.

## 2.1 Feature sampling: foundations and modeling perspectives

The *feature sampling* framework fundamentally differs from the *species sampling* framework discussed in Chapter 1 in the mechanism by which data are collected, although the two are closely related, as will become apparent shortly. In the *feature sampling* context, each subject in the population is associated with a *collection of species*, here referred to as *features*. To make the analogy concrete, consider an ecological setting in which subjects correspond to traps or plots within a region, and the features assigned to each subject represent the species of animals or plants found there. Given a random sample $(Z_1, \ldots, Z_n)$ drawn from the population, the collection of $n_i$ features associated with the $i$th subject is represented by $Z_i = \{X_{i,1}, \ldots, X_{i,n_i}\}$, with $X_{i,\ell} \in \mathbb{X}$, $\ell = 1, \ldots, n_i$, where $\mathbb{X}$ should be regarded as an abstract set of labels used to identify the various features. As in the species sampling case, it is assumed that all subjects are collected under identical experimental conditions, implying that their order is irrelevant for statistical analysis. This assumption naturally leads to the requirement of *exchangeability* among the $Z_i$'s.

Analogously to the species sampling setting, a number of inferential questions arise once a sample $(Z_1, \ldots, Z_n)$ has been observed. These are collectively referred to as *feature sampling problems*. The most prominent among them is the *unseen feature problem*, a natural generalization of the unseen species problem, which concerns the estimation of

$$K_m^{(n)} := |\{Z_{n+1}, \ldots, Z_{n+m}\} \setminus \{Z_1, \ldots, Z_n\}|, \tag{2.1}$$

namely the number of hitherto unseen (distinct) features that would be observed if $m \geq 1$ additional subjects $(Z_{n+1}, \ldots, Z_{n+m})$ were collected from the same population. It is worth emphasizing that the expression in (2.1) formally coincides with the corresponding quantity in (1.1) for the *unseen species problem*, with the key distinction that here each $Z_i$ may contain multiple features. A recent contribution addressing the unseen feature problem is provided by Masoero et al. (2022). Applications of this problem have been explored in several domains, including genomic studies (Masoero et al., 2021, 2022; Camerlenghi et al., 2024; Shen et al., 2024), user activity prediction in online A/B testing (Beraha et al., 2024; Masoero et al., 2024). Other feature sampling problems naturally parallel their species sampling counterparts. For instance, one may be interested in predicting the number of previously unseen features that will appear exactly $r \geq 1$ times in future samples.

A convenient and equivalent representation of the information contained in the $Z_i$'s can be constructed as follows. We may regard the feature labels appearing in the $Z_i$'s as being

drawn from a common (at most countable) set of feature tags, denoted by the sequence $(\tilde{X}_j)_{j\geq 1}$. For a rigorous treatment, assume that each $\tilde{X}_j \in \mathbb{X}$, where $\mathbb{X}$ is a Polish space. Each individual $i$ is then characterized by the collection of features it expresses, that is, by the sequence of pairs $((\tilde{X}_j, \tilde{A}_{ij}))_{j\geq 1}$, where $\tilde{A}_{ij} = 1$ if the $i$th individual exhibits feature $\tilde{X}_j$, and $\tilde{A}_{ij} = 0$ otherwise. These pairs can be organized into a counting measure $Z_i$ on $\mathbb{X}$ defined as

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

This formulation is particularly convenient, as it introduces a shared collection of potential feature labels $(\tilde{X}_j)_{j\geq 1}$ for all subjects, while the binary indicators $\tilde{A}_{ij}$'s specify which features are actually expressed by each individual.

Among the exchangeable probabilistic models for a sequence $(Z_i)_{i\geq 1}$, the Bayesian literature has focused on the assumption that, conditionally to a sequence of random probabilities $(\tilde{q}_j)_{j\geq 1}$, the variables $\tilde{A}_{ij}$'s are independent Bernoulli random variables. Formally, it is assumed that

$$\tilde{A}_{ij} \,|\, \tilde{q}_j \overset{\text{ind}}{\sim} \text{Bernoulli}(\tilde{q}_j).$$

The probabilities $(\tilde{q}_j)_{j\geq 1}$ can in turn be organized into a random measure $\tilde{\mu}$ on $\mathbb{X}$, defined by

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

We will say that, conditionally on $\tilde{\mu}$, the random measures $Z_i$'s are i.i.d. *Bernoulli processes* with base measure $\tilde{\mu}$, denoted as

$$Z_i \,|\, \tilde{\mu} \overset{\text{iid}}{\sim} \text{BP}(\tilde{\mu}), \qquad i \geq 1.$$

This leads naturally to the definition of a *feature frequency model*, originally introduced in Broderick et al. (2013), which we report next.

**Definition 5** (Feature frequency model)**.** *A feature frequency model consists of a sequence of random measures $(Z_i)_{i\geq 1}$ and a random measure $\tilde{\mu}$ such that*

$$Z_i \,|\, \tilde{\mu} \overset{iid}{\sim} BP(\tilde{\mu}), \qquad \tilde{\mu} = \sum_{j\geq 1} \tilde{q}_j\delta_{\tilde{X}_j}, \tag{2.2}$$

*where $\tilde{X}_j \overset{iid}{\sim} G_0$, with $G_0$ a diffuse probability distribution on $\mathbb{X}$, and $(\tilde{q}_j)_{j\geq 1}$ is an independent random sequence such that $\tilde{q}_j \in (0, 1], j \geq 1$.*

In this framework, $\tilde{q}_j$ represents the frequency of feature $\tilde{X}_j$ in the population. As in the species sampling context, assuming that the feature labels $\tilde{X}_j$'s are i.i.d. from a diffuse measure $G_0$ is not restrictive: these labels simply serve as unique identifiers and have no intrinsic meaning. Finally, note that while any feature frequency model induces an exchangeable sequence of random measures $(Z_i)_{i\geq 1}$, the converse does not hold. Exchangeability alone does not imply the specific structure of a feature frequency model.

Feature sampling problems, analogously to the species sampling problems discussed in Chapter 1, are primarily concerned with prediction for future observations. The most popular example is the unseen feature problem, which entails predicting the number

$K_m^{(n)}$ in (2.1), representing the number of distinct unseen features in the observed sample $(Z_1, \ldots, Z_n)$ that would appear in an additional sample $(Z_{n+1}, \ldots, Z_{n+m})$ of size $m \geq 1$. More generally, all prediction tasks in this framework ultimately reduce to the problem of determining the distribution of the next observation $Z_{n+1}$, conditional on the available data $\boldsymbol{Z} := (Z_1, \ldots, Z_n)$, for any $n \geq 1$. By exchangeability of the sequence $(Z_i)_{i \geq 1}$, the relevant information contained in the observed sample for prediction purposes can be summarized as follows. Let $K_n = k$ denote the number of distinct observed features in $\boldsymbol{Z}$, with corresponding feature labels $X_\ell$, as $\ell = 1, \ldots, k$, which form a subset of the complete list of labels $(\tilde{X}_j)_{j \geq 1}$. Define binary indicators $A_{i\ell} := Z_i(\{X_\ell\})$, so that $A_{i\ell} = 1$ if the $i$th individual displays feature $X_\ell$ and $A_{i\ell} = 0$ otherwise. Let $\boldsymbol{m} = (m_1, \ldots, m_k)$ denote the feature frequencies, with $m_\ell = \sum_{i=1}^n A_{i\ell}$ indicating how many subjects display feature $X_\ell$, for $\ell = 1, \ldots, k$. It is important to note that, within the feature sampling framework, it is typically assumed that each individual exhibits only a finite number of features: as a result, the total number of observed features $K_n$ is almost surely finite for any $n \geq 1$.

Under a feature frequency model of the form (2.2), the predictive distributions take the form $Z_1 \sim \text{BP}(\tilde{\mu})$ and, for $n \geq 1$,

$$Z_{n+1} \mid \boldsymbol{Z} \stackrel{\text{d}}{=} \sum_{\ell=1}^{k} A_{n+1,\ell} \delta_{X_\ell} + Z'_{n+1}, \tag{2.3}$$

where:

(i) conditionally on a vector of random parameters $(q_1, \ldots, q_k)$, $(A_{n+1,1}, \ldots, A_{n+1,k})$ consists of independent Bernoulli random variables with success probabilities $(q_1, \ldots, q_k)$;

(ii) $Z'_{n+1}$ is a random measure composed of $Y_{n+1}$ atoms $\tilde{X}'_j \stackrel{\text{iid}}{\sim} G_0$, where the random number of atoms $Y_{n+1}$ generally depends on $(q_1, \ldots, q_k)$.

This predictive form follows as a special case of the general results in Chapter 4, specialized to i.i.d. atoms that are independent of the probabilities $(\tilde{q}_j)_{j \geq 1}$. The rule in (2.3) admits an intuitive interpretation: conditionally to $(q_1, \ldots, q_k)$, the next subject expresses each of the *old* features $X_\ell$ independently with probability $q_\ell$, and also exhibits a random number $Y_{n+1}$ of *new*, previously unobserved features. In general, prediction in feature frequency models is governed by the random parameters $(q_1, \ldots, q_k)$ and by the conditional distribution of $Y_{n+1}$ given these parameters. The joint distribution of $(q_1, \ldots, q_k)$ depends on the observed data through the sample size $n$ and the feature frequencies $\boldsymbol{m} = (m_1, \ldots, m_k)$, consequently, the distribution of $Y_{n+1}$ depends on these same quantities marginally. As in the case of species sampling models, classes of priors for $\tilde{\mu}$ can be characterized in terms of specific properties they induce on the predictive rule in (2.3). This is the main contribution in Chapter 4, Section 4.3. We briefly introduce these ideas in Section 2.3, drawing a parallel with the *sufficientness postulates* in the species sampling framework reviewed in Section 1.3, and present the main theoretical results in Section 4.3.

The predictive rule in (2.3) describes the sequential procedure through which a sample $(Z_1, \ldots, Z_n)$ from a feature frequency model is generated. The essential information contained in a sample of size $n$ can be expressed by the number $K_n = k$ of distinct observed features and, for each observed feature, by the subset of subjects exhibiting it.

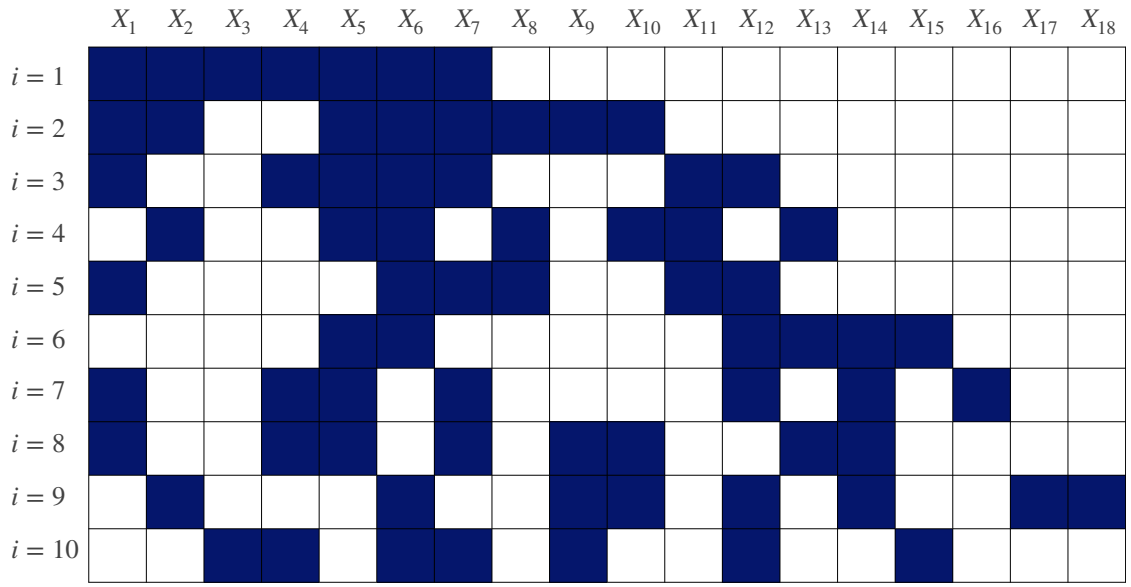



Figure 2.1: Matrix representation of a feature allocation, with $n = 10$ individuals and $K_n = 18$ observed features. Features are in *order of appearance* (Broderick et al., 2013). The blue squares correspond to $A_{i\ell} = 1$ (the $i$th individual displays feature $X_\ell$), the white squares correspond to $A_{i\ell} = 0$ (the $i$th does not express feature $X_\ell$).

This pattern of features is mathematically represented by the notion of *feature allocation*. Specifically, the ordered *feature allocation* is defined as the sequence $F_n = (B_{n,1}, \ldots, B_{n,k})$ of non-empty sets $B_{n,\ell} \subseteq [n]$, for $\ell = 1, \ldots, k$, where $B_{n,\ell}$ identifies the set of individuals displaying the $\ell$th feature. A feature allocation can be conveniently represented as a binary matrix, as illustrated in Figure 2.1, where $n = 10$ individuals express $K_n = 18$ features. Each column of the binary matrix corresponds to one set $B_{n,\ell}$, with the $i$th entry of the $\ell$th column representing $A_{i\ell}$. Since each individual is assumed to exhibit only finitely many features, every individual belongs to finitely many sets $B_{n,\ell}$. The use of feature labels in feature frequency models (2.2), or equivalently in their predictive representation (2.3), is a convenient modeling trick, analogous to the labeling mechanism in species sampling models. Indeed, a sample $(Z_1, \ldots, Z_n)$ induces the associated ordered feature allocation $F_n = (B_{n,1}, \ldots, B_{n,k})$ through $B_{n,\ell} = \{i \in [n] : Z_i(\{X_\ell\} = 1\}$, and the distinct feature labels $X_\ell$'s associated to the sets of the feature allocation. In the pure feature sampling framework, the feature labels $X_\ell$ are irrelevant: the fundamental object of interest is the feature allocation $F_n$ itself, particularly the number of features $K_n$. However, when feature models are applied beyond the classical feature sampling framework (see, e.g., Section 4.5), the feature labels $X_\ell$ become meaningful and play a central role in inference.

In terms of probability distribution induced by a feature frequency model on $F_n = (B_{n,1}, \ldots, B_{n,k})$, $n \geq 1$, exchangeability of $(Z_i)_{i\geq 1}$ implies that the law depends on the sample size $n$ and on the feature frequencies $\boldsymbol{m} = (m_1, \ldots, m_k)$, where $m_\ell = \#B_{n,\ell}$ is the frequency of the $\ell$th feature. In particular, for any feature frequency model, the probability distribution of $F_n$ writes as

$$\mathsf{P}(F_n = f_n) = \pi_n(m_1, \ldots, m_k), \tag{2.4}$$

for every $f_n$ and $n \geq 1$, where $\pi_n$ 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 $\pi_n$ is referred to as the *exchangeable feature probability function* (EFPF) and it encapsulates all the relevant properties of the model. Its role is directly analogous to that of the exchangeable partition probability function (EPPF) in the species sampling framework, as carefully discussed in Broderick et al. (2013).

**Remark 2.1.** *A feature allocation model is a probability distribution on the ordered feature allocation $F_n = (B_{n,1}, \ldots, B_{n,K_n})$. Alternatively, one may consider the probability distribution of the 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 $B_{n,1}, \ldots, B_{n,K_n}$, with $\tilde{K}_{n,h}$ being the number of sets equal to $\tilde{B}_{n,h}$. Although $\tilde{F}_n$ is sometimes used to define feature allocation models (see Broderick et al., 2013), it is often more convenient to work with the ordered representation $F_n$. In any case, it is assumed that the probability of observing an unordered allocation $\tilde{F}_n = \tilde{f}_n$ is uniformly distributed among the $K_n!/\prod_{h=1}^{H_n} \tilde{K}_{n,h}!$ possible orderings of the sets $B_{n,1}, \ldots, B_{n,K_n}$, that is,*

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

*where $f_n$ denotes one such ordering of $\tilde{f}_n$.*

Feature frequency models, and consequently their associated EFPFs, are closely connected to the broader notion of *infinite exchangeable random feature allocations*. However, the nature of this connection differs substantially from that between species sampling models and infinite exchangeable random partitions, although there are evident analogies between the two settings. The concept of exchangeable random feature allocations is formally introduced by Broderick et al. (2013), extending the theory of exchangeable random partitions developed by Aldous (1985).

**Definition 6** (Exchangeable random feature allocation)**.** *A random feature allocation $F_n$ of $\{1, \ldots, n\}$ is said to be exchangeable if, for any permutation $\sigma$ of $\{1, \ldots, n\}$ and any feature allocation $f_n = (B_{n,1}, \ldots, B_{n,k})$, it holds that*

$$\mathbb{P}(F_n = f_n) = \mathbb{P}(F_n = (\sigma(B_{n,1}), \ldots, \sigma(B_{n,k}))).$$

*An infinite exchangeable random feature allocation is a consistent sequence $(F_n)_{n\geq 1}$ of exchangeable random feature allocations, where consistency means that for every $n$, $F_{n-1}$ is equal almost surely to the feature allocation obtained from $F_n$ by removing element $n$.*

Feature allocation models $(F_n)_{n\geq 1}$ that admit EFPFs as in (2.4) are exchangeable random feature allocations. It is important to note, however, that the class of exchangeable feature allocation models admitting EFPFs forms a strict subset of all infinite exchangeable feature allocation models (Broderick et al., 2013, Proposition 7). This situation contrasts with the case of infinite exchangeable random partitions and species sampling models with EPPFs (see Chapter 1), where the two classes are equivalent.

Feature frequency models induce feature allocation models that admit EFPFs. In fact, they are essentially equivalent to the class of such models, up to a mild additional assumption. This assumption corresponds to the notion of regular feature allocation models,

which excludes degenerate features that appear in only a single observation across the population. Formally, this requires that if a feature is observed in a finite sample $F_n$, i.e., $B_{n,\ell}$ is non-empty, then it is displayed by infinitely many individuals with probability one, i.e., $\lim_{n\to\infty} \#B_{n,\ell} = \infty$ almost surely. (Broderick et al., 2013, Theorem 18) established that the class of regular feature allocation models with EFPFs is equivalent to the class of feature frequency models.

Remark 1.1 in Chapter 1 outlined three complementary perspectives through which SSMs can be conveniently described in the species sampling framework. The natural analogue of SSMs in the feature setting, comparable in popularity, tractability, and flexibility, is the class of *(regular) feature allocation models with EFPFs*, namely the class of *feature frequency models*. From now on, we will omit the specification *regular*, though assumed implicitly. Drawing a parallel with Remark 1.1, we conclude this general introduction to the feature sampling framework by summarizing the discussion above in the following remark.

**Remark 2.2.** *Any feature frequency model can be equivalently characterized through any of the following three pairs of objects:*

*(i) the prior specification of $\tilde{\mu}$ in (2.2), i.e., $G_0$ and the distribution of $(\tilde{q}_j)_{j\geq 1}$;*

*(ii) the predictive distributions in (2.3), i.e., $G_0$, the law of $(q_1, \ldots, q_k)$ and the conditional law of $Y_{n+1}$ given $(q_1, \ldots, q_k)$;*

*(iii) the infinite exchangeable feature allocation model $(F_n)_{n\geq 1}$ and the law of the distinct feature labels $X_\ell$'s, i.e., $G_0$ and the EFPF $\pi := (\pi_n)_{n\geq 1}$.*

*As in the species sampling case, each perspective emphasizes different aspects of the model and may prove more convenient depending on the inferential or modeling context.*

The class of *feature frequency models* has been defined in Broderick et al. (2013). However, statistical analysis has so far largely focused on two prominent examples: the Indian buffet process (IBP, Griffiths and Ghahramani, 2005) and its generalizations, and the beta Bernoulli (BB) models. We will review these models in Section 2.2. Building on these foundations, Battiston et al. (2018) introduced a broader family of feature allocation models with EFPFs in product form—directly analogous to the Gibbs-type random partitions in the species setting (see Section 1.2). In Section 2.2, we introduce this class, referred to as *Gibbs-type feature allocation models*, and review the notable examples of IBP and BB models. Then, in Chapter 3, we develop an extensive investigation of their statistical properties and establish results for the feature framework that mirror those of Lijoi et al. (2007) in the species sampling context. From Chapter 3, it will become evident that the Gibbs-type class provides an appealing compromise between analytical tractability and modeling flexibility, paralleling its role in the species sampling literature.

## 2.2 Gibbs-type feature allocation models

Gibbs-type feature models are introduced by Battiston et al. (2018) under the name of *feature allocation models with EFPF in product form*. The term *Gibbs-type* highlights the close

analogy with exchangeable Gibbs-type random partitions described in Section 1.2. Specifically, this class contains all feature allocation models whose EFPF depends multiplicatively on the sizes of the feature allocation sets, as

$$\pi_n(m_1, \ldots, m_k) = V_{n,k} \prod_{\ell=1}^{k} W_{m_\ell} U_{n-m_\ell},$$

for any $n$, $(m_1, \ldots, m_k) \in [n]^k$, 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, Battiston et al. (2018) show that Gibbs-type feature allocation models necessarily take the form

$$\pi_n(m_1, \ldots, m_k) = V_{n,k} \prod_{\ell=1}^{k} (1-\alpha)_{m_\ell - 1} (\theta + \alpha)_{n-m_\ell}, \tag{2.5}$$

for $-\infty < \alpha < 1$ and $-\alpha < \theta < \infty$. The array $V$ must satisfy the recurrence relationship

$$V_{n,k} = \sum_{j=0}^{\infty} \binom{k+j}{j} \{(\theta + \alpha)_n\}^j (\theta + n)^k V_{n+1,k+j}, \tag{2.6}$$

which ensures the consistency of the resulting EFPF. For fixed parameters $(\theta, \alpha)$, multiple arrays $V$ may satisfy the recursion, each corresponding to a distinct Gibbs-type feature allocation model.

By pairing a Gibbs-type feature allocation model with a diffuse probability measure $G_0$, one obtains the corresponding feature frequency model, as discussed in point (iii) of Remark 2.2. To lighten the notation, we will use the term *Gibbs-type feature model* to refer interchangeably to both the Gibbs-type feature allocation model and its associated feature frequency model, with the intended meaning clear from context. The most prominent and widely studied members of this class trace back to the work of Griffiths and Ghahramani (2005), who represented feature allocation models as stochastic mechanisms for generating binary matrices, as illustrated in Figure 2.1. In particular, Griffiths and Ghahramani (2005) introduced the Indian buffet process (IBP) as a stochastic process over binary matrices, obtained as the infinite-limit case of the beta Bernoulli (BB) models. We review these foundational models next.

**Example 2.1** (Beta Bernoulli model (Griffiths and Ghahramani, 2005))**.** The beta Bernoulli (BB) model with parameters $(M, \alpha, \theta)$, where $M \in \mathbb{N}$, $\alpha < 0$ and $\theta > -\alpha$ is a Gibbs-type feature model such that the $V_{n,k}$'s in (2.5) are given by

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

where $\mathbb{1}_C$ denotes the indicator function of a set $C$. This model imposes an upper bound $M$ on the number of distinct observed features $K_n$.

From the perspective of the associated feature frequency model, the random measure $\tilde{\mu}$ in (2.2) is given by

$$\tilde{\mu} = \sum_{j=1}^{M} \tilde{q}_j \delta_{\tilde{X}_j},$$

where $\tilde{X}_j \overset{\text{iid}}{\sim} G_0$ and $\tilde{q}_j \overset{\text{iid}}{\sim} \text{Beta}(-\alpha, \theta+\alpha)$, for $j = 1, \ldots, M$.

**Example 2.2** (Indian buffet process and its generalizations (Griffiths and Ghahramani, 2011; Teh and Gorur, 2009))**.** The most general formulation, known as the three-parameter Indian buffet process (IBP), is introduced in Teh and Gorur (2009). It is parametrized by $(\gamma, \alpha, \theta)$ satisfying $\gamma > 0$, $0 \leq \alpha < 1$, $\theta > -\alpha$, and is characterized by the coefficients

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

This definition generalizes the two-parameter IBP introduced in the pioneering work of Griffiths and Ghahramani (2005), which is recovered by setting $\alpha = 0$, and further reduces to the one-parameter model when $\theta = 1$. Notably, the two-parameter IBP was derived by Griffiths and Ghahramani (2005) as an infinite-limit of BB models, viewed as a stochastic mechanism for binary matrices or, equivalently, as a feature allocation model. A representation of the two-parameter IBP as a feature frequency model was later established in the illuminating contribution of Thibaux and Jordan (2007), who showed that $\tilde{\mu}$ in (2.2) follows a beta process (Hjort, 1990). The three-parameter IBP extends this construction by replacing the beta process with the stable-beta process (Teh and Gorur, 2009). We now recall the construction of the stable-beta process, with the beta process arising as a special case. In contrast to the BB model in Example 2.1, the random measure $\tilde{\mu}$ in the three-parameter IBP is more elaborate, as it involves infinitely many feature labels $\tilde{X}_j$ and associated probabilities $\tilde{q}_j$. Let us first define the class of homogeneous completely random measures (CRMs, Kingman, 1967) without fixed atoms, characterized by a Laplace functional of the type

$$\mathsf{E}[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}] = \exp\left\{-\int_{\mathbb{X}}\int_0^\infty \left[1 - e^{-sf(x)}\right]\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\},$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}_+$. Here, $\rho(\mathrm{d}s)$ is the Lévy intensity measure on $\mathbb{R}_+$, determining the distribution of the probabilities $(\tilde{q}_j)_{j\geq 1}$, and $G_0$ is a diffuse distribution on $\mathbb{X}$ from which the labels $\tilde{X}_j$ are sampled. We write $\tilde{\mu} \sim \text{CRM}(\rho; G_0)$. Comprehensive treatments of CRMs are provided by Daley and Vere-Jones (2008), and their role as a unifying framework in Bayesian nonparametrics is discussed in Lijoi and Prünster (2010). In the generative model (2.2), the random measure $\tilde{\mu}$ must have jumps $\tilde{q}_j \in (0, 1]$; thus, we consider $\rho(\mathrm{d}s)$ with support in $(0, 1]$. As shown by Teh and Gorur (2009), in the IBP the measure $\tilde{\mu}$ follows a stable-beta process, a specific type of CRM with Lévy intensity

$$\rho(\mathrm{d}s) = \gamma\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}s. \tag{2.7}$$

The parameter $\gamma$ is often referred to as the *total mass* since $\gamma = \mathsf{E}[\tilde{\mu}(\mathbb{X})] = \sum_{j\geq 1}\mathsf{E}[\tilde{q}_j]$ represents the expected sum of all the probabilities. The special case $\alpha = 0$ yields the beta process of Hjort (1990).

Unlike the BB model in Example 2.1, where the number $K_n$ of distinct observed features is bounded by the fixed value $M$, in the IBP case $K_n$ has unbounded support. In particular, as $n$ diverges, the two-parameter IBP (Griffiths and Ghahramani, 2011) exhibits a logarithmic growth rate of $K_n$, a property that motivated the introduction of the three-parameter

generalization by Teh and Gorur (2009), which yields instead a power-law behavior. These behaviors naturally emerge from the general theory of Gibbs-type feature models we discuss in Chapter 3, to which both the two- and three-parameter IBP belong.

Remarkably, a characterization theorem by Battiston et al. (2018) shows that the IBP and the BB models constitute the building blocks of all Gibbs-type feature models. More precisely, for fixed values of $(\theta, \alpha)$, the set of coefficients $V_{n,k}$ in (2.5) can always be expressed as mixtures with respect to the parameters $\gamma$ and $M$ of the IBP and BB models, respectively. This result is formalized below.

**Proposition 2.1** (Theorem 1.1 of Battiston et al. (2018))**.** *For fixed $(\theta, \alpha)$, with $\theta > -\alpha$, the set of solutions to the recursive equations for the $V_{n,k}$'s is given by:*

*(i) for $0 \leq \alpha < 1$, mixtures over $\gamma \in \mathbb{R}_+$ of the $V_{n,k}$'s of IBPs, with respect to a mixing distribution $P_\gamma$;*

*(ii) for $\alpha < 0$, mixtures over $M \in \mathbb{N}$ of the $V_{n,k}$'s of BBs, with respect to a mixing distribution $P_M$.*

Hence, any Gibbs-type feature model is obtained by placing a prior distribution on the parameters $\gamma$ or $M$ of the IBP and BB models, respectively. This results draws an elegant parallel between Gibbs-type feature models and Gibbs-type partitions of Gnedin and Pitman (2005): in both frameworks, product form exchangeable distributions emerge as mixtures over simpler, fundamental models. Further analogies will be developed in Section 3.2, where we will show that the parameter $\alpha$ 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 $\alpha$-diversity of Pitman (2003) for feature allocations.

### 2.2.1 Predictive distributions of Gibbs-type models

Predictive distributions in the feature sampling framework play a central role in describing how a sample evolves as the number of individuals increases. For a generic feature frequency model (2.2), these distributions are given in (2.3) and further discussed in point (ii) of Remark 2.2. In the parallel context of SSMs, the general predictive distributions presented in Lemma 1.2 specialize to the Gibbs-type models through the form in (1.6). By contrast, explicit predictive expressions for Gibbs-type feature models have so far been absent from the literature. This lack represents an important open point in the literature, which we resolve in Chapter 3 (Section 3.2.1), through the derivation of general predictive laws valid for any Gibbs-type feature model. In the same section, we also introduce the *buffet metaphor*, a natural extension of the well-known Indian buffet metaphor of Griffiths and Ghahramani (2011), which provides an intuitive interpretation of the sequential sampling mechanism for this broader class of models.

While the general theory is developed in Chapter 3 (Section 3.2.1), the predictive laws for the special cases of BB model and IBP model, introduced in Examples 2.1 and 2.2, are well established in the literature and reported below.

**Example 2.3** (Predictive rules for the BB and IBP models)**.** For the BB model with parameters $(M, \alpha, \theta)$, the predictive rule follows (2.3), with

$$A_{n+1,\ell} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{m_\ell - \alpha}{\theta + n}\right), \quad Y_{n+1} \sim \text{Binomial}\left(M - k, -\frac{\alpha}{\theta + n}\right), \tag{2.8}$$

for $n \geq 0$, where Binomial$(n_0, p)$ denotes a binomial random variable with parameters $n_0 \in \mathbb{N}$ and $p \in (0, 1)$. In this case, the feature probabilities in (2.3) $q_\ell = (m_\ell - \alpha)/(\theta + n)$ are deterministic. Moreover, no new features $Y_{n+1}$ can appear once the maximum number $K_n = M$ has been reached. Because $M$ is fixed a priori, this represents the major limitation of the BB model, as the total number of features in the population is assumed to be known in advance.

For the (three-parameter) IBP with parameters $(\gamma, \alpha, \theta)$, the predictive rule follows (2.3), with

$$A_{n+1,\ell} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{m_\ell - \alpha}{\theta + n}\right), \quad Y_{n+1} \sim \text{Poisson}\left(\gamma \frac{(\theta + \alpha)_n}{(\theta + 1)_n}\right), \tag{2.9}$$

for $n \geq 0$. Here, the number $Y_{n+1}$ of new features in the next subject depends on the initial sample $\boldsymbol{Z}$ only through the sample size $n$, and not on its specific composition. This makes the predictive rule of the IBP rather limited.

Further discussion of the predictive behavior of the BB and IBP models, as well as their role as specific instances of the broader Gibbs-type feature class, is provided in Chapter 3.

### 2.2.2 Unseen feature problems via Gibbs-type models

A central goal in the feature allocation framework is to study the number of distinct features observed in a sample, or expected to appear in future samples. This quantity provides a natural analogue to biodiversity measures in ecological applications. In the context of SSMs, the distribution of the number $K_n$ of distinct species in a sample of size $n$ for a Gibbs-type model is given by (1.9); see (Gnedin and Pitman, 2005; De Blasi et al., 2015). The related unseen species problem, concerning the distribution of $K_m^{(n)}$ defined in (1.1), is addressed by Lijoi et al. (2007), leading to the general expression in (1.10). In the parallel setting of feature sampling, the focus shifts to two analogous quantities under Gibbs-type feature models: (i) the distribution of the number $K_n$ of distinct features observed in a sample of size $n$, which is intended as a prior distribution for such a quantity; (ii) the distribution of $K_m^{(n)}$, defined in (2.1), representing the number of new features which will be observed in an additional sample $(Z_{n+1}, \ldots, Z_{n+m})$ of size $m \geq 1$, given the observed sample $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$. Despite their fundamental importance, these distributions have not been investigated in the existing literature, and no general results analogous to those available for Gibbs-type SSMs have been established. In Chapter 3 (Section 3.2.2) we resolve this important issue by deriving explicit expressions for the distributions of both $K_n$ and $K_m^{(n)}$, which hold true for any Gibbs-type feature model. These results constitute a substantial methodological advancement in the feature allocation literature, providing the first comprehensive analysis of feature diversity and discovery within the Gibbs-type feature framework. In addition, we present novel examples within this class that extend beyond the classical BB and IBP models, demonstrating both the flexibility and the applicability of the Gibbs-type family.

For completeness, we recall below the results for the BB and IBP models introduced in Examples 2.1 and 2.2. To the best of our knowledge, the distribution of $K_m^{(n)}$ for the BB model has not previously appeared in the literature and can be obtained as a special case of the general theory we develop in Section 3.2.2.

**Example 2.4** (Distribution of number of features for the BB and IBP models)**.** For the BB model with parameters $(M, \alpha, \theta)$, it holds that

$$K_n \sim \text{Binomial}\left(M, p_n(\theta, \alpha)\right), \qquad K_m^{(n)} \mid \boldsymbol{Z} \sim \text{Binomial}\left(M - k, p_m(\theta + n, \alpha)\right),$$

where $p_n(\theta, \alpha) = 1 - (\theta + \alpha)_n/(\theta)_n$. As already noted from the predictive rule in (2.8), no new features can appear in additional samples once all the $M$ features have been observed.

For the (three-parameter) IBP with parameters $(\gamma, \alpha, \theta)$, it holds that

$$K_n \sim \text{Poisson}\left(\gamma g_n(\theta, \alpha)\right), \qquad K_m^{(n)} \mid \boldsymbol{Z} \sim \text{Poisson}\left(\gamma\left(g_{n+m}(\theta, \alpha) - g_n(\theta, \alpha)\right)\right),$$

where $g_n(\theta, \alpha)$ is defined in Example 2.2. Consistent with the predictive rule in (2.9), the distribution of $K_m^{(n)}$ depends on the initial sample $\boldsymbol{Z}$ only through the sample size $n$, confirming that the IBP provides limited predictive structure. This behavior mirrors that of the Dirichlet process in the species sampling framework, whose predictive distribution for the number of new species also depends solely on $n$, see Example 1.4.

Further discussion on the BB and IBP models is provided in Chapter 3, together with a detailed asymptotic analysis of the distributions of $K_n$ and $K_m^{(n)}$ for general Gibbs-type models. The asymptotic results recover the known properties of the BB and IBP cases and highlight the broader modeling flexibility offered by the Gibbs-type family.

## 2.3 Sufficientness postulates for feature sampling models

In the context of feature sampling, predictive distributions for unobserved samples play a central role, as they govern the stochastic mechanism driving sequential data evolution. Under a feature frequency model, the predictive distribution is given by (2.3), which assigns positive probability to both previously observed (*old*) and unobserved (*new*) features. This structure is intuitively appealing, since the quantities involved have a clear interpretation, enabling straightforward comparison across different modeling choices. According to (2.3) and point (ii) of Remark 2.2, the prediction rule for $Z_{n+1}$ may depend on the initial sample $\boldsymbol{Z}$ through the sample size $n$ and the whole information in the vector of feature frequencies $\boldsymbol{m} = (m_1, \ldots, m_k)$, that is the number $k$ of distinct observed features and their frequencies $m_\ell$, as a consequence of exchangeability. Specific model assumptions determine how the predictive distribution depends on this information. For instance, as shown in Example 2.3, in the BB model the number $Y_{n+1}$ of new features depends on both $n$ and $k$ (see (2.8)), while in the IBP it depends solely on the sample size $n$ (see (2.9)). This limited dependence of $Y_{n+1}$ on the sample size alone represents a major drawback for the IBP, as it hinders meaningful inference on unseen features because Bayesian estimators in such settings become independent of other sample statistics. However, this limitation is not unique to the IBP model. Indeed, James (2017) showed that the distribution of $Y_{n+1}$ depends solely on the sample size for any CRM prior for $\tilde{\mu}$ in (2.2). Hence, to obtain richer predictive structures, one must move beyond the class of CRMs when specifying priors

for $\tilde{\mu}$. This observation highlights the need to design priors guided by desired predictive properties. More broadly, this viewpoint reflects a growing trend in Bayesian statistics: assigning a system of predictive distributions directly, rather than specifying priors and likelihoods. As argued by Fortini and Petrone (2012) and Fong et al. (2023), reasoning in terms of predictive laws is often more intuitive and transparent than working with priors directly. A systematic characterization of the classes of priors that induce predictive laws with prescribed dependence structures therefore serves a dual purpose: it advances theoretical understanding of complex, often nonparametric, models, and it provides practitioners with principled tools for prior elicitation.

In the species sampling setting, such predictive characterizations are well established and known as *sufficientness postulates* (see Section 1.3). By contrast, analogous results for feature allocation models remain considerably less explored, and only limited contributions are available. Camerlenghi and Favaro (2021) and Camerlenghi et al. (2024) focused on feature models based on scaled process priors (James et al., 2015), a class of random measures obtained via suitable transformations of CRMs. They introduce stronger dependence across the measure weights and, when used in the context of feature models, yield predictive distributions that depend on the entire sampling information. As shown by Camerlenghi et al. (2024), only scaled processes derived from stable subordinators lead to predictive distributions for new features that depend solely on the sample size and the number of distinct features in the sample. While these findings provide valuable insights into scaled processes, they offer limited guidance for broader classes of feature models. Indeed, as we show in Chapter 3, predictions depending on both sample size and the number of distinct features also arise in feature allocation models admitting an EFPF in product form (Battiston et al., 2018). This observation motivates a more general and systematic investigation of sufficientness postulates in the feature sampling framework. We undertake this task in Chapter 4 (Section 4.3) where we develop the first comprehensive treatment of sufficientness postulates for feature allocation models, thereby complementing a foundational concept of species sampling theory within the feature sampling framework.

The discussion so far has focused on how richer predictive structures can emerge from different distributional assumptions on the weights $\tilde{q}_j$ of the random measure $\tilde{\mu}$. A complementary important question, however, concerns the atoms of $\tilde{\mu}$, that is, the feature labels $\tilde{X}_j$. In all feature frequency models defined by (2.2), these labels are assumed to be i.i.d. and independent of the occurrence probabilities $\tilde{q}_j$. As a consequence, the predictive rule (2.3) cannot depend on the observed feature labels $X_\ell$, for $\ell = 1, \ldots, k$. This assumption is natural in standard feature allocation settings considered so far, where labels serve merely as identifiers without intrinsic meaning. However, in many modern applications, such as Bayesian factor analysis and nonnegative matrix factorization (Griffiths and Ghahramani, 2005; Knowles and Ghahramani, 2011; Ayed and Caron, 2021; Zhou et al., 2012, 2016), or spatial models where the atoms $\tilde{X}_j$ encode meaningful quantities like spatial locations, the feature labels themselves carry essential information. In these contexts, it becomes desirable to move beyond the independence assumption, allowing for richer interactions between the feature labels and possibly their occurrence probabilities. We formalize this broader class of models in Chapter 4, introducing the *extended feature models*. Within this generalized framework, predictive distributions can depend not only on $n$ and $\boldsymbol{m}$, but also

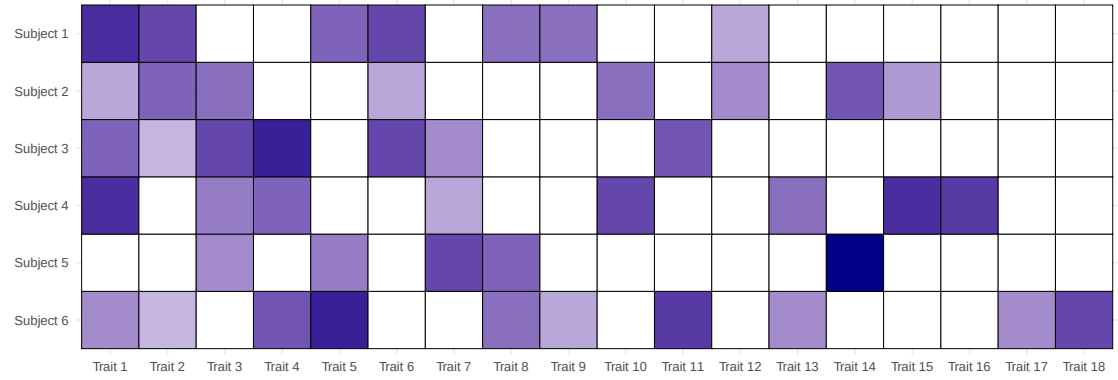



Figure 2.2: Observed data from an exchangeable trait model: matrix of counts $\boldsymbol{A}$, with $n = 6$ subjects and $K_n = 18$ observed traits. Rows and columns are arranged in no particular order. White cells, such as $A_{13} = 0$, indicate the absence of a trait for a given subject, while darker shades of blue represent higher values of the corresponding counts $A_{i\ell} \in \{1, 2, \dots\}$.

on the specific values of feature labels $X_\ell$. We finally remark that the predictive characterizations of Chapter 4 will be developed within the setting of extended feature models, and that the sufficientness postulates for standard feature frequency models emerge naturally as a special case.

## 2.4 From features to traits: extending the feature allocation framework

Trait allocation models (James, 2017; Campbell et al., 2018) generalize feature allocation models by allowing each subject's characteristics, called *traits*, to be associated with non-negative values. In this broader setting, the presence of a trait is coupled with a quantitative measurement, often reflecting its strength, abundance, or degree of expression. Notable examples of trait allocation models are described in Broderick et al. (2015).

We begin by reviewing the trait allocation setting (Campbell et al., 2018). In this framework, we observe $n$ subjects and $K_n = k$ distinct traits. The data can be represented by an $n \times k$ matrix $\boldsymbol{A}$, where each entry $A_{i\ell} \in \{0, 1, 2, \dots\}$ denotes the count of the $\ell$th trait (column) for the $i$th subject (row), as illustrated in Figure 2.2. Each trait is associated with a distinct *label* $X_\ell \in \mathbb{X}$, where $\mathbb{X}$ denotes the space of the trait labels. As in the feature setting, it is typically assumed that each subject exhibits only finitely many traits, ensuring that the total number of distinct traits $K_n$ is finite in any given sample. The trait framework thus provides a natural generalization of the feature setting introduced in Section 2.1: binary indicators $A_{i\ell} \in \{0, 1\}$ are replaced by nonnegative integer counts $A_{i\ell} \in \{0, 1, 2, \dots\}$.

From a modeling perspective, the trait allocation framework closely parallels that of feature allocations (Campbell et al., 2018). In the feature setting, we focus on the subclass of infinite exchangeable models known as *feature frequency models* (see Section 2.1). Analogously, in the trait setting, the natural counterpart is the class of *frequency models*, a generalization of feature frequency models that allows for non-negative measurements. In

these models, each subject $i$ is represented by a random counting measure $Z_i$ on $\mathbb{X}$, namely

$$Z_i(\,\cdot\,) = \sum_{j\geq 1} \tilde{A}_{ij}\delta_{\tilde{X}_j}(\,\cdot\,), \qquad i \geq 1, \tag{2.10}$$

where the $\tilde{X}_j$'s denote the (possibly infinite) set of trait labels, and the random variables $\tilde{A}_{ij} \in \{0, 1, 2, \dots\}$ represent the abundance of trait $\tilde{X}_j$ in subject $i$. Moreover, the defining assumption of *frequency models* is that, conditionally to a sequence of parameters $(\theta_j)_{j\geq 1}$, the random variables $\tilde{A}_{ij}$ are i.i.d. across subjects for fixed $j$, and independent across traits. Formally,

$$\tilde{A}_{ij} \,|\, \theta_j \overset{\text{iid}}{\sim} P(\cdot\,;\theta_j), \qquad i \geq 1, j \geq 1, \tag{2.11}$$

where $P(\cdot\,;\theta)$ is a parametric distribution supported on the non-negative integers, such as a Poisson distribution, depending on a positive parameter $\theta > 0$. The parameters $(\theta_j)_{j\geq 1}$ can be organized in a discrete measure $\tilde{\mu}$ on $\mathbb{X}$, defined as

$$\tilde{\mu}(\,\cdot\,) = \sum_{j\geq 1} \theta_j \delta_{\tilde{X}_j}(\,\cdot\,). \tag{2.12}$$

Note that the atoms of the discrete measure $\tilde{\mu}$, i.e., the trait labels $\tilde{X}_j$, are common across all subjects, so that the same traits are allowed to be observed in multiple subjects. Summarizing, the full Bayesian specification is

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \text{CP}(\tilde{\mu}), \qquad i \geq 1, \\ \tilde{\mu} &\sim \mathcal{Q}, \end{aligned} \tag{2.13}$$

which means that $Z_i$ in (2.10) are i.i.d. from a *process of counts* (CP) with parameter $\tilde{\mu}$ defined by (2.11)-(2.12). Here $\mathcal{Q}$ denotes the de Finetti measure, i.e., the prior distribution of the random measure $\tilde{\mu}$. In the special case of binary traits, the frequency model (2.13) reduces to the feature frequency model in (2.2).

To conclude this brief review of the trait models, we present two relevant examples for the distribution $P(\cdot;\theta)$.

**Example 2.5** (Exchangeable binary traits)**.** When the count measurements $\tilde{A}_{ij}$ are binary, the trait allocation framework reduces to the special case of feature allocation framework of Chapter 2. Thus, we let $\tilde{A}_{ij} \,|\, \theta_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta_j)$ for $i \geq 1$ and fixed $j$, with success probabilities $\theta_j \in (0, 1]$, so that $Z_i \,|\, \tilde{\mu} \overset{\text{iid}}{\sim} \text{CP}(\tilde{\mu})$ are i.i.d. Bernoulli processes (see Section 2.1).

**Example 2.6** (Exchangeable Poisson counts)**.** When the data $\tilde{A}_{ij}$ take values in $\{0, 1, 2, \dots\}$, as those depicted in Figure 2.2, a natural choice is $P(a;\theta) = (a!)^{-1}\theta^a e^{-\theta}$ for $a \in \{0, 1, \dots\}$, that is a Poisson distribution with mean $\theta > 0$. In other words, we assume $\tilde{A}_{ij} \,|\, \theta_j \overset{\text{iid}}{\sim} \text{Poisson}(\theta_j)$ for $i \geq 1$ and fixed $j$. This specification has been less explored compared to the binary case.

## 2.5 Extending the feature and trait frameworks to multiple populations

Most existing works on feature and trait allocation models assume full exchangeability among subjects, implying a form of homogeneity across all observations. This has been

the standing assumption throughout Chapter 2. However, in many applied areas, such as multi-study analyses, biological experiments across environments or patient cohorts, this assumption can be restrictive. In such contexts, data naturally arise from multiple related but distinct populations, for which a weaker assumption of *partial exchangeability* is more appropriate. Under partial exchangeability, observations are divided into groups that are exchangeable within but not across groups. This framework allows for dependence and heterogeneity across subpopulations while preserving exchangeability within each. The notion of partial exchangeability has a long tradition in Bayesian nonparametrics, especially in the modeling of dependent species populations (see Section 1.5).

The aim of this section is to recall the general framework of partial exchangeability in the trait allocation setting and to position our contributions within this context. We do not provide the full technical background here; rather, we introduce the setting in which our new methodological contributions are developed, leaving their formal treatment to Chapter 5. Analogously to the exchangeable case for the sequence $(Z_i)_{i\geq 1}$ in (2.13), partial exchangeability can be achieved by introducing a vector of (dependent) random measures $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$, such that

$$Z_{iq} \,|\, \tilde{\mu}_1, \ldots, \tilde{\mu}_d \overset{\text{ind}}{\sim} \text{CP}(\tilde{\mu}_q), \qquad i \geq 1, q = 1, \ldots, d,$$

where $Z_{iq}$ denotes observation $i$ in group $q$, and $d$ is the number of groups. From a Bayesian perspective, specifying a partially exchangeable model corresponds to placing a prior on the vector of dependent random measures. Introducing dependence among the group-specific random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ enables borrowing of information across subpopulations, but also raises the challenge of constructing priors that are both computationally and analytically tractable. Only a few recent contributions have investigated this setting for feature and trait allocation models. For instance, Masoero et al. (2018); Beraha and Favaro (2025); James et al. (2024) introduced hierarchical constructions to induce dependence among groups, while Shen et al. (2024) proposed a bivariate beta process. In contrast, in the related literature on species sampling, countless models have been developed; see Section 1.5 for a review.

In this thesis, we extend this line of research by developing a general and tractable class of Bayesian nonparametric priors for partially exchangeable trait allocation models, based on *completely random vectors*. The detailed formulation and a comprehensive theoretical analysis of these models are deferred to Chapter 5, where we pursue inferential goals such as: (i) estimation of trait- and group-specific parameters, and (ii) estimation of the number of unseen traits in the observed sample. In particular, we provide closed-form expressions for marginal and posterior distributions, and illustrate the tractability of our framework in the cases of binary (Example 2.5) and Poisson-distributed traits (Example 2.6). Building on these results, we also develop a novel mixture model that infers the group partition structure from the data, effectively clustering trait allocations. This extension can be viewed as a generalization of Bayesian nonparametric latent class models, and we show that it effectively mitigates the systematic overclustering that arises when the number of traits is fixed. Finally, we illustrate the usefulness of our approach through an application to the 'Ndrangheta criminal network from the *Operazione Infinito* investigation, where our model provides insights into the organization of illicit activities.

# 3. Bayesian analysis of product feature allocation models

In this chapter, 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 and Gorur, 2009) and the beta Bernoulli (BB, Griffiths and Ghahramani, 2005). 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 behavior. 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 (Gnedin and Pitman, 2005; De Blasi et al., 2015). 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 and Yor, 1997), the normalized generalized gamma process (Lijoi et al., 2007), and mixtures of Dirichlet multinomial processes (Gnedin, 2010; De Blasi et al., 2013). 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 and 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 and Colwell, 2001). A Bayesian nonparametric inferential framework for predicting unseen species has been laid down by Lijoi et al. (2007) 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. Colwell et al., 2012; Chiu et al., 2014; Chao et al., 2014; Chakraborty et al., 2019; Chiu, 2022, 2023), the Bayesian nonparametric literature remains much more limited, except for the recent works of Masoero et al. (2022); 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 chapter, particularly relevant for ecological applications, is the definition of the $\alpha$-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 $\alpha$-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 (Ionita-Laza et al., 2009; Gravel, 2014; 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 and Mizukami, 2021) and microbiome analysis (Sanders et al., 2019).

The chapter is structured as follows. Section 3.1 provides a concise review of Gibbs-type feature allocation models. Although this material is discussed in greater detail in Chapter 2, we include a brief overview here to make the chapter self-contained. In Section 3.2 we develop general theory for the class of Gibbs-type feature models. In Sections 3.3-3.4 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 3.5, while Section 3.6 illustrates our methodology by analyzing two real datasets. The chapter ends with a discussion; proofs, additional theorems, simulation studies and additional details about the applications are collected in the Appendix.

## 3.1 Review on Gibbs-type Feature allocation models

### 3.1.1 Exchangeable Gibbs-type feature allocation models

Among the exchangeable and consistent models (defined in Section 2.1), 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 and 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

$$\pi_n(m_1,\ldots,m_k) = V_{n,k}\prod_{\ell=1}^{k}(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}, \tag{3.1}$$

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 recurrence relationship in (2.6). Let $M_{n,\ell}$ denote the number of individuals displaying feature $\ell$ in a sample of size $n$. 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 $(\gamma,\alpha,\theta)$ satisfying $\gamma > 0$, $0 \le \alpha < 1$, and $\theta > -\alpha$. The EFPF is in product form (3.1) and the $V_{n,k}$'s are given by

$$V_{n,k} = \frac{1}{k!}\left\{\frac{\gamma}{(\theta+1)_{n-1}}\right\}^{k}\exp\left\{-\gamma g_n(\theta,\alpha)\right\}, \quad \text{with} \quad g_n(\theta,\alpha) := \sum_{i=1}^{n}\frac{(\theta+\alpha)_{i-1}}{(\theta+1)_{i-1}}. \tag{3.2}$$

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 $(M,\alpha,\theta)$ such that $M \in \mathbb{N}$, $\alpha < 0$ and $\theta > -\alpha$ (Griffiths and Ghahramani, 2011). The EFPF of a BB is also in product form (3.1), with the $V_{n,k}$'s given by

$$V_{n,k} = \binom{M}{k}\left\{\frac{-\alpha}{(\theta+\alpha)_n}\right\}^{k}\left\{\frac{(\theta+\alpha)_n}{(\theta)_n}\right\}^{M}\mathbb{1}_{\{0,1,\ldots,M\}}(k), \tag{3.3}$$

where $\mathbb{1}_C$ denotes the indicator function of a set $C$. The BB model prescribes that the observed number of features $K_n$ is bounded by $M$. 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 $(\theta,\alpha)$, the set of Gibbs coefficients $V_{n,k}$ satisfying the aforementioned consistency condition are necessarily mixtures of the $\gamma$ and $M$ parameters of the IBP and the BB, respectively. This is better clarified in the following result.

**Proposition 3.1** (Theorem 1.1 of Battiston et al. (2018))**.** *For fixed values of $(\theta,\alpha)$ such that $\theta > -\alpha$, the set of solutions of the recursions for the $V_{n,k}$'s is:*

*(i) for $0 \le \alpha < 1$, mixtures over $\gamma \in \mathbb{R}^+$ of the $V_{n,k}$'s of IBPs with respect to a distribution $P_\gamma$;*

*(ii) for $\alpha < 0$, mixtures over $M \in \mathbb{N}$ of the $V_{n,k}$'s of* BBs *with respect to a distribution $P_M$.*

Hence, any Gibbs-type feature model is obtained by considering a prior distribution for the $\gamma$ and $M$ 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 3.2, where we will show that the parameter $\alpha$ 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 $\alpha$-diversity of Pitman (2003) for feature allocations.

### 3.1.2 Hierarchical representations and random measures

Like all feature frequency models, Gibbs-type feature models admit a hierarchical representation in terms of Bernoulli processes (BPs) and random measures, which we now recall. (see Section 2.1 for a detailed introduction). Let $(\tilde{X}_j)_{j\geq 1}$ denote the sequence of all possible feature labels, where $\tilde{X}_j \overset{\text{iid}}{\sim} G_0$ for any $j \geq 1$. The $i$th individual is represented by a counting measure $Z_i$ on the space $\mathbb{X}$, which is given by

$$Z_i(\cdot) = \sum_{j\geq 1} \tilde{A}_{ij}\delta_{\tilde{X}_j}(\cdot), \tag{3.4}$$

where $\tilde{A}_{ij} = 1$ if the $i$th individual exhibits feature $\tilde{X}_j$, and $\tilde{A}_{ij} = 0$ otherwise. The feature labels $X_1, \ldots, X_{K_n}$ observed in a sample of $n$ individuals are a subset of the sequence $(\tilde{X}_j)_{j\geq 1}$. 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 $A_{i\ell} = Z_i(\{X_\ell\})$. Moreover, the variables $\tilde{A}_{ij}$ are conditionally independent Bernoulli random variables given a sequence of random probabilities $(\tilde{q}_j)_{j\geq 1}$, that is, $\tilde{A}_{ij} \,|\, \tilde{q}_j \overset{\text{ind}}{\sim} \text{Bernoulli}(\tilde{q}_j)$. We organize these probabilities through a random measure $\tilde{\mu}$ on $\mathbb{X}$, namely

$$\tilde{\mu}(\cdot) := \sum_{j\geq 1} \tilde{q}_j\delta_{\tilde{X}_j}(\cdot). \tag{3.5}$$

Conditionally on $\tilde{\mu}$, the $Z_i$'s are i.i.d. draws from a Bernoulli process with base measure $\tilde{\mu}$, written $Z_i \,|\, \tilde{\mu} \overset{\text{iid}}{\sim} \text{BP}(\tilde{\mu})$ for any $i \geq 1$. Summarizing, the following hierarchical representation holds:

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \text{BP}(\tilde{\mu}), \qquad i \geq 1, \\ \tilde{\mu} &\sim \mathscr{Q}, \end{aligned} \tag{3.6}$$

where $\mathscr{Q}$ is the *prior* distribution of $\tilde{\mu}$, i.e., the *de Finetti measure*. Equation (3.6) defines a feature frequency model. In particular, both the BB and IBP models admit this hierarchical construction: in the BB model, this follows directly with the random measure denoted $\tilde{\mu} \,|\, M$, while for the IBP model the representation arises from the seminal works of Thibaux and Jordan (2007); Teh and Gorur (2009), with random measure denoted $\tilde{\mu} \,|\, \gamma$. Thanks to Proposition 3.1, the law of $\tilde{\mu}$ for any Gibbs-type feature model is a mixture over $\gamma$ or $M$ of the corresponding law for the IBP or BB model. Therefore, all Gibbs-type feature models are indeed feature frequency models.

We now provide details on this hierarchical construction for the BB and IBP models. Let us firstly consider the BB model with parameters $(M, \alpha, \theta)$, in which there are $M$ possible features $\tilde{X}_1, \ldots, \tilde{X}_M$. The hierarchical representation of the beta Bernoulli process is straightforward as we have $\tilde{\mu} \,|\, M = \sum_{j=1}^{M} \tilde{q}_j \delta_{\tilde{X}_j}$, with $\tilde{q}_j \overset{\text{iid}}{\sim} \text{Beta}(-\alpha, \theta+\alpha)$ and $\tilde{X}_j \overset{\text{iid}}{\sim} G_0$, for $j = 1, \ldots, M$, recalling that $\alpha < 0$ and $\theta > -\alpha$. See Lemma 3.4 in the Appendix 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

$$\mathsf{E}[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}] = \exp\left\{-\int_{\mathbb{X}}\int_{0}^{\infty}\left[1 - e^{-sf(x)}\right]\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\},$$

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 $G_0$ is a diffuse distribution on $\mathbb{X}$ from which the labels $\tilde{X}_j$ are sampled. We will write $\tilde{\mu} \sim \text{CRM}(\rho; G_0)$. 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 (3.6) 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 $\tilde{\mu} \,|\, \gamma$ is distributed as a completely random measure and, more precisely, it follows a stable-beta process, whose intensity measure $\rho(\mathrm{d}s)$ is

$$\rho(\mathrm{d}s) = \gamma \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}s. \tag{3.7}$$

Note that $\gamma$ is sometimes called the *total mass* parameter because $\gamma = \mathsf{E}[\tilde{\mu}(\mathbb{X})] = \sum_{j\geq 1} \mathsf{E}[\tilde{q}_j]$ 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 and 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).

## 3.2 Predictive structure of Gibbs-type feature models

### 3.2.1 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 3.1.2, we study the conditional distribution of $Z_{n+1}$, given a random sample $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, 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 $\boldsymbol{Z}$; (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 $X_\ell$ 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

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

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 3.1** (Predictive law)**.** *Suppose the* EFPF *is in product form* (3.1)*, 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 $\boldsymbol{Z}$, 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 (3.1). In the second place, the count $Y_{n+1}$ of new features depends on the sample $\boldsymbol{Z}$ 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 3.3–3.4.

As an immediate consequence of Theorem 3.1, 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 and 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 3.1** (Sample coverage)**.** *Suppose the* EFPF *is in product form* (3.1)*, then the probability that $Z_{n+1}$ does not show any new features, given $\boldsymbol{Z}$, is*

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

The predictive distribution presented in Theorem 3.1 can be likened to the Indian buffet metaphor (Griffiths and 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 3.1. 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

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

and $K_1 = Y_1$. The $K_1$ selected dishes are associated with labels $X_\ell$, for $\ell = 1, \ldots, K_1$, provided that $K_1 > 0$. Then, the $(n+1)$th customer enters and selects dishes in two steps.

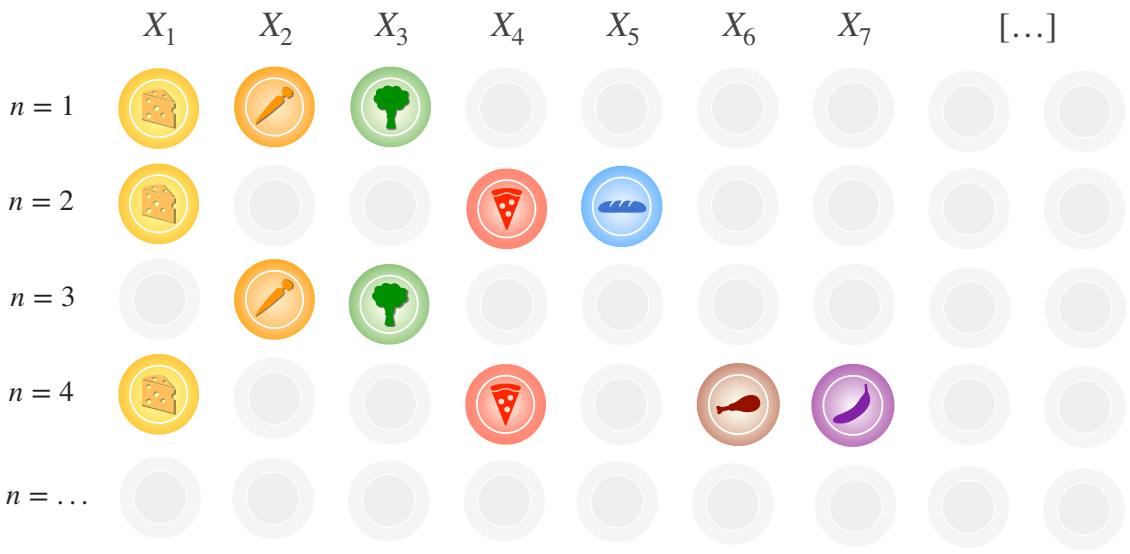



Figure 3.1: The buffet metaphor for Gibbs-type feature models. In this example there are $n = 4$ customers and $K_n = 7$ dishes. The observed frequencies of the dishes are $(m_1, \ldots, m_7) = (3, 2, 2, 2, 1, 1, 1)$. The numbers of new dishes picked by each customer (from top to bottom) are $Y_1 = 3$, $Y_2 = 2$, $Y_3 = 0$, and $Y_4 = 2$.

Firstly, the customer picks $Y_{n+1}$ new dishes (not chosen by the previous $n$ customers) according to the distribution

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

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 $X_\ell$, for $\ell = K_n + 1, \ldots, K_{n+1}$. Secondly, 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

$$A_{n+1,\ell} \,|\, \boldsymbol{Z} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{m_\ell - \alpha}{\theta + n}\right), \tag{3.8}$$

where $m_\ell$ corresponds to the number of previous customers who selected dish $X_\ell$. Higher values of $m_\ell$ correspond to a higher probability of dish $X_\ell$ 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 $\theta$ and $\alpha$. Essentially, the posterior probability of observing the $\ell$th 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 $\ell = 1, \ldots, k$:

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

where $\hat{p}_\ell = m_\ell / n$ denotes the empirical frequency of the $\ell$th observed feature. When $\alpha < 0$, the quantity $-\alpha/\theta \in (0, 1)$ can be conveniently interpreted as the prior frequency of a feature, while $\theta$ 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 $\alpha \in [0, 1)$, $\theta$

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 $Y_1 \,|\, \gamma \sim \text{Poisson}(\gamma)$ and $Y_{n+1} \,|\, K_n, \gamma \sim \text{Poisson}\,(\gamma(\theta+\alpha)_n/(\theta+1)_n)$ for $n \geq 1$. Moreover, as can be verified via Theorem 3.1 using $V_{n,k}$ from (3.3), the BB model leads to $Y_1 \,|\, M \sim \text{Binomial}\,(M, -\alpha/\theta)$ and $Y_{n+1} \,|\, K_n, M \sim \text{Binomial}\,(M-k, -\alpha/(\theta+n))$, 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 $K_n = M$ features are observed.

### 3.2.2 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 $\mathsf{E}(K_1), \ldots, \mathsf{E}(K_n)$ should be regarded as a model-based *rarefaction* curve (Gotelli and 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. (2007). 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 3.2** (Number of features $K_n$)**.** *Suppose the EFPF is in product form* (3.1) *and let $\alpha < 0$. Then, the number of distinct features observed in a sample of $n$ individuals, $K_n$, satisfies, for any $k \geq 0$,*

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

*Alternatively, let $g_n(\theta, \alpha)$ be defined as in* (3.2)*, if $\alpha \in [0, 1)$ then for any $k \geq 0$,*

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

A preliminary version of this result is also presented in Battiston et al. (2018), up to some further simplifications. In Sections 3.3–3.4, we will specialize the formulas of Theorem 3.2 for specific choices of $V_{n,k}$'s, recovering well-known distributions. In Proposition 3.9 of the Appendix, 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 3.2 and Proposition 3.9 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 $\boldsymbol{Z}$, i.e., *a posteriori*. In ecological contexts, this leads to the *extrapolation*

of the accumulation curve (Zito et al., 2023), namely the expectations $\mathsf{E}(K_{n+m} \mid \boldsymbol{Z}) = k + \mathsf{E}(K_m^{(n)} \mid \boldsymbol{Z})$ for any $m \geq 1$. What follows is a key result of this work since we provide the distribution of $K_m^{(n)}$ for any Gibbs-type feature model, analogous to the main finding of Lijoi et al. (Proposition 1, 2007) for Gibbs-type species sampling models.

**Theorem 3.3** (Number of hitherto unobserved features)**.** *Suppose the* EFPF *is in product form* (3.1)*. Then, the distribution of* $K_m^{(n)} \mid \boldsymbol{Z}$ *is such that, for any* $y \geq 0$*,*

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

*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 3.3, it is evident that the distribution of $K_m^{(n)}$ depends on the data $\boldsymbol{Z}$ only through the sample size $n$ and the number of distinct features $K_n$, but not on the feature counts $m_\ell$. 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., 2007). We refer once again to Sections 3.3–3.4 for tractable special cases of Theorem 3.3.

### 3.2.3 Asymptotic behavior and $\alpha$-diversity

The $\alpha$ parameter of a Gibbs-type feature model plays a key role, as hinted by Proposition 3.1. We show that $\alpha$ identifies the asymptotic behavior 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 behavior is well known for BB and IBP models (e.g. Griffiths and Ghahramani, 2011; Teh and Gorur, 2009), but it is in fact a general property of Gibbs-type feature models, as the following proposition illustrates.

**Proposition 3.2** ($\alpha$-diversity)**.** *Suppose the* EFPF *is in product form* (3.1) *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* $K_n \xrightarrow{\mathrm{d}} M$*,*

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

*(iii) if* $0 < \alpha < 1$*, then* $K_n / n^\alpha \xrightarrow{\mathrm{d}} \gamma\, \Gamma(\theta+1) / \{\alpha \Gamma(\theta+\alpha)\}$*,*

*where the random variables* $M$ *and* $\gamma$ *have distributions* $P_\gamma$ *and* $P_M$ *as in the mixture representation of the weights* $V_{n,k}$*'s described in Proposition 3.1.*

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, $K_n/c_\alpha(n) \xrightarrow{\mathrm{d}} S_\alpha$, as $n \to \infty$. We call random variable $S_\alpha$ the *$\alpha$-diversity* of a feature allocation model, analogous to the $\alpha$-diversity introduced by Pitman (2003). The random variable $S_\alpha$ is often of direct interest in ecological problems, as it represents a synthetic biodiversity measurement. Naturally, comparing $\alpha$-diversities across different datasets makes sense only if they are based on the same growth rate. Note that, for fixed values of $\alpha$ and $\theta$, in the BB and IBP models the $\alpha$-diversity is deterministic. In practice, the $\alpha$-diversity is unknown, and it may be estimated employing a prior distribution for $M$ or $\gamma$, which results in a Gibbs-type feature model thanks to Proposition 3.1. Moreover, the posterior law of $\gamma$ and $M$ may be obtained by combining the prior density $p_\gamma(\mathrm{d}\gamma)$ associated to $P_\gamma$, or the prior probability distribution $p_M(y)$ associated to $P_M$, with the EFPF of equations (3.2)-(3.3), giving respectively

$$\begin{aligned} p_\gamma(\mathrm{d}\gamma \,|\, \boldsymbol{Z}) &\propto p_\gamma(\mathrm{d}\gamma)\gamma^k \exp\left\{-\gamma g_n(\theta,\alpha)\right\}, \\ p_M(y \,|\, \boldsymbol{Z}) &\propto p_M(y)\frac{y!}{(y-k)!}\left(\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^y, \end{aligned} \tag{3.9}$$

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

**Proposition 3.3** (Posterior law of the $\alpha$-diversity)**.** *Suppose the EFPF is in product form* (3.1)*. Let $K_m^{(n)} \,|\, \boldsymbol{Z}$ be the number of hitherto unseen features and let $S_\alpha$ be the $\alpha$-diversity. Then, as $m \to \infty$*

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

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

### 3.2.4 Posterior characterization

We conclude our theoretical investigation with another pivotal result: the determination of the posterior distribution arising from model (3.6) of $\tilde{\mu} = \sum_{j\geq 1} \tilde{q}_j \delta_{\tilde{X}_j}$, given $\boldsymbol{Z}$. 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 3.3-3.4, where we discuss specific choices for the priors of $\gamma$ and $M$.

**Theorem 3.4** (Posterior distribution of $\tilde{\mu}$)**.** *Suppose $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ follows model (3.6), then the posterior distribution of $\tilde{\mu}$, given $\boldsymbol{Z}$, satisfies the following decomposition*

$$\tilde{\mu} \,|\, \boldsymbol{Z} \stackrel{d}{=} \mu' + \mu^*, \tag{3.10}$$

*where $\mu^* \stackrel{d}{=} \sum_{\ell=1}^{k} q_\ell \delta_{X_\ell}$ is a random measure such that $q_\ell \stackrel{ind}{\sim} \text{Beta}(m_\ell - \alpha, \alpha + \theta + n - m_\ell)$, for $\ell = 1, \ldots, k$, and $X_1, \ldots, X_k$ denote the observed features. Moreover, the random measure $\mu'$ in (3.10) is independent of $\mu^*$, and its distribution depends on the value of $\alpha$, as specified below.*

*(i) If $\alpha < 0$, then the random measure $\mu' \,|\, M' \stackrel{d}{=} \sum_{j=1}^{M'} q'_j \delta_{\tilde{X}'_j}$, where $q'_j \stackrel{iid}{\sim} \text{Beta}(-\alpha, \theta + \alpha + n)$ and $\tilde{X}'_j \stackrel{iid}{\sim} G_0$, for $j = 1, \ldots, M'$. Moreover, $M' + k \stackrel{d}{=} M \,|\, \boldsymbol{Z}$ as in (3.9).*

*(ii) If $\alpha \in [0, 1)$, then the random measure $\mu' \,|\, \gamma' \sim \text{CRM}(\rho'; G_0)$ 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, $\gamma' \stackrel{d}{=} \gamma \,|\, \boldsymbol{Z}$ as in (3.9).*

The previous distributional equality (3.10) decomposes the posterior distribution of $\tilde{\mu}$ into two parts: $\mu'$ accounts for the newly observed features, while $\mu^*$ deals with those previously observed. Regarding the latter, the $(n+1)$th individual exhibits an existing feature $X_\ell$ if $A_{n+1,\ell} = 1$, where each $A_{n+1,\ell} \,|\, q_\ell \stackrel{\text{ind}}{\sim} \text{Bernoulli}(q_\ell)$ and $q_\ell \stackrel{\text{ind}}{\sim} \text{Beta}(m_\ell - \alpha, \alpha + \theta + n - m_\ell)$. By integrating out the $q_\ell$'s, we obtain $A_{n+1,\ell} \,|\, \boldsymbol{Z} \stackrel{\text{ind}}{\sim} \text{Bernoulli}\left((m_\ell - \alpha)/(\theta + n)\right)$, consistent with the predictive structure (3.8). It is worth highlighting that the distribution of $\mu^*$ remains the same across all Gibbs-type feature models. However, $\mu'$ 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 $\mu' \,|\, \gamma' \sim \text{CRM}(\rho'; G_0)$ 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 $M \,|\, \boldsymbol{Z}$ or $\gamma \,|\, \boldsymbol{Z}$, which typically follow highly tractable distributions. Then, conditioned on $M$ or $\gamma$, both random measures $\mu'$ and $\mu^*$ are simple to sample from: $\mu^*$ involves a finite number of beta random variables, as does $\mu'$ when $\alpha < 0$. Even simulating $\mu'$ when $\alpha \in [0, 1)$ is straightforward, as due to conjugacy, $\mu'$ follows a stable-beta process, for which efficient sampling algorithms exist; see, for example, Teh and Gorur (2009).

## 3.3 Gamma mixture of Indian buffet processes

### 3.3.1 Predictive structure, number of features, and $\alpha$-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 $\gamma$. Upon examining the posterior distribution in (3.9), 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 (3.2) with respect to the prior density, follows a product form (3.1) and the weights are

$$V_{n,k} = \frac{1}{k!} \frac{b^a \, (a)_k}{\{(\theta+1)_{n-1}\}^k \{b + g_n(\theta, \alpha)\}^{a+k}}. \tag{3.11}$$

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 (3.11). In particular, a stable-beta scaled process with parameters $(\alpha, c, d)$ 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 (3.11), utilizing the general findings from Section 3.2. The predictive laws of the IBP and the gamma mixture of IBPs follow by specializing Theorem 3.1, substituting the $V_{n,k}$'s of equations (3.2) and (3.11), respectively, into the general formulas. As discussed in Section 3.2.1, 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 (3.8), 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 $\mathscr{N}(y; n_0, \mu_0) \propto p^{n_0}(1-p)^y$ is defined for any $y \in \mathbb{N}_0$, with $p = n_0/(\mu_0 + n_0)$, so that $\mathsf{E}(Y) = \mu_0$ and $\text{Var}(Y) = \mu_0 + \mu_0^2/n_0$. Moreover, let $Y_{n+1} \,|\, K_n, \gamma$ be the number of new features for the $(n+1)$th individual in the IBP case, and let $Y_{n+1} \,|\, K_n$ be the same quantity for the gamma mixture model. Then, simple calculus yields

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

Additional distributional properties can be derived by specializing the results of Section 3.2 for the IBP and gamma mixture of IBPs. We summarize some of these properties in the following proposition and refer to the Appendix for additional findings and simplifications, such as the distribution of the number of shared features $K_{n,r}$ or the sample coverage.

**Proposition 3.4.** *Suppose the EFPF is in product form* (3.1) *with* $\alpha \in [0, 1)$. *Let* $K_n \,|\, \gamma$ *and* $K_n$ *be the number of features observed in a sample of* $n$ *individuals for the IBP and the gamma mixture in* (3.11). *Then, we have*

$$K_n \,|\, \gamma \sim \text{Poisson}\left(\gamma g_n(\theta, \alpha)\right), \qquad K_n \sim \text{NegBinomial}\left(a, (a/b) g_n(\theta, \alpha)\right). \tag{3.12}$$

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

$$K_m^{(n)} \,|\, \boldsymbol{Z}, \gamma \sim \text{Poisson}\left(\gamma \left(g_{n+m}(\theta, \alpha) - g_n(\theta, \alpha)\right)\right), \tag{3.13}$$

*whereas for the gamma mixture*

$$K_m^{(n)} \,|\, \boldsymbol{Z} \sim \text{NegBinomial}\left(a+k, \frac{a+k}{b+g_n(\theta,\alpha)} \left(g_{n+m}(\theta, \alpha) - g_n(\theta, \alpha)\right)\right). \tag{3.14}$$

The results regarding the IBP have been deduced from the general theory outlined in Theorem 3.2 and Theorem 3.3, 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 $K_m^{(n)} \mid \boldsymbol{Z}, \gamma$ 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 $\mathsf{E}(K_n \mid \gamma) = \gamma g_n(\theta, \alpha)$ and $\mathsf{E}(K_n) = (a/b) g_n(\theta, \alpha)$. 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 and Yor, 1997) with parameters $(\alpha, \theta)$. 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 $K_m^{(n)} \mid \boldsymbol{Z}, \gamma$, 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 $K_m^{(n)} \mid \boldsymbol{Z}$ 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:

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

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

Recall that, as stated in Proposition 3.2, the parameter $\alpha$ 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 $\gamma$, and hence the $\alpha$-diversity, is deterministic (Masoero et al., 2022). Conversely, by definition, $\gamma$ follows a priori a Gamma$(a, b)$ in the gamma mixture model (3.11), which allows the Bayesian learning of the $\alpha$-diversity through its posterior.

**Proposition 3.5.** *Suppose the EFPF is in product form* (3.1) *with* $\alpha \in [0, 1)$ *and the* $V_{n,k}$*'s defined as in* (3.11)*, so that a priori* $\gamma \sim \mathrm{Gamma}(a, b)$*. Then, the posterior is* $\gamma \mid \boldsymbol{Z} \sim \mathrm{Gamma}(a + k, b + g_n(\theta, \alpha))$ *and therefore the* $\alpha$*-diversity, for* $\alpha = 0$*, is given by*

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

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

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

A consequence of the deterministic $\alpha$-diversity in the IBP is that the width of the credible intervals for $K_m^{(n)}$ grows at a rate slower than $m^{\alpha}$. In contrast, the mixtures of IBPs yield larger credible intervals, whose widths grow proportionally to $m^{\alpha}$, as highlighted by the previous proposition. This difference can be observed in simulation study B of the

Appendix: Figure 3.D.8 suggests that the IBP underestimates the uncertainty in predicting the number of unseen features. Proposition 3.5 presents some of the first results concerning the posterior distribution of the $\alpha$-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).

### 3.3.2 Posterior characterizations and negative binomial processes

We specialize here the posterior characterization of Theorem 3.4 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 $G_0$ on $\mathbb{X}$, if $\tilde{\mu}$ has Laplace functional

$$\mathsf{E}[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}] = \left\{1 + \int_{\mathbb{X}}\int_0^{\infty} \left[1 - e^{-sf(x)}\right] \rho(\mathrm{d}s) G_0(\mathrm{d}x)\right\}^{-a}, \tag{3.15}$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}_+$. We will write $\tilde{\mu} \sim \text{NB}(a, \rho; G_0)$ 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

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \text{BP}(\tilde{\mu}), \qquad\qquad i \geq 1, \\ \tilde{\mu} &\sim \text{NB}(a, \rho; G_0), \end{aligned} \tag{3.16}$$

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 3.B, 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 3.4 characterizes the posterior distribution of the process $\tilde{\mu}$, given $\boldsymbol{Z}$, in terms of negative binomial random measures.

**Corollary 3.2.** *Suppose $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ follows model (3.16), then the posterior distribution of $\tilde{\mu}$, given $\boldsymbol{Z}$, satisfies the decomposition $\tilde{\mu} \,|\, \boldsymbol{Z} \overset{d}{=} \mu' + \mu^*$ in (3.10), where $\mu'$ and $\mu^*$ are independent random measures such that $\mu^*$ is distributed as in Theorem 3.4, whereas $\mu' \sim \text{NB}(a + k, \rho'; G_0)$, with updated intensity $\rho'(\mathrm{d}s) = 1/\{b + g_n(\theta, \alpha)\}\Gamma(1+\theta)/\{\Gamma(1-\alpha)\Gamma(\theta+\alpha)\}s^{-\alpha-1}(1-s)^{n+\theta+\alpha-1}\mathrm{d}s$.*

## 3.4 Gibbs-type feature models with finitely many features

### 3.4.1 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); De Blasi et al. (2013). In feature allocation models, a similar role is played by mixtures of BB models with $M$ features, which corresponds to the $\alpha < 0$

case. These feature allocation models assume a finite number of features $M$, representing the *richness* in ecological problems. The standard BB model assumes that $M$ 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) $M$ is a Poisson random variable with parameter $\lambda > 0$, referred to as *Poisson mixture of BBs*; (ii) $M$ is a negative binomial random variable with parameters $(n_0, \mu_0)$, 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 3.6.** *If* $M \sim \text{Poisson}(\lambda)$ *in the mixture representation of Proposition 3.1, then the model has EFPF in product form* (3.1) *and the* $V_{n,k}$*'s are given by*

$$V_{n,k} = \frac{1}{k!} \exp\left\{-\lambda\left(1 - \frac{(\theta+\alpha)_n}{(\theta)_n}\right)\right\}\left\{\frac{-\lambda\alpha}{(\theta)_n}\right\}^k. \tag{3.17}$$

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

$$V_{n,k} = \binom{k+n_0-1}{k}\left\{\frac{-\alpha}{(\theta)_n}\frac{\mu_0}{\mu_0+n_0}\right\}^k\left(1 - \frac{\mu_0}{\mu_0+n_0}\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^{-n_0-k}\left(\frac{n_0}{\mu_0+n_0}\right)^{n_0}. \tag{3.18}$$

Clearly, the negative binomial mixture of BBs allows for a higher degree of prior uncertainty regarding the total number of features $M$ 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 $\lambda$ parameter of the Poisson mixture of BBs. Specifically, assuming $M \,|\, \lambda \sim \text{Poisson}(\lambda)$, and $\lambda \sim \text{Gamma}(a, b)$ is equivalent to having $M \sim \text{NegBinomial}(n_0, \mu_0)$, 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 $\lambda$, 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 3.2 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 $K_m^{(n)} \,|\, \boldsymbol{Z}$, leading to the extrapolation of the accumulation curves. It is worth reiterating that the predictive distribution $Y_{n+1} \,|\, \boldsymbol{Z}$ involved in the buffet metaphor can be derived as a special case, setting $m = 1$ in the distribution of $K_m^{(n)} \,|\, \boldsymbol{Z}$, which is a trivial task in the following formulas.

**Proposition 3.7.** *Suppose the EFPF is in product form* (3.1) *with* $\alpha < 0$ *and let* $p_n(\theta, \alpha) = 1 - (\theta+\alpha)_n/(\theta)_n$*. If* $M$ *is fixed, corresponding to the BB model* (3.3)*, then*

$$K_n \,|\, M \sim \text{Binomial}\left(M, p_n(\theta, \alpha)\right), \qquad K_m^{(n)} \,|\, \boldsymbol{Z}, M \sim \text{Binomial}\left(M - k, p_m(\theta + n, \alpha)\right).$$

*If instead* $M \sim \text{Poisson}(\lambda)$*, corresponding to model* (3.17)*, then*

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

*Finally, if* $M \sim \text{NegBinomial}(n_0, \mu_0)$*, corresponding to model* (3.18)*, then*

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

Proposition 3.7 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 $M$ is deterministic the prior expectation for $\mathsf{E}(K_n \mid M) = Mp_n(\theta, \alpha)$ 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 $M$ possible features. This elegant interpretation holds true also for the Poisson and negative binomial cases, as $\mathsf{E}(K_n) = \lambda p_n(\theta, \alpha)$ and $\mathsf{E}(K_n) = \mu_0 p_n(\theta, \alpha)$, 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 $\lambda$ and $\mu_0$. 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 $M$ is deterministic, the Bayesian estimator for the number of unseen features is $\mathsf{E}(K_m^{(n)} \mid \boldsymbol{Z}, M) = (M - k)p_m(\theta + n, \alpha)$, in which $p_m(\theta + n, \alpha)$ is the expected fraction of features in a future sample of size $m$, out of the remaining $M - k$ features. In the Poisson and negative binomial cases, where the total number of features $M$ 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 $M$, are

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

Note that $\mathsf{E}(\lambda \mid \boldsymbol{Z}) = (n_0 + k)/\{n_0/\mu_0 + p_n(\theta, \alpha)\}$ is the posterior expectation of $\lambda$ in the Poisson-gamma representation of the negative binomial, where $\mu_0/n_0$ denotes the overdispersion. It is important to emphasize how the sampling information $\boldsymbol{Z}$ affects the distribution of the statistic $K_m^{(n)}$. In the Poisson mixture, the distribution of $K_m^{(n)}$ depends on the initial sample $\boldsymbol{Z}$ 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. Lastly, Proposition 3.7 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 $M - k$. This eventual shrinkage of uncertainty is a very undesirable behavior. 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 $M$, i.e., the richness, which coincides with the $\alpha$-diversity, since $K_n \xrightarrow{\text{d}} M$ 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 3.3, it may be equivalently obtained by extrapolating the accumulation curve, specifically by considering $\lim_{m\to\infty} K_m^{(n)} + k \mid \boldsymbol{Z}$. For the BB model, the posterior distribution of $M$ is deterministic, since $M$ is known a priori. This yields critical issues as highlighted in simulation study A of the Appendix (Figures 3.D.4 and 3.D.5). For the proposed mixtures of BBs, the following result can be easily proved using Proposition 3.7 and noting that $\lim_{m\to\infty} p_m(\theta+n, \alpha) = 1$.

**Proposition 3.8.** *Suppose the* EFPF *is in product form* (3.1) *with* $\alpha < 0$. *Then* $K_m^{(n)} \mid \boldsymbol{Z} \xrightarrow{\text{d}} M'$ *with* $M' + k \overset{d}{=} M \mid \boldsymbol{Z}$, *as* $m \to \infty$. *Let* $p_n(\theta, \alpha) = 1-(\theta+\alpha)_n/(\theta)_n$, *if* $M \sim \text{Poisson}(\lambda)$,

*then*

$$M' \sim \text{Poisson}\left(\lambda(1 - p_n(\theta, \alpha))\right),$$

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

$$M' \sim \text{NegBinomial}\left(n_0 + k, \frac{n_0 + k}{n_0/\mu_0 + p_n(\theta, \alpha)}\{1 - p_n(\theta, \alpha)\}\right). \tag{3.19}$$

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 $M$, are the posterior expectations

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

To make these formulas operative, one needs to specify values for $\lambda$, $\theta$, and $\alpha$ and possibly for $n_0$ and $\mu_0$. 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 Appendix we investigate a fully Bayesian procedure in which we specify suitable priors for the hyperparameters.

### 3.4.2 Hierarchical formulation for mixtures of beta Bernoulli models

We now specialize the general posterior characterization in Theorem 3.4 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 (3.6) can be described as $\tilde{\mu} \,|\, M = \sum_{j=1}^{N} \tilde{q}_j \delta_{\tilde{X}_j}$, with $\tilde{q}_j \overset{\text{iid}}{\sim} \text{Beta}(-\alpha, \theta + \alpha)$ and $\tilde{X}_j \overset{\text{iid}}{\sim} G_0$, for some prior distribution $M \sim P_M$. When $M$ 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 (3.17) may be written as

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \text{BP}(\tilde{\mu}), \qquad i \geq 1 \\ \tilde{\mu} &\sim \text{CRM}(\rho; G_0), \end{aligned} \tag{3.20}$$

where the intensity measure $\rho(\mathrm{d}s)$ is finite since $\alpha < 0$ and is proportional to the density of a $\text{Beta}(-\alpha, \theta + \alpha)$ distribution $\rho(\mathrm{d}s) = \lambda\Gamma(\theta)/\{\Gamma(-\alpha)\Gamma(\theta + \alpha)\}s^{-\alpha-1}(1 - s)^{\theta+\alpha-1}\mathrm{d}s$. This result follows by a construction of completely random measures with finitely many jumps $\tilde{q}_j$, whose number has a Poisson distribution (Daley and Vere-Jones, 2008). As for the negative binomial mixture, the EFPF (3.18) is associated with the following statistical model involving negative binomial processes:

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \text{BP}(\tilde{\mu}), \qquad i \geq 1, \\ \tilde{\mu} &\sim \text{NB}(n_0, \rho; G_0), \end{aligned} \tag{3.21}$$

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 3.C for completeness. The following corollary is a consequence of Theorem 3.4, and it characterizes the posterior distribution of $\tilde{\mu}$ in both cases.

**Corollary 3.3.** *Suppose* $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ *follows models* (3.20) *or* (3.21)*, then the posterior distribution of* $\tilde{\mu}$*, given* $\boldsymbol{Z}$*, satisfies the decomposition* $\tilde{\mu} \,|\, \boldsymbol{Z} \stackrel{d}{=} \mu' + \mu^*$ *in* (3.10)*, where* $\mu'$ *and* $\mu^*$ *are independent random measures such that* $\mu^*$ *is distributed as in Theorem 3.4. Under model* (3.20) $\mu' \sim \textsc{crm}(\rho'; G_0)$*, 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* (3.21) $\mu' \sim \textsc{nb}(n_0 + k, \rho'; G_0)$*, 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 3.4, given the availability of the posterior for $M$ given in Proposition 3.8. 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).

## 3.5 Model fitting and simulation studies

### 3.5.1 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 (3.1) of a BB model, obtained with $V_{n,k}$'s as in (3.3). This procedure provides us with an estimate $\hat{\alpha}$ and $\hat{\theta}$ of the values of $\alpha$ and $\theta$, together with an estimate $\hat{M}$ for the total number of features $M$. Then, in the Poisson mixture we set $\lambda = \mathsf{E}(M) = \hat{M}$, whereas we set $\mu_0 = \mathsf{E}(M) = \hat{M}$ 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 $M$. We argue that such variance should just reflect the practitioner's degree of uncertainty around the prior guess $\mathsf{E}(M)$, possibly performing a sensitivity analysis for it. In our simulated scenarios, available in the Appendix, we will consider additional choices of the prior expectation $\mathsf{E}(M)$ for the mixtures of BBs, instead of specifying it through the data-driven approach just described. In such cases, we estimate the parameters $\alpha$ and $\theta$ by maximizing the model-specific EFPF, that is (3.17) for the Poisson mixture, with $\lambda = \mathsf{E}(M)$, and (3.18) for the negative binomial mixture, with $\mu_0 = \mathsf{E}(M)$ and $n_0$ such that the desired prior variance is obtained.

Along a similar argument, for the mixtures of IBPs we firstly propose to maximize the EFPF (3.1) with $V_{n,k}$'s as in (3.2) to find an estimate $\hat{\alpha}$ and $\hat{\theta}$ for the parameters $\alpha$, $\theta$, together with an estimate $\hat{\gamma}$ for the total mass $\gamma$. Secondly, we choose the parameters of the prior for $\gamma$ by enforcing the condition $\mathsf{E}(\gamma) = \hat{\gamma}$. In particular, for the gamma mixture of IBPs, we assume $\gamma \sim \text{Gamma}(a, b)$ and we set $\hat{\gamma} = \mathsf{E}(\gamma) = a/b$. Similarly to the negative binomial mixture of BBs, the prior variance of $\gamma$ 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 $\alpha$ and $\theta$ for all the mixtures. We exploit such a fully Bayesian approach in the real data analysis of Section 3.E.2, where we also show that posterior inferences obtained with the

two procedures are coherent.

### 3.5.2 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 $K_1, \dots, K_n$ with the expected values $\mathsf{E}(K_1), \dots, \mathsf{E}(K_n)$ under different models. Since the observed values $K_1, \dots, K_n$ refer to a particular ordering of the observations, in place of $K_1, \dots, K_n$, we will consider the in-sample accumulation curve $K_1', \dots, K_n'$, 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 $K_{n,1}, \dots, K_{n,\bar{r}}$ and the expected values $\mathsf{E}(K_{n,1}), \dots, \mathsf{E}(K_{n,\bar{r}})$, 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 $\alpha$, $\theta$ and the prior mean of $M$ (resp. $\gamma$) if a mixture of BBs (resp. IBPs) is selected. As a consequence, if we adopt the empirical Bayes approach described in Section 3.5.1 for parameters elicitation, then all the mixtures of BBs (resp. IBPs) have the same rarefaction curve and the same curve $\mathsf{E}(K_{n,r})$, for $r = 1, \dots, n$. 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 $\boldsymbol{Z}$ and a model described by parameters $\boldsymbol{\theta}$ (referred to as hyperparameters in our Bayesian setting), the (minimum) deviance is defined as $D(\hat{\boldsymbol{\theta}}) = -2 \log \mathcal{L}(\boldsymbol{Z} \,|\, \hat{\boldsymbol{\theta}}) = -2 \log \pi_n(m_1, \dots, m_k \,|\, \hat{\boldsymbol{\theta}})$, 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.

### 3.5.3 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 $M$,

corresponding to the $\alpha$-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 3.D, 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 3.5.1. As a second step, we apply the model-checking approaches we presented in Section 3.5.2 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 $M$ 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 $M$.

## 3.6 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 $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, for $m = 1, 2, \ldots$, and the associated $\alpha$-diversity (refer to Proposition 3.3 of Section 3.2.3). 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 3.5.1. Prediction and inference are then faced using the most appropriate model, selected through the model-checking described in Section 3.5.2. For a more exhaustive assessment of the models' predictive ability, we perform a data-holdout experiment in Section 3.E.1, 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 3.E.2, 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 3.5.1.

### 3.6.1 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 3.E.1 of the Appendix, 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 3.5.2 and the quantitative assessment through the comparison of deviances. From the plots of Figure 3.2, 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 on the gamma mixture of IBPs, and we consider two possible prior variances for $\gamma$, i.e., $\text{Var}(\gamma) \in \{1, 100\}$. From Proposition 3.5, the asymptotic growth rate of the curve $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, for $m = 1, 2, \ldots$, is of order $m^{\alpha}$, with an estimated rate of $\hat{\alpha} = 0.17$, and the posterior $\alpha$-diversity $S'_{\alpha}$ is gamma distributed; the expected value of $S'_{\alpha}$ equals 186.48 for both the choices of the prior variance of $\gamma$, 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 3.3, 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., $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, with $k = 215$. The posterior point estimates are similar for the two selected prior variances, while the variability increases as the prior variance of $\gamma$ increases.

To provide a more quantitative answer to the extrapolation problem ecologists might be interested in, we report in Table 3.1 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

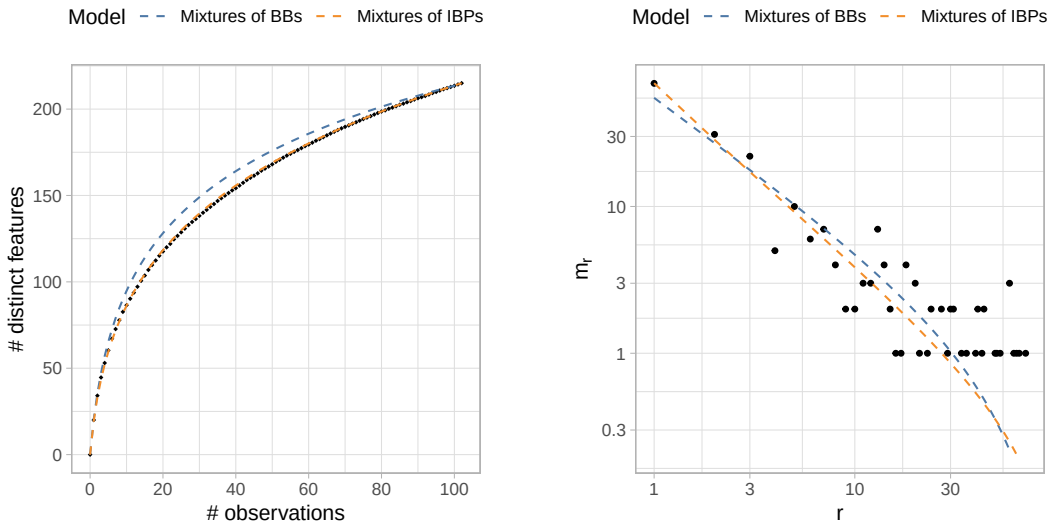



Figure 3.2: Left panel: the empirical accumulation curve (black dots) and the rarefaction curve of the models (blue and orange dashed lines). Right panel: the observed values of $K_{n,r}$ (black dots) compared with the expected curve $\mathsf{E}(K_{n,r})$ of the models (blue and orange dashed lines). The right plot is in log-log scale; the orange (resp. blue) curves are identical for all the mixtures of IBPs (resp. BBs).

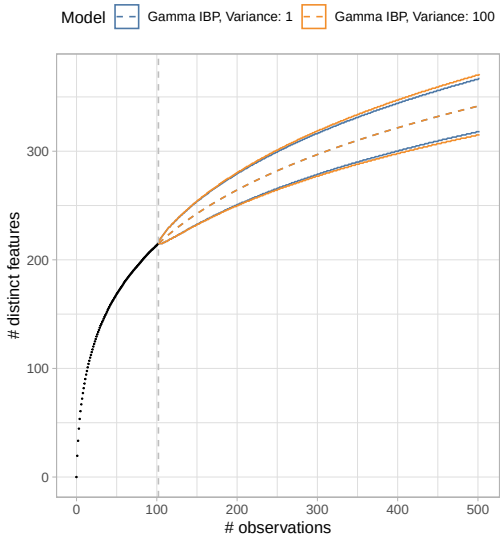



Figure 3.3: Expected values and the 95% credible intervals for $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, with $k = 215$, for the gamma mixture of IBPs. The extrapolation horizon is $m = 1, \ldots, 400$.

examined, for some values of $m$. Both the choices for the prior variance of $\gamma$ 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 $m \in \{100, 1000\}$ in terms of their credible intervals.

Table 3.1: Expected values $\mathsf{E}(K_m^{(n)} \mid \boldsymbol{Z})$ and 95% credible intervals (in brackets) for the statistic $K_m^{(n)} \mid \boldsymbol{Z}$, for $m \in \{1, 10, 100, 1000\}$, for the vascular plants in Mazziotta et al. (2016a).

| Gamma mixture of IBPs ($n = 102, k = 215$) | $m = 1$ | $m = 10$ | $m = 100$ | $m = 1000$ |
|---|---|---|---|---|
| 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] |

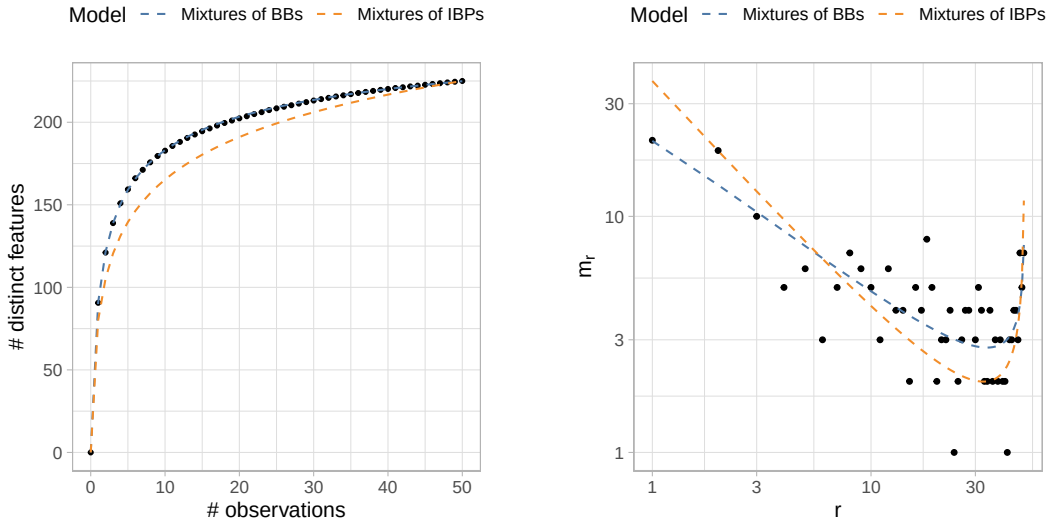



Figure 3.4: Left panel: the empirical accumulation curve (black dots) and the rarefaction curve of the models (blue and orange dashed lines). Right panel: the observed values of $K_{n,r}$ (black dots) compared with the expected curve $\mathsf{E}(K_{n,r})$ of the models (blue and orange dashed lines). The right plot is in log-log scale; the orange (resp. blue) curves are identical for all the mixtures of IBPs (resp. BBs).

### 3.6.2 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 3.E.4 of the Appendix, 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 3.5.2. The visual inspection of Figure 3.4 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 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 $M$, i.e., $\mathrm{Var}(M) = \mu_0 \times c$, 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 3.5, we report the posterior distribution of the species richness $M$, 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 $[278, 316]$.

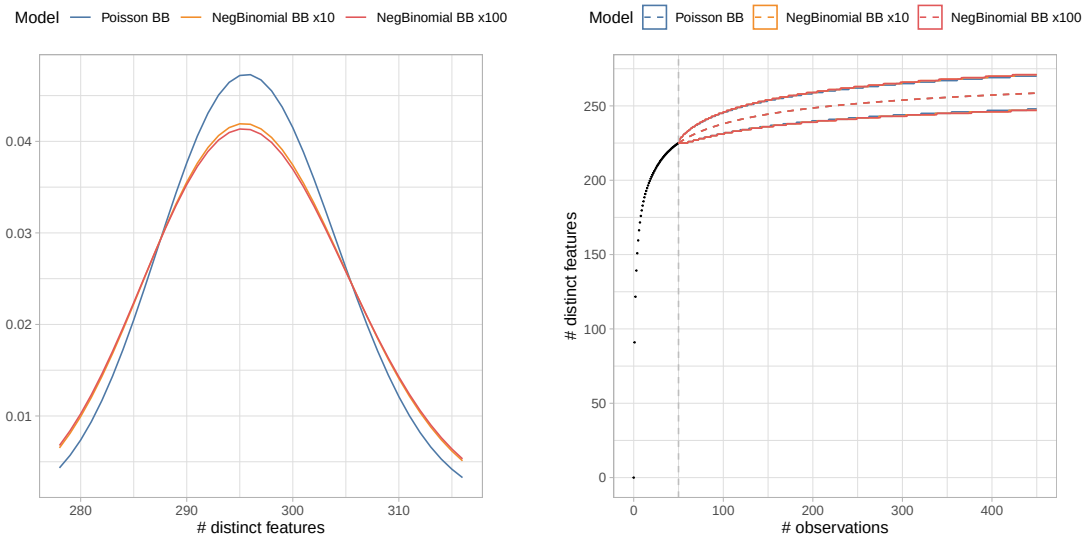



Figure 3.5: Left panel: posterior distributions of the species richness $M$ for different mixtures of BBs. Right panel: expected values and 95% credible intervals for $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, for different mixtures of BBs.

As far as the extrapolation problem is concerned, Figure 3.5 (right panel) reports the expected values and the 95% credible intervals for the posteriors $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, 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 3.3, we remark that such a sequence $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, $m = 1, 2, \ldots$, converges to the posterior distribution of the species richness $M$. As we have just discussed, all the mixtures of BBs that we have fitted provide an expected richness of around 296.

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 $m = 1000$: 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.

## 3.7 Discussion

In the present chapter, we analyzed Gibbs-type feature models, namely those models exhibiting an EFPF in product form (3.1). 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 paves the way for several 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 research involves defining and investigating feature allocation models in the presence of multi-sample data. This constitutes the main focus of our contribution in Chapter 5. Although numerous models are available for partially exchangeable data within the species setting (Quintana et al., 2022), analogous developments within the feature allocation framework remain relatively scarce; see, e.g., Teh and Jordan (2010) for a few early examples. Secondly, Gibbs-type feature models are natural tools for modeling biodiversity when focusing 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. Thirdly, 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 $M$, 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 and 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 work 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.

# Appendix

## Organization of the Appendix

The Appendix is organized as follows. In Section 3.A, we prove all the general results of Section 3.2, valid for any Gibbs-type feature model. In addition, in Section 3.A.4 we provide general distributional results for the statistic $K_{n,r}$, denoting the number of features appearing exactly $r$ times across $n$ individuals. Section 3.A.5 specializes the general distribution theory for $K_{n,r}$ under the Poisson and negative binomial mixtures of BBs and the gamma mixture of IBPs. Section 3.B (resp. Section 3.C) contains all the proofs and details of Section 3.3 (resp. Section 3.4). Our simulation studies are reported in Section 3.D. Finally, Section 3.E includes additional plots referring to the ecological applications presented in the main body; moreover, in Section 3.E.2, we also propose a fully Bayesian approach for parameters' elicitation, alternative to the empirical Bayes one described in Section 3.5.1. We apply such a fully Bayesian approach to the two real data scenarios of Section 3.6 in comparison with the inference obtained via the empirical Bayes procedure.

## 3.A Proofs of Section 3.2 and additional results

### 3.A.1 Proof of Theorem 3.1

In order to determine the predictive distribution, we have to pay attention to the ordering of the new features. More precisely, if $K_n = k$, there are $\binom{k+y}{k}$ possible ways to order the new $Y_{n+1} = y$ observed features with respect to the first $K_n$ ordered features. By taking into account this combinatorial coefficient, the predictive law equals

$$p_n(y, a_1, \ldots, a_k) = \frac{\binom{k+y}{k}\pi_{n+1}(m_1 + a_1, \ldots, m_k + a_k, 1, \ldots, 1)}{\pi_n(m_1, \ldots, m_k)}$$

where the number 1 at the numerator is repeated $y$ times. The numerator corresponds to the probability of the feature allocation induced by the observed sample $\boldsymbol{Z}$ and $Z_{n+1}$, by considering all the possible orders of the $y$ newly observed features. Besides, the denominator is the probability distribution of the feature allocation induced by the sample $\boldsymbol{Z}$. By substituting the specific expression of the EFPF in product form (3.1), we obtain

$$p_n(y, a_1, \ldots, a_k) = \binom{k+y}{k}\frac{V_{n+1,k+y}}{V_{n,k}}(\theta+\alpha)_n^y \cdot \prod_{\ell=1}^{k}\frac{(1-\alpha)_{m_\ell+a_\ell-1}}{(1-\alpha)_{m_\ell-1}}\frac{(\theta+\alpha)_{n+1-m_\ell-a_\ell}}{(\theta+\alpha)_{n-m_\ell}}.$$

By observing that

$$\frac{(1-\alpha)_{m_\ell+a_\ell-1}}{(1-\alpha)_{m_\ell-1}} = (m_\ell-\alpha)^{a_\ell}, \quad \frac{(\theta+\alpha)_{n+1-m_\ell-a_\ell}}{(\theta+\alpha)_{n-m_\ell}} = (\theta+\alpha+n-m_\ell)^{1-a_\ell},$$

we get the following expression:

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

and the thesis follows.

### 3.A.2 Proof of Corollary 3.1

The corollary easily follows from Theorem 3.1. Indeed, the sample coverage equals

$$\mathsf{P}(Y_{n+1} = 0 \mid \boldsymbol{Z}) = \sum_{(a_1,\dots,a_k)\in\{0,1\}^k} p_n(0, a_1, \dots, a_k) = \frac{V_{n+1,k}}{V_{n,k}} (\theta+n)^k \prod_{\ell=1}^{k} \sum_{a_\ell=0}^{1} \mathscr{B}\left(a_\ell; \frac{m_\ell - \alpha}{\theta+n}\right)$$
$$= \frac{V_{n+1,k}}{V_{n,k}} (\theta+n)^k$$

where we have marginalized out the probability of observing old features from the predictive distribution.

### 3.A.3 Proof of Theorem 3.2

We start by proving the following lemma, which provides the probability distribution of $K_n$ for any exchangeable feature allocation model admitting an EFPF.

**Lemma 3.1.** *For any feature allocation model admitting an EFPF, the probability distribution of $K_n$ is*

$$\mathsf{P}(K_n = k) = \sum_{\substack{m_\ell \in \{1,\dots,n\} \\ \ell = 1,\dots,k}} \pi_n(m_1, \dots, m_k) \binom{n}{m_1} \cdots \binom{n}{m_k}.$$

*Proof.* The event $\{K_n = k\}$ corresponds to the union of all feature allocations of the type $F_n = (B_{n,1}, \dots, B_{n,k})$, where $m_\ell = |B_{n,\ell}|$ takes any value in the set $\{1, \dots, n\}$, for any $\ell = 1, \dots, k$. For a specific configuration of $(m_1, \dots, m_k)$, there are several feature allocations $F_n$ such that $m_\ell = |B_{n,\ell}|$, as $\ell = 1, \dots, k$. Indeed, there are $\binom{n}{m_1}$ different ways to choose the indexes in the set $[n]$ to define $B_{n,1}$, similarly there are $\binom{n}{m_2}$ different ways to choose the set $B_{n,2}$, etc.. Hence, the probability of observing all feature allocations having predetermined block sizes $(m_1, \dots, m_k)$ equals $\pi_n(m_1, \dots, m_k)\binom{n}{m_1} \cdots \binom{n}{m_k}$. As a consequence, the probability $\mathsf{P}(K_n = k)$ can be obtained by summing over all the possible configurations of the vector $(m_1, \dots, m_k) \in \{1, \dots, n\}^k$. □

We now exploit Lemma 3.1 to find the probability distribution of $K_n$ for any Gibbs-type

feature allocation model:

$$\begin{aligned}\mathsf{P}(K_n = k) &= \sum_{\substack{m_\ell \in \{1,\dots,n\} \\ \ell = 1,\dots,k}} V_{n,k} \prod_{\ell=1}^{k} (1-\alpha)_{m_\ell - 1} (\theta + \alpha)_{n - m_\ell} \binom{n}{m_\ell} \\ &= V_{n,k} \prod_{\ell=1}^{k} \sum_{m_\ell = 1}^{n} \binom{n}{m_\ell} (1-\alpha)_{m_\ell - 1} (\theta + \alpha)_{n - m_\ell} \\ &= V_{n,k} \left[ \sum_{i=1}^{n} \binom{n}{i} (1-\alpha)_{i-1} (\theta + \alpha)_{n-i} \right]^k .\end{aligned} \tag{3.22}$$

We now consider the case $\alpha < 0$ and $\alpha \in [0, 1)$ separately.

**Case $\boldsymbol{\alpha < 0}$**. For these values of $\alpha$, the last sum in (3.22) can be expressed as

$$\sum_{i=1}^{n} \binom{n}{i} (1-\alpha)_{i-1} (\theta + \alpha)_{n-i} = \frac{\Gamma(-\alpha)}{\Gamma(1-\alpha)} \sum_{i=0}^{n} \binom{n}{i} (-\alpha)_i (\theta + \alpha)_{n-i} + \frac{(\theta + \alpha)_n}{\alpha}.$$

Thanks to the Chu-Vandermonde identity, we obtain

$$\frac{\Gamma(-\alpha)}{\Gamma(1-\alpha)} \sum_{i=0}^{n} \binom{n}{i} (-\alpha)_i (\theta + \alpha)_{n-i} + \frac{(\theta + \alpha)_n}{\alpha} = -\frac{(\theta)_n}{\alpha} + \frac{(\theta + \alpha)_n}{\alpha}$$

and the thesis follows for $\alpha < 0$.

**Case $\boldsymbol{\alpha \in [0, 1)}$.** For positive values of $\alpha$, the sum in (3.22) can be expressed as

$$\begin{aligned}\sum_{i=1}^{n} \binom{n}{i} (1-\alpha)_{i-1} (\theta + \alpha)_{n-i} &= \sum_{i=1}^{n} \binom{n}{i} \frac{\Gamma(i-\alpha)}{\Gamma(1-\alpha)} \frac{\Gamma(\theta + \alpha + n - i)}{\Gamma(\theta + \alpha)} \cdot \frac{\Gamma(\theta + n)}{\Gamma(\theta + n)} \\ &= \frac{\Gamma(\theta + n)}{\Gamma(1-\alpha)\Gamma(\theta + \alpha)} \sum_{i=1}^{n} \binom{n}{i} B(i - \alpha, \theta + \alpha + n - i),\end{aligned}$$

where $B(a, b)$ denotes the Euler's beta function evaluated at $a, b > 0$. Thanks to the integral representation of the beta function, one obtains:

$$\begin{aligned}\sum_{i=1}^{n} \binom{n}{i} (1-\alpha)_{i-1} (\theta + \alpha)_{n-i} &= \frac{\Gamma(\theta + n)}{\Gamma(1-\alpha)\Gamma(\theta + \alpha)} \sum_{i=1}^{n} \binom{n}{i} \int_0^1 x^{i-\alpha-1} (1-x)^{\theta+\alpha+n-i-1} \mathrm{d}x \\ &= \frac{\Gamma(\theta + n)}{\Gamma(1-\alpha)\Gamma(\theta + \alpha)} \int_0^1 x^{-\alpha-1} (1-x)^{\theta+\alpha+n-1} \sum_{i=1}^{n} \binom{n}{i} \left(\frac{x}{1-x}\right)^i \mathrm{d}x \\ &= \frac{\Gamma(\theta + n)}{\Gamma(1-\alpha)\Gamma(\theta + \alpha)} \int_0^1 x^{-\alpha-1} (1-x)^{\theta+\alpha+n-1} \left[ \left( \frac{x}{1-x} + 1 \right)^n - 1 \right] \mathrm{d}x \\ &= \frac{\Gamma(\theta + n)}{\Gamma(1-\alpha)\Gamma(\theta + \alpha)} \int_0^1 x^{-\alpha-1} (1-x)^{\theta+\alpha+n-1} \left[ \left( \frac{1}{1-x} \right)^n - 1 \right] \mathrm{d}x \\ &= \frac{\Gamma(\theta + n)}{\Gamma(1-\alpha)\Gamma(\theta + \alpha)} \int_0^1 x^{-\alpha-1} (1-x)^{\theta+\alpha-1} \left[ 1 - (1-x)^n \right] \mathrm{d}x.\end{aligned}$$

By virtue of the following equality

$$1 - (1-x)^n = x \sum_{i=0}^{n-1} (1-x)^i,$$

we obtain

$$\begin{aligned}
&\frac{\Gamma(\theta+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\int_0^1 x^{-\alpha-1}(1-x)^{\theta+\alpha-1}\left[1-(1-x)^n\right]\mathrm{d}x \\
&\qquad=\frac{\Gamma(\theta+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\int_0^1 x^{-\alpha+1-1}(1-x)^{\theta+\alpha-1}\sum_{i=0}^{n-1}(1-x)^i\mathrm{d}x \\
&\qquad=\frac{\Gamma(\theta+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\sum_{i=0}^{n-1}\int_0^1 x^{-\alpha+1-1}(1-x)^{\theta+\alpha+i-1}\mathrm{d}x \\
&\qquad=\frac{\Gamma(\theta+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\sum_{i=0}^{n-1}\frac{\Gamma(i+\theta+\alpha)\Gamma(1-\alpha)}{\Gamma(i+\theta+1)} \\
&\qquad=\frac{\Gamma(\theta+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\sum_{i=1}^{n}\frac{\Gamma(i+\theta+\alpha-1)\Gamma(1-\alpha)}{\Gamma(i+\theta)} \\
&\qquad=\frac{\Gamma(\theta+n)}{\Gamma(\theta+1)}\sum_{i=1}^{n}\frac{(\theta+\alpha)_{i-1}}{(\theta+1)_{i-1}}=(\theta+1)_{n-1}\sum_{i=1}^{n}\frac{(\theta+\alpha)_{i-1}}{(\theta+1)_{i-1}}
\end{aligned}$$

and the thesis follows by substituting the previous expression in (3.22).

### 3.A.4 Distributional results for the number of $r$-shared features $K_{n,r}$

Here we provide distributional results for the statistic $K_{n,r}$, where $r \in \{1,\dots,n\}$, denoting the number of features appearing exactly $r$ times among $n$ individuals. We start by proving the following lemma, which provides the probability distribution of $K_{n,r}$ for any feature allocation model admitting an EFPF.

**Lemma 3.2.** *For any feature allocation model admitting an EFPF, the distribution of $K_{n,r}$ equals*

$$\begin{aligned}
&\mathsf{P}(K_{n,r}=y) \\
&\quad=\binom{n}{r}^y\left\{\pi_n(r,\dots,r)+\sum_{k=y+1}^{\infty}\binom{k}{y}\sum_{\substack{m_\ell\in\{1,\dots,n\}\\ m_\ell\neq r\\ \ell=1,\dots,k-y}}\pi_n(r,\dots,r,m_1,\dots,m_{k-y})\prod_{\ell=1}^{k-y}\binom{n}{m_\ell}\right\}
\end{aligned} \tag{3.23}$$

*where $(r,\dots,r)$ in the previous expression is a vector of length $y$.*

*Proof.* The event $\{K_{n,r}=y\}$ corresponds to all random ordered feature allocations of the type $F_n=(B_{n,1},\dots,B_{n,k})$, where $k\geq y$ and there are exactly $y$ sets out of the $B_{n,\ell}$'s with cardinality $r$.
Consider $k>y$, and suppose that the first $y$ sets are those with cardinality $r$, i.e., $|B_{n,\ell}|=r$, for all $\ell=1,\dots,y$, and that the remaining sets $B_{n,y+\ell}$, as $\ell=1,\dots,k-y$, have cardinalities $m_\ell\in\{1,\dots,n\}$, where $m_\ell\neq r$. There are $\binom{n}{r}$ different ways to choose the indexes in $[n]$ to construct each of the first $y$ blocks $B_{n,\ell}$, as $\ell=1,\dots,y$. Analogously, there are $\binom{n}{m_\ell}$ different ways to define $B_{n,y+\ell}$, as $\ell=1,\dots,k-y$. Thus, the probability of observing all feature allocations having the first $y(<k)$ blocks with cardinality $r$ equals

$$\pi_n(r,\dots,r,m_1,\dots,m_{k-y})\binom{n}{r}^y\binom{n}{m_1}\cdots\binom{n}{m_{k-y}}.$$

The previous probability refers to all the feature allocations where the first $y$ blocks have cardinality $r$, it is apparent that we have to multiply by the factor $\binom{k}{y}$ to take into account all the possible rearrangements of the $y$ blocks with cardinality $r$. When $k = y$, the probability that all the blocks of the feature allocation have cardinality $r$ equals $\pi_n(r,\dots,r)\binom{n}{r}^y$, where $(r,\dots,r)$ is a vector of length $y$. The expression of $\mathsf{P}(K_{n,r} = y)$ can be obtained by summing the previous probabilities when $k \geq y$:

$$\mathsf{P}(K_{n,r} = y) = \pi_n(r,\dots,r)\binom{n}{r}^y + \sum_{k=y+1}^{\infty} \sum_{\substack{m_\ell \in \{1,\dots,n\} \\ m_\ell \neq r \\ \ell=1,\dots,k-y}} \binom{k}{y} \pi_n(r,\dots,r,m_1,\dots,m_{k-y}) \binom{n}{r}^y \prod_{\ell=1}^{k-y} \binom{n}{m_\ell}$$

and the thesis now follows. □

We now state our main result of the section by specializing the previous lemma to the class of Gibbs-type feature allocation models.

**Proposition 3.9** (Number of $r$-shared features $K_{n,r}$)**.** *Suppose the* EFPF *is in product form* (3.1)*. Then, for any* $k \geq 0$*,*

$$\mathsf{P}(K_{n,r} = k) = \left\{ \binom{n}{r} (1-\alpha)_{r-1} (\theta+\alpha)_{n-r} \right\}^k \sum_{j=0}^{\infty} \binom{j+k}{k} V_{n,j+k} \{t_{n,r}(\theta,\alpha)\}^k,$$

*where, if* $\alpha < 0$*,*

$$t_{n,r}(\theta,\alpha) = -\frac{(\theta)_n}{\alpha} + \frac{(\theta+\alpha)_n}{\alpha} - \binom{n}{r}(1-\alpha)_{r-1}(\theta+\alpha)_{n-r},$$

*whereas, if* $\alpha \in [0,1)$*,*

$$t_{n,r}(\theta,\alpha) = (\theta+1)_{n-1}\, g_n(\theta,\alpha) - \binom{n}{r}(1-\alpha)_{r-1}(\theta+\alpha)_{n-r}.$$

*Proof.* We exploit Lemma 3.2 to evaluate $\mathsf{P}(K_{n,r} = y)$ for the Gibbs-type feature allocation models. A plain application of Equation (3.23) leads to

$$\begin{aligned}
\mathsf{P}(K_{n,r} = y) &= \left[ \binom{n}{r} (1-\alpha)_{r-1} (\theta+\alpha)_{n-r} \right]^y \\
&\quad \times \left\{ V_{n,y} + \sum_{z=1}^{\infty} \binom{z+y}{y} V_{n,z+y} \sum_{\substack{m_\ell \in \{1,\dots,n\} \\ m_\ell \neq r \\ \ell=1,\dots,k-y}} \prod_{\ell=1}^{z} \binom{n}{m_\ell} (1-\alpha)_{m_\ell - 1} (\theta+\alpha)_{n-m_\ell} \right\} \\
&= \left[ \binom{n}{r} (1-\alpha)_{r-1} (\theta+\alpha)_{n-r} \right]^y \times \\
&\quad \times \sum_{z=0}^{\infty} \binom{z+y}{y} V_{n,z+y} \left[ \sum_{\substack{i \in \{1,\dots,n\} \\ i \neq r}} \binom{n}{i} (1-\alpha)_{i-1} (\theta+\alpha)_{n-i} \right]^z \\
&= \left[ \binom{n}{r} (1-\alpha)_{r-1} (\theta+\alpha)_{n-r} \right]^y \times \\
&\quad \times \sum_{z=0}^{\infty} \binom{z+y}{y} V_{n,z+y} \left[ \sum_{i=1}^{n} \binom{n}{i} (1-\alpha)_{i-1} (\theta+\alpha)_{n-i} - \binom{n}{r} (1-\alpha)_{r-1} (\theta+\alpha)_{n-r} \right]^z.
\end{aligned}$$

We can now apply a similar strategy as in the proof of Theorem 3.2 to get the expressions for the case $\alpha < 0$ and $\alpha \in [0, 1)$. □

### 3.A.5 Additional specialized results for novel models

The general expression for the number of $r$-shared features $K_{n,r}$ presented in Proposition 3.9 can be specialized in the case of gamma mixture of IBPs and Poisson and negative binomial mixtures of BBs, leading to standard and well-known probability distributions. In the following proposition, we report the specialized results for the mixtures of BBs.

**Proposition 3.10.** *Suppose the EFPF is in product form* (3.1) *with* $\alpha < 0$ *and let* $b_{n,r}(\theta, \alpha) := -\alpha\binom{n}{r}(1-\alpha)_{r-1}(\theta+\alpha)_{n-r}/(\theta)_n$. *If* $M$ *is fixed, corresponding to the BB model* (3.3), *then*

$$K_{n,r} \,|\, M \sim \text{Binomial}\,(M, b_{n,r})\,.$$

*If instead* $M \sim \text{Poisson}(\lambda)$, *corresponding to model* (3.17), *then*

$$K_{n,r} \sim \text{Poisson}\,(\lambda \cdot b_{n,r})\,.$$

*Finally, if* $M \sim \text{NegBinomial}(n_0, \mu_0)$, *corresponding to model* (3.18), *then*

$$K_{n,r} \sim \text{NegBinomial}\,(n_0, \mu_0 \cdot b_{n,r})\,.$$

For the gamma mixture of IBPs, the prior distribution of $K_{n,r}$ is described in the next proposition.

**Proposition 3.11.** *Suppose the EFPF is in product form* (3.1) *with* $\alpha \in [0, 1)$ *and* $d_{n,r} := \binom{n}{r}(1-\alpha)_{r-1}(\theta+\alpha)_{n-r}\Gamma(\theta+1)/\Gamma(\theta+n)$. *If* $\gamma$ *is fixed, corresponding to the IBP model* (3.2), *then*

$$K_{n,r} \,|\, \gamma \sim \text{Poisson}\,(\gamma \cdot d_{n,r})\,.$$

*If instead* $\gamma \sim \text{Gamma}(a, b)$, *then*

$$K_{n,r} \sim \text{NegBinomial}\,(a, a/b \cdot d_{n,r})\,.$$

### 3.A.6 Proof of Theorem 3.3

We start by proving the following lemma, which provides the probability distribution of $K_m^{(n)} \,|\, \boldsymbol{Z}$ for any feature allocation model admitting an EFPF.

**Lemma 3.3.** *For any feature allocation model admitting an EFPF, the distribution of* $K_m^{(n)} \,|\, \boldsymbol{Z}$ *equals*

$$\begin{aligned}\mathsf{P}(K_m^{(n)} = y \,|\, \boldsymbol{Z}) = \binom{k+y}{k} \frac{1}{\pi_n(m_1, \ldots, m_k)} \\ \times \sum_{\substack{r_t \in \{1,\ldots,m\} \\ t=1,\ldots,y}} \sum_{\substack{c_\ell \in \{0,\ldots,m\} \\ \ell=1,\ldots,k}} \pi_{n+m}(m_1 + c_1, \ldots, m_k + c_k, r_1, \ldots, r_y) \prod_{r=1}^{y} \binom{m}{r_t} \prod_{\ell=1}^{k} \binom{m}{c_\ell}.\end{aligned}$$

*Proof.* We denote by $F_n = (B_{n,1}, \ldots, B_{n,K_n})$ the random ordered feature allocation corresponding to the random sample $\boldsymbol{Z}$, and by $f_n$ the observed value of $F_n$. Thanks to the correspondence between $\boldsymbol{Z}$ and $F_n$, as explained by Broderick et al. (2013), we have

$$\mathsf{P}(K_m^{(n)} = y \,|\, \boldsymbol{Z}) = \frac{\mathsf{P}(K_m^{(n)} = y, F_n = f_n)}{\mathsf{P}(F_n = f_n)}, \tag{3.24}$$

where we observe that $\mathsf{P}(F_n = f_n) = \pi_n(m_1, \ldots, m_k)$. Denote by $(B_{m+n,1}, \ldots, B_{m+n,k})$ the random ordered feature allocation which refers to the old features, and by $(B_{m+n,k+1}, \ldots, B_{m+n,k+y})$ the random ordered feature allocation corresponding to the $y$ new features. In order to evaluate the probability at the numerator in (3.24), we need to evaluate the probability that the aforementioned random ordered feature allocations satisfy:

(i) $(B_{m+n,1}, \ldots, B_{m+n,k})$ is consistent with the initial $F_n$, in the sense that each set $B_{m+n,\ell}$ contains the $m_\ell$ indexes of the corresponding $B_{n,\ell}$, as $\ell = 1, \ldots, k$;

(ii) each set in $(B_{m+n,1}, \ldots, B_{m+n,k})$ contains $c_\ell$ indexes from $\{n+1, \ldots, n+m\}$, for all the possible choices of $c_\ell \in \{0, \ldots, m\}$, as $\ell = 1, \ldots, k$;

(iii) the sets $B_{m+n,k+1}, \ldots, B_{m+n,k+y}$, corresponding to the *new* features, contain $r_t$ indexes out of the set $\{n+1, \ldots, n+m\}$, for all the possible choices of $r_t \in \{1, \ldots, m\}$, as $t = 1, \ldots, y$.

We note that $B_{m+n,\ell}$ can be chosen in $\binom{m}{c_\ell}$ different ways, as $\ell = 1, \ldots, k$, because this is the number of possible ways to select $c_\ell$ indexes among $\{n+1, \ldots, n+m\}$ of the subjects displaying feature $\ell$. Besides, $B_{m+n,k+t}$ can be chosen in $\binom{m}{r_t}$ different ways, as $t = 1, \ldots, y$. The probability of observing random ordered feature allocations satisfying (i)–(iii) equals

$$\pi_{n+m}(m_1 + c_1, \ldots, m_k + c_k, r_1, \ldots, r_y) \prod_{t=1}^{y} \binom{m}{r_t} \prod_{\ell=1}^{k} \binom{m}{c_\ell}.$$

We finally observe that there are $\binom{k+y}{y}$ possible ways for ordering the $y$ new features among the $k$ old features, as a consequence one has:

$$\begin{aligned}\mathsf{P}(K_m^{(n)} = y \,|\, \boldsymbol{Z}) &= \frac{1}{\pi_n(m_1, \ldots, m_k)} \\ &\times \sum_{\substack{r_t \in \{1,\ldots,m\} \\ t=1,\ldots,y}} \sum_{\substack{c_\ell \in \{0,\ldots,m\} \\ \ell=1,\ldots,k}} \pi_{n+m}(m_1 + c_1, \ldots, m_k + c_k, r_1, \ldots, r_y) \prod_{r=1}^{y} \binom{m}{r_t} \prod_{\ell=1}^{k} \binom{m}{c_\ell} \binom{k+y}{k}\end{aligned}$$

and the thesis follows. □

We now concentrate on the posterior probability distribution of $K_m^{(n)}$ in Gibbs-type feature allocation models. By virtue of Lemma 3.3, one has:

$$\begin{aligned}\mathsf{P}(K_m^{(n)} = y \,|\, \boldsymbol{Z}) = \binom{k+y}{k} \frac{V_{n+m,k+y}}{\pi_n(m_1, \ldots, m_k)} \sum_{\substack{r_t \in \{1,\ldots,m\} \\ t=1,\ldots,y}} \prod_{t=1}^{y} \binom{m}{r_t} (1-\alpha)_{r_t - 1} (\theta + \alpha)_{n+m-r_t} \\ \times \sum_{\substack{c_\ell \in \{0,\ldots,m\} \\ \ell=1,\ldots,k}} \prod_{\ell=1}^{k} \binom{m}{c_\ell} (1-\alpha)_{m_\ell + c_\ell - 1} (\theta + \alpha)_{n+m-m_\ell-c_\ell}.\end{aligned} \tag{3.25}$$

First, consider the following sum

$$\sum_{\substack{c_\ell\in\{0,\dots,m\}\\ \ell=1,\dots,k}}\prod_{\ell=1}^{k}\binom{m}{c_\ell}(1-\alpha)_{m_\ell+c_\ell-1}(\theta+\alpha)_{n+m-m_\ell-c_\ell}=\prod_{\ell=1}^{k}\sum_{c_\ell=0}^{m}\binom{m}{c_\ell}(1-\alpha)_{m_\ell+c_\ell-1}(\theta+\alpha)_{n+m-m_\ell-c_\ell},$$

by observing that

$$(1-\alpha)_{m_\ell+c_\ell-1}=(1-\alpha)_{m_\ell-1}(m_\ell-\alpha)_{c_\ell}$$

and

$$(\theta+\alpha)_{n+m-m_\ell-c_\ell}=(\theta+\alpha)_{n-m_\ell}(\theta+\alpha+n-m_\ell)_{m-c_\ell},$$

the Chu-Vandermonde identity implies

$$\prod_{\ell=1}^{k}\sum_{c_\ell=0}^{m}\binom{m}{c_\ell}(1-\alpha)_{m_\ell+c_\ell-1}(\theta+\alpha)_{n+m-m_\ell-c_\ell}=\prod_{\ell=1}^{k}\left[(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}(\theta+n)_m\right].$$

By exploiting the previous identity and the expression of the EFPF (3.1), the probability of interest (3.25) boils down to

$$\mathsf{P}(K_m^{(n)}=y\,|\,\boldsymbol{Z})=\binom{k+y}{k}\frac{V_{n+m,k+y}}{V_{n,k}}(\theta+n)_m^k\left[\sum_{i=1}^{m}\binom{m}{i}(1-\alpha)_{i-1}(\theta+\alpha)_{n+m-i}\right]^y. \tag{3.26}$$

We now consider separately the two cases $\alpha<0$ and $\alpha\in[0,1)$.

**Case $\boldsymbol{\alpha<0}$.** For such values of $\alpha$, we observe that

$$(1-\alpha)_{i-1}=-\frac{1}{\alpha}(-\alpha)_i\quad\text{and}\quad(\theta+\alpha)_{n+m-i}=(\theta+\alpha)_n(\theta+\alpha+n)_{m-i},$$

and, thanks to the Chu-Vandermonde identity, one has

$$\sum_{i=1}^{m}\binom{m}{i}(1-\alpha)_{i-1}(\theta+\alpha)_{n+m-i}=-\frac{1}{\alpha}(\theta+\alpha)_n\left[(\theta+n)_m-(\theta+\alpha+n)_m\right].$$

The result for $\alpha<0$ follows by substituting the previous expression in (3.26).

**Case $\boldsymbol{\alpha\in[0,1)}$.** As for $\alpha\in[0,1)$, the sum appearing in (3.26) can be rewritten as follows:

$$\begin{aligned}\sum_{i=1}^{m}\binom{m}{i}(1-\alpha)_{i-1}(\theta+\alpha)_{n+m-i}&=\sum_{i=1}^{m}\binom{m}{i}\frac{\Gamma(i-\alpha)}{\Gamma(1-\alpha)}\frac{\Gamma(\theta+\alpha+m+n-i)}{\Gamma(\theta+\alpha)}\cdot\frac{\Gamma(\theta+m+n)}{\Gamma(\theta+m+n)}\\&=\frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\sum_{i=1}^{m}\binom{m}{i}B(i-\alpha,\theta+\alpha+m+n-i).\end{aligned}$$

We can now use the integral representation of the beta function to get

$$\begin{aligned}\sum_{i=1}^{m}\binom{m}{i}(1-\alpha)_{i-1}(\theta+\alpha)_{n+m-i}&=\frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\sum_{i=1}^{m}\binom{m}{i}\int_0^1 x^{i-\alpha-1}(1-x)^{\theta+\alpha+m+n-i-1}\mathrm{d}x\\&=\frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\int_0^1 x^{-\alpha-1}(1-x)^{\theta+\alpha+m+n-1}\sum_{i=1}^{m}\binom{m}{i}\left(\frac{x}{1-x}\right)^i\mathrm{d}x\\&=\frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\int_0^1 x^{-\alpha-1}(1-x)^{\theta+\alpha+m+n-1}\left[\left(\frac{x}{1-x}+1\right)^m-1\right]\mathrm{d}x\\&=\frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\int_0^1 x^{-\alpha-1}(1-x)^{\theta+\alpha+n-1}\left[1-(1-x)^m\right]\mathrm{d}x.\end{aligned}$$

By virtue of the simple identity

$$1-(1-x)^m = x\sum_{i=0}^{m-1}(1-x)^i,$$

the previous expression becomes

$$\begin{aligned}
&\sum_{i=1}^{m}\binom{m}{i}(1-\alpha)_{i-1}(\theta+\alpha)_{n+m-i} \\
&\qquad = \frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\int_0^1 x^{-\alpha+1-1}(1-x)^{\theta+\alpha+n-1}\sum_{i=0}^{m-1}(1-x)^i\mathrm{d}x \\
&\qquad = \frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\sum_{i=0}^{m-1}\int_0^1 x^{-\alpha+1-1}(1-x)^{\theta+\alpha+n+i-1}\mathrm{d}x \\
&\qquad = \frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\sum_{i=0}^{m-1}\frac{\Gamma(i+\theta+\alpha+n)\Gamma(1-\alpha)}{\Gamma(i+\theta+n+1)} \\
&\qquad = \frac{\Gamma(\theta+m+n)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\sum_{i=1}^{m}\frac{\Gamma(i+\theta+\alpha+n-1)\Gamma(1-\alpha)}{\Gamma(i+\theta+n)} \\
&\qquad = \frac{\Gamma(\theta+m+n)}{\Gamma(\theta+1)}\sum_{i=1}^{m}\frac{(\theta+\alpha)_{n+i-1}}{(\theta+1)_{n+i-1}} = (\theta+1)_{m+n-1}\sum_{i=1}^{m}\frac{(\theta+\alpha)_{n+i-1}}{(\theta+1)_{n+i-1}} \\
&\qquad = (\theta+1)_{n-1}(\theta+n)_m\sum_{i=1}^{m}\frac{(\theta+\alpha)_{n+i-1}}{(\theta+1)_{n+i-1}}
\end{aligned}$$

and the thesis for $\alpha\in[0,1)$ follows by substituting this expression in (3.26).

### 3.A.7 Proof of Proposition 3.2

The asymptotic behavior of $K_n$ follows by specializing the asymptotic distribution of $K_m^{(n)}$, determined in Proposition 3.3, when $n=0$ and $m$ is replaced with $n$. One may also prove this result by working along the same lines as in the proof of Proposition 3.3.

### 3.A.8 Proof of Proposition 3.3

We prove the theorem for $\alpha<0$, $\alpha=0$ and $\alpha\in(0,1)$ separately.
**Case $\boldsymbol{\alpha<0}$.** This choice corresponds to a mixture over $M$ of a BB model. We first observe that, in the case of a BB model with parameter $M$, the posterior distribution of the number of hitherto unseen features $K_m^{(n)}\,|\,\boldsymbol{Z}, M$ has a binomial distribution given by

$$K_m^{(n)}\,|\,\boldsymbol{Z}, M \sim \text{Binomial}\Big(M-k, 1-\frac{(\theta+\alpha+n)_m}{(\theta+n)_m}\Big).$$

This result is a simple consequence of Theorem 3.3 that can be obtained by substituting the expression of the $V_{n,k}$'s of a BB model with parameter $M$, displayed in Equation (3.3). Thus, by letting $p_m(\theta+n,\alpha)=1-(\theta+\alpha+n)_m/(\theta+n)_m$, the characteristic function of $K_m^{(n)}\,|\,\boldsymbol{Z}, M$, denoted here as $\Psi_m^{(n)}$, equals

$$\Psi_m^{(n)}(t) = \big[1-p_m(\theta+n,\alpha)+p_m(\theta+n,\alpha)e^{it}\big]^{M-k}, \tag{3.27}$$

for any $t \in \mathbb{R}$, and $i$ is the imaginary unit. Note that $\lim_{m\to\infty} p_m(\theta+n,\alpha) = 1$, indeed we have:

$$\begin{aligned} p_m(\theta+n,\alpha) &= 1 - \frac{(\theta+\alpha+n)_m}{(\theta+n)_m} = 1 - \frac{\Gamma(\theta+\alpha+n+m)}{\Gamma(\theta+\alpha+n)} \cdot \frac{\Gamma(\theta+n)}{\Gamma(\theta+n+m)} \\ &= 1 - \frac{\Gamma(\theta+n)}{\Gamma(\theta+\alpha+n)} \cdot m^{\alpha} + o(m^{\alpha}), \end{aligned}$$

where we have used the asymptotic expansion of ratio of gamma functions (Erdélyi and Tricomi, 1951), and the little-$o$ notation. Since $\alpha < 0$, the limit of $p_m(\theta+n,\alpha)$ goes to 1, as $m \to +\infty$. Thus, for all $t \in \mathbb{R}$, the limit of the characteristic function in (3.27) equals

$$\lim_{m\to\infty} \Psi_m^{(n)}(t) = e^{it(M-k)},$$

in other words, it holds $K_m^{(n)} \,|\, \boldsymbol{Z}, M \xrightarrow{\mathrm{d}} M-k$.
Now, consider a prior distribution for the parameter $M$ having probability mass function $p_M$, and denote by $\Phi_m^{(n)}$ the characteristic function of $K_m^{(n)} \,|\, \boldsymbol{Z}$ for this statistical model. We can now compute the limit of the characteristic function for any $t \in \mathbb{R}$:

$$\lim_{m\to\infty} \Phi_m^{(n)}(t) = \lim_{m\to\infty} \mathsf{E}\left[e^{itK_m^{(n)}} \,|\, \boldsymbol{Z}\right] = \lim_{m\to\infty} \mathsf{E}\left[\mathsf{E}\left[e^{itK_m^{(n)}} \,|\, \boldsymbol{Z}, M\right] \,|\, \boldsymbol{Z}\right]$$

and note that $\Psi_m^{(n)}(t) = \mathsf{E}\left[e^{itK_m^{(n)}} \,|\, \boldsymbol{Z}, M\right]$ is the characteristic function under a BB model. Since $|\, \Psi_m^{(n)}(t) \,| \leq 1$ and $\lim_{m\to\infty} \Psi_m^{(n)}(t) = e^{it(M-k)}$, the dominated convergence theorem implies

$$\lim_{m\to\infty} \Phi_m^{(n)}(t) = \mathsf{E}\left[e^{it(M-k)} \,|\, \boldsymbol{Z}\right].$$

Therefore, $K_m^{(n)} \,|\, \boldsymbol{Z}$ does converge in distribution to the random variable $M-k \,|\, \boldsymbol{Z} \stackrel{\mathrm{d}}{=} M'$, as $m \to +\infty$. Note that $M'$ is almost surely finite since $M$ is a priori almost surely finite, and the posterior distribution of $M$ equals

$$p_M(y \,|\, \boldsymbol{Z}) \propto \frac{y!}{(y-k)!} \left(\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^y \cdot \mathbb{1}_{\{k,k+1,\ldots\}}(y) p_M(y).$$

**Case $\alpha = 0$.** This case corresponds to a mixture over $\gamma$ of a two-parameter IBP model. We first assume that $\gamma$ is a fixed parameter, and we focus on the asymptotic behaviour of the sequence of random variables $K_m^{(n)}/\log(m) \,|\, \boldsymbol{Z}, \gamma$, as $m \to +\infty$. As a consequence of Theorem 3.3, by substituting the expression of the $V_{n,k}$'s of a two-parameter IBP displayed in Equation (3.2), it follows that

$$K_m^{(n)} \,|\, \boldsymbol{Z}, \gamma \sim \text{Poisson}\left(\gamma\theta \sum_{i=1}^{m} (n+\theta+i-1)^{-1}\right).$$

Denoting with $\Psi_m^{(n)}$ the characteristic function of $K_m^{(n)}/\log(m) \,|\, \boldsymbol{Z}, \gamma$, it holds that

$$\begin{aligned} \Psi_m^{(n)}(t) &= \mathsf{E}\left[\exp\left\{itK_m^{(n)}/\log(m)\right\} \,|\, \boldsymbol{Z}, \gamma\right] \\ &= \exp\left\{\gamma\theta \sum_{i=1}^{m} (n+\theta+i-1)^{-1} \cdot \left(e^{it/\log(m)} - 1\right)\right\}, \end{aligned}$$

for any $t \in \mathbb{R}$, and $i$ stands for the imaginary unit. We observe that, as $m \to +\infty$

$$\Psi_m^{(n)}(t) = \exp\left\{\gamma\theta \sum_{i=1}^{m} (n+\theta+i-1)^{-1} \cdot \left[it/\log(m) + \mathcal{O}(\log(m)^{-2})\right]\right\},$$

where we have used the big-$\mathcal{O}$ notation. Moreover, as $m \to +\infty$, it is easy to see that

$$\sum_{i=1}^{m} (n+\theta+i-1)^{-1}/\log(m) = 1 + \mathcal{O}(\log(m)^{-1}),$$

as a consequence, for any $t \in \mathbb{R}$, the characteristic function satisfies

$$\lim_{m\to\infty} \Psi_m^{(n)}(t) = e^{it\gamma\theta}.$$

In other words, we have shown that $K_m^{(n)}/\log(m) \,|\, \boldsymbol{Z}, \gamma \xrightarrow{\mathrm{d}} \gamma\theta$.

Now, we consider a prior distribution for the parameter $\gamma$ having density function $p_\gamma$, and we focus on the point-wise convergence of the characteristic function of the random variable $K_m^{(n)}/\log(m) \,|\, \boldsymbol{Z}$, denoted here as $\Phi_m^{(n)}$. To this end, let $t \in \mathbb{R}$ and evaluate the following limit

$$\begin{aligned}\lim_{m\to\infty} \Phi_m^{(n)}(t) &= \lim_{m\to\infty} \mathsf{E}\left[\exp\left\{itK_m^{(n)}/\log(m)\right\} \,|\, \boldsymbol{Z}\right] \\ &= \lim_{m\to\infty} \mathsf{E}\left[\mathsf{E}\left[\exp\left\{itK_m^{(n)}/\log(m)\right\} \,|\, \boldsymbol{Z}, \gamma\right] \,|\, \boldsymbol{Z}\right].\end{aligned}$$

Note that $\Psi_m^{(n)}(t) = \mathsf{E}\left[\exp\left\{itK_m^{(n)}/\log(m)\right\} \,|\, \boldsymbol{Z}, \gamma\right]$ is the characteristic function under the two-parameter IBP model. Since $|\,\Psi_m^{(n)}(t)\,| \leq 1$ and $\lim_{m\to\infty} \Psi_m^{(n)}(t) = e^{it\gamma\theta}$, the dominated convergence theorem implies

$$\lim_{m\to\infty} \Phi_m^{(n)}(t) = \mathsf{E}\left[e^{it\gamma\theta} \,|\, \boldsymbol{Z}\right],$$

therefore $K_m^{(n)}/\log(m) \,|\, \boldsymbol{Z}$ does converge in distribution to the random variable $\gamma\theta \,|\, \boldsymbol{Z}$, and the thesis follows.

**Case $\boldsymbol{\alpha \in (0,1)}$.** We consider the limit of the characteristic function of $K_m^{(n)}/m^\alpha \,|\, \boldsymbol{Z}$, denoted with $\Phi_m^{(n)}$, for a mixture over $\gamma$ of a three-parameter IBP model with parameters $(\gamma, \alpha, \theta)$:

$$\begin{aligned}\lim_{m\to\infty} \Phi_m^{(n)}(t) &= \lim_{m\to\infty} \mathsf{E}\left[\exp\left\{itK_m^{(n)}/m^\alpha\right\} \,|\, \boldsymbol{Z}\right] \\ &= \lim_{m\to\infty} \mathsf{E}\left[\mathsf{E}\left[\exp\left\{itK_m^{(n)}/m^\alpha\right\} \,|\, \boldsymbol{Z}, \gamma\right] \,|\, \boldsymbol{Z}\right].\end{aligned}$$

Note that $\mathsf{E}\left[\exp\left\{itK_m^{(n)}/m^\alpha\right\} \,|\, \boldsymbol{Z}, \gamma\right]$ corresponds to the characteristic function of the random variable $K_m^{(n)}/m^\alpha \,|\, \boldsymbol{Z}, \gamma$ for the three-parameter IBP. (Masoero et al., 2022, Proposition 2) show that $K_m^{(n)}/m^\alpha \,|\, \boldsymbol{Z}, \gamma \xrightarrow{a.s.} \xi$, with $\xi = \gamma\Gamma(\theta+1)/(\alpha\Gamma(\theta+\alpha))$. Thus, as $m \to \infty$, we get

$$\mathsf{E}\left[\exp\left\{itK_m^{(n)}/m^\alpha\right\} \,|\, \boldsymbol{Z}, \gamma\right] \longrightarrow e^{it\xi}.$$

By an application of the dominated convergence theorem, we obtain

$$\lim_{m\to\infty} \Phi_m^{(n)}(t) = \mathsf{E}\left[e^{it\xi} \,|\, \boldsymbol{Z}\right] = \mathsf{E}\left[\exp\left\{it\frac{\gamma\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)}\right\} \,|\, \boldsymbol{Z}\right] = \Phi_{\gamma\,|\,\boldsymbol{Z}}\left(t\frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)}\right)$$

where $\Phi_{\gamma\,|\,\boldsymbol{Z}}$ denotes the characteristic function of $\gamma \,|\, \boldsymbol{Z}$, and the thesis follows for $\alpha \in (0,1)$.

### 3.A.9 Proof of Theorem 3.4

We prove the theorem for $\alpha < 0$ and $\alpha \in [0, 1)$ separately.

**Case $\boldsymbol{\alpha < 0}$.**

We preface a Lemma to provide the hierarchical representation of the BB process. We remark that the following is an alternative construction of the BB model in Battiston et al. (2018) and Griffiths and Ghahramani (2011).

**Lemma 3.4.** *The BB model with parameters $(M, \alpha, \theta)$ may be equivalently described in the following hierarchical form*

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{iid}{\sim} BP(\tilde{\mu}), \qquad i \geq 1 \\ \tilde{\mu} &= \sum_{j=1}^{M} \tilde{q}_j \delta_{\tilde{X}_j}, \end{aligned} \tag{3.28}$$

*where $\tilde{q}_j$ are i.i.d. beta random variables with parameters $(-\alpha, \theta + \alpha)$ and $\tilde{X}_j \overset{iid}{\sim} G_0$, for $j = 1, \ldots, M$.*

*Proof.* We show that the EFPF induced by the model (3.28) is the EFPF in (3.1) with $V_{n,k}$'s in (3.3) characterizing the BB model. Assuming model (3.28), introduce the $n \times M$ binary matrix $\tilde{\boldsymbol{Z}}$, whose generic element $\tilde{Z}_{ij}$ is defined as $\tilde{Z}_{ij} = Z_i(\tilde{X}_j)$. In particular, $\tilde{Z}_{ij} = 1$ if subject $i$ possesses feature $j$, $\tilde{Z}_{ij} = 0$ otherwise. It holds that the entries of the matrix $\tilde{\boldsymbol{Z}}$ are distributed as

$$\begin{aligned} \tilde{Z}_{ij} \,|\, \tilde{q}_j &\overset{\text{ind}}{\sim} \text{Bernoulli}(\tilde{q}_j), \quad i \geq 1,\, j = 1, \ldots, M, \\ \tilde{q}_j &\overset{\text{iid}}{\sim} \text{Beta}(-\alpha, \theta + \alpha), \quad j = 1, \ldots, M. \end{aligned}$$

Indeed, let $\tilde{\boldsymbol{z}} \in \{0, 1\}^{n \times M}$ a generic realization of $\tilde{\boldsymbol{Z}}$, then

$$\mathsf{P}(\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}} \,|\, \tilde{q}_1, \ldots, \tilde{q}_M) = \prod_{j=1}^{M} \prod_{i=1}^{n} \tilde{q}_j^{\tilde{z}_{ij}} (1 - \tilde{q}_j)^{1 - \tilde{z}_{ij}} = \prod_{j=1}^{M} \tilde{q}_j^{m_j} (1 - \tilde{q}_j)^{n - m_j}, \tag{3.29}$$

where $m_j = \sum_{i=1}^{n} \tilde{z}_{ij}$. The marginal distribution of $\tilde{\boldsymbol{Z}}$ results in

$$\begin{aligned} \mathsf{P}(\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}}) &= \int \mathsf{P}(\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}} \,|\, \tilde{q}_1, \ldots, \tilde{q}_M) \cdot \prod_{j=1}^{M} \tilde{q}_j^{-\alpha - 1} (1 - \tilde{q}_j)^{\alpha + \theta - 1} \mathrm{d}\tilde{q}_1 \cdots \mathrm{d}\tilde{q}_M \\ &= \left( \frac{\Gamma(\theta)}{\Gamma(-\alpha)\Gamma(\theta + \alpha)\Gamma(n + \theta)} \right)^M \prod_{j=1}^{M} \Gamma(m_j - \alpha)\Gamma(n - m_j + \alpha + \theta). \end{aligned}$$

The EFPF corresponding to $\tilde{\boldsymbol{z}}$, can be evaluated by taking into account the $\binom{M}{k}$ possible ways to place the null columns of $\tilde{\boldsymbol{z}}$, thus obtaining:

$$\begin{aligned} \pi_n(m_1, \ldots, m_k) &= \binom{M}{k} \mathsf{P}(\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}}) \\ &= \binom{M}{k} \left( \frac{\Gamma(\theta)}{\Gamma(-\alpha)\Gamma(\theta + \alpha)\Gamma(n + \theta)} \right)^M \prod_{j=1}^{M} \Gamma(m_j - \alpha)\Gamma(n - m_j + \alpha + \theta), \end{aligned}$$

which corresponds to the EFPF in (3.1) with $V_{n,k}$'s as in (3.3). □

We now need to provide a posterior representation of the latent measure $\tilde{\mu}$ in Lemma 3.4. In particular, assume that $k$ features have been observed in $\boldsymbol{Z}$, with labels $X_1, \ldots, X_k$. Moreover, assume that each observed feature $X_\ell$ has been displayed in $m_\ell$ subjects, for $\ell = 1, \ldots, k$. Among the $M$ atoms of $\tilde{\mu}$ in model (3.28), $k$ atoms of the posterior random measure $\tilde{\mu} \,|\, \boldsymbol{Z}$ are necessarily $X_1, \ldots, X_k$. The remaining $M - k$ atoms of $\tilde{\mu} \,|\, \boldsymbol{Z}$ are drawn independently from the prior. Specifically, the posterior distribution of $\tilde{\mu} \,|\, \boldsymbol{Z}$ may be characterized as

$$\tilde{\mu} \,|\, \boldsymbol{Z} \stackrel{\mathrm{d}}{=} \mu' + \mu^*,$$

where $\mu^* = \sum_{\ell=1}^{k} q_\ell \delta_{X_\ell}$ accounts for the $k$ observed features and $\mu' = \sum_{j=1}^{M-k} q'_j \delta_{\tilde{X}'_j}$ describes the remaining $M - k$ possible features, with $\tilde{X}'_j \stackrel{\mathrm{iid}}{\sim} G_0$, for $j = 1, \ldots, M - k$. The joint distribution of $q'_j$'s and $q_\ell$'s can be obtained by multiplying the likelihood function (3.29) with the prior distribution of the $\tilde{q}_j$'s, which are independent beta random variables with parameters $(-\alpha, \theta + \alpha)$. Thus, an application of the Bayes theorem leads to $q_\ell \stackrel{\mathrm{ind}}{\sim} \mathrm{Beta}(m_\ell - \alpha, \alpha + \theta + n - m_\ell)$, as $\ell = 1, \ldots, k$, and $q'_j \stackrel{\mathrm{iid}}{\sim} \mathrm{Beta}(-\alpha, \alpha + \theta + n)$, for $j = 1, \ldots, M - k$.

Point (i) of Theorem 3.4, valid for any mixture of BBs, follows by conditioning on $M$ and applying the just discussed distributional equality for the posterior distribution of $\tilde{\mu}$.
**Case $\boldsymbol{\alpha} \in [\mathbf{0}, \mathbf{1})$.** This corresponds to the IBP case, where $\tilde{\mu} \,|\, \gamma$ is a stable-beta process with intensity measure given by (3.7), namely:

$$\rho(\mathrm{d}s) = \gamma \frac{\Gamma(1 + \theta)}{\Gamma(1 - \alpha)\Gamma(\theta + \alpha)} s^{-\alpha - 1}(1 - s)^{\theta + \alpha - 1} \mathrm{d}s. \tag{3.30}$$

Thus, conditionally on $\gamma$, the posterior representation in (ii) follows from (James, 2017, Theorem 3.1), which provides a characterization for general CRMs. More precisely, one has

$$\tilde{\mu} \,|\, \boldsymbol{Z} \stackrel{\mathrm{d}}{=} \mu^* + \mu'$$

where $\mu^*$ and $\mu'$ are independent random measures such that:

(1) $\mu' \sim \mathrm{CRM}(\rho'; G_0)$, with $\rho'(\mathrm{d}s) = (1 - s)^n \rho(\mathrm{d}s)$;

(2) the random measure $\mu^* = \sum_{\ell=1}^{k} q_\ell \delta_{X_\ell}$ is almost surely discrete, where the $X_\ell$'s are the distinct feature labels out of the observed sample $\boldsymbol{Z}$, and $q_\ell$'s are independent random variables with density

$$f_{q_\ell}(\mathrm{d}s) \propto (1 - s)^{n - m_\ell} s^{m_\ell} \rho(\mathrm{d}s)$$

as $\ell = 1, \ldots, k$.

By substituting the specific expression of $\rho$ (3.30) in the previous characterization, part (ii) of the theorem easily follows.

## 3.B Proofs of Section 3.3

### 3.B.1 Details for the determination of (3.11)

Here we show that the gamma mixture of IBPs has a product form EFPF with weights $V_{n,k}$'s as in (3.11). To this end, we integrate the product form EFPF in (3.1)-(3.2) with respect

to the parameter $\gamma \sim \text{Gamma}(a, b)$:

$$\begin{aligned}
\pi_n(m_1, \ldots, m_k) &= \int_0^\infty \frac{1}{k!} \left( \frac{\gamma}{(\theta+1)_{n-1}} \right)^k e^{-\gamma g_n(\theta,\alpha)} \prod_{\ell=1}^k (1-\alpha)_{m_\ell - 1} (\theta+\alpha)_{n-m_\ell} \frac{b^a}{\Gamma(a)} \gamma^{a-1} e^{-\gamma b} \mathrm{d}\gamma \\
&= \frac{b^a}{k!\Gamma(a)\{(\theta+1)_{n-1}\}^k} \int_0^\infty \gamma^{k+a-1} \exp\{-\gamma(g_n(\theta,\alpha)+b)\} \mathrm{d}\gamma \prod_{\ell=1}^k (1-\alpha)_{m_\ell - 1} (\theta+\alpha)_{n-m_\ell} \\
&= \frac{b^a}{k!\Gamma(a)\{(\theta+1)_{n-1}\}^k} \cdot \frac{\Gamma(k+a)}{(g_n(\theta,\alpha)+b)^{k+a}} \prod_{\ell=1}^k (1-\alpha)_{m_\ell - 1} (\theta+\alpha)_{n-m_\ell} \\
&= \frac{b^a \, (a)_k}{k!\{(\theta+1)_{n-1}\}^k (g_n(\theta,\alpha)+b)^{k+a}} \prod_{\ell=1}^k (1-\alpha)_{m_\ell - 1} (\theta+\alpha)_{n-m_\ell}.
\end{aligned}$$

It is easy to observe that the last expression is an EFPF of type (3.1) where

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

### 3.B.2 Proof of Proposition 3.4

The distributions of $K_n$ in (3.12) follow by specializing Theorem 3.2 with the two specifications (3.2) and (3.11) for $V_{n,k}$. Analogously, the posterior distributions for $K_m^{(n)}$ in (3.13) and (3.14) follow from Theorem 3.3.

### 3.B.3 Proof of Proposition 3.5

From Equation (3.9), it is easy to see that the posterior distribution of $\gamma \,|\, \boldsymbol{Z}$ is a gamma with parameters $(k+a, g_n(\theta,\alpha)+b)$, since $p_\gamma$ is a gamma density with parameters $(a, b)$. Thus, we also remind that the characteristic function of $\gamma \,|\, \boldsymbol{Z}$ equals

$$\Phi_{\gamma \,|\, \boldsymbol{Z}}(t) = \left( \frac{g_n(\theta,\alpha)+b}{g_n(\theta,\alpha)+b-it} \right)^{k+a}.$$

We now consider the case $\alpha = 0$ and $\alpha \in (0,1)$ separately.

**Case $\boldsymbol{\alpha \in (0,1)}$.** By an application of Proposition 3.3, $K_m^{(n)}/m^\alpha \,|\, \boldsymbol{Z}$ converges in distribution to $\gamma\Gamma(\theta+1)/(\alpha\Gamma(\theta+\alpha)) \,|\, \boldsymbol{Z}$, whose characteristic function equals

$$\Phi_{\gamma \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)} \,|\, \boldsymbol{Z}}(t) = \Phi_{\gamma \,|\, \boldsymbol{Z}} \left( t \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)} \right) = \left( \frac{g_n(\theta,\alpha)+b}{g_n(\theta,\alpha)+b-it\frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)}} \right)^{k+a}$$

which is the characteristic function of a gamma random variable with parameters $(k+a, (g_n(\theta,\alpha)+b)\Gamma(\theta+\alpha)\alpha/\Gamma(\theta+1))$, as stated.

**Case $\boldsymbol{\alpha = 0}$.** By virtue of Proposition 3.3, one has that $K_m^{(n)}/\log(m) \,|\, \boldsymbol{Z}$ converges in distribution to $\gamma\theta \,|\, \boldsymbol{Z}$, whose characteristic function equals

$$\Phi_{\gamma\theta \,|\, \boldsymbol{Z}}(t) = \Phi_{\gamma \,|\, \boldsymbol{Z}}(t\theta) = \left( \frac{g_n(\theta,0)+b}{g_n(\theta,0)+b-it\theta} \right)^{k+a}$$

which is the characteristic function of a gamma random variable with parameters $(k+a, (g_n(\theta,0)+b)/\theta)$, as stated.

### 3.B.4 Proof of Equation (3.16)

We need to show the equivalence between the hierarchical formulation of the gamma mixture of IBPs and the formulation (3.16), which relies on the negative binomial process. More specifically, we consider the stable-beta process with parameters $(\gamma, \alpha, \theta)$, i.e., $\tilde{\mu} \,|\, \gamma \sim \text{CRM}(\rho'; G_0)$, where $\rho'$ is as in (3.7)

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

and we choose a gamma prior with parameters $(a, b)$ for $\gamma$. The Laplace functional of the resulting random measure $\tilde{\mu}$ equals

$$\begin{aligned}\mathcal{L}_{\tilde{\mu}}(g) &= \mathsf{E}\left[e^{-\int_{\mathbb{X}} g(x)\tilde{\mu}(\mathrm{d}x)}\right] = \mathsf{E}\left[\mathsf{E}\left[e^{-\int_{\mathbb{X}} g(x)\tilde{\mu}(\mathrm{d}x)} \,|\, \gamma\right]\right] = \mathsf{E}\left[\exp\left\{-\int_0^1 \int_{\mathbb{X}} \left(1-e^{-sg(x)}\right)\rho'(\mathrm{d}s)G_0(\mathrm{d}x)\right\}\right] \\ &= \int_0^\infty \exp\left\{-\gamma \int_0^1 \int_{\mathbb{X}} \left(1-e^{-sg(x)}\right) \frac{\Gamma(1+\theta)s^{-\alpha-1}(1-s)^{\theta+\alpha-1}}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\mathrm{d}sG_0(\mathrm{d}x)\right\} \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}\mathrm{d}\gamma,\end{aligned}$$

for any measurable function $g : \mathbb{X} \to \mathbb{R}_+$. Thus, by integrating with respect to $\gamma$, the Laplace functional of $\tilde{\mu}$ boils down to

$$\mathcal{L}_{\tilde{\mu}}(g) = \left(1 + \int_0^1 \int_{\mathbb{X}} \left(1-e^{-sg(x)}\right) \frac{1}{b} \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}sG_0(\mathrm{d}x)\right)^{-a}.$$

Therefore, $\tilde{\mu}$ is distributed as a negative binomial process with parameters $(a, \rho; G_0)$, where

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

### 3.B.5 Proof of Corollary 3.2

It is sufficient to specialize part (ii) of Theorem 3.4 and to show that the measure $\mu'$, which appears in the posterior representation, is a negative binomial process. In order to do this, we now compute the Laplace functional of $\mu'$, and we show it coincides with the Laplace functional of a negative binomial process of type (3.15). For any measurable function $g : \mathbb{X} \to \mathbb{R}_+$, we evaluate

$$\begin{aligned}&\mathsf{E}\left[e^{-\int_{\mathbb{X}} g(x)\mu'(\mathrm{d}x)}\right] = \mathsf{E}\left[\mathsf{E}\left[e^{-\int_{\mathbb{X}} g(x)\mu'(\mathrm{d}x)} \,|\, \gamma'\right]\right] \\ &\quad = \mathsf{E}\left[\exp\left\{-\gamma' \int_{\mathbb{X}} \int_0^1 \left(1-e^{-sg(x)}\right) \frac{\Gamma(1+\theta)s^{-\alpha-1}(1-s)^{\theta+\alpha+n-1}\mathrm{d}s}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} G_0(\mathrm{d}x)\right\}\right],\end{aligned}$$

where we used the fact that $\mu' \,|\, \gamma'$ is a CRM characterized by the intensity measure

$$\gamma' \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} s^{-\alpha-1}(1-s)^{n+\theta+\alpha-1}\mathrm{d}s.$$

We now integrate out $\gamma'$, which equals in distribution to the posterior $\gamma \,|\, \boldsymbol{Z} \sim \text{Gamma}(a+k, b+g_n(\theta,\alpha))$. Therefore, the Laplace functional under study boils down to

$$\begin{aligned}
&\mathsf{E}\left[e^{-\int_{\mathbb{X}} g(x)\mu'(\mathrm{d}x)}\right] \\
&\quad= \int_0^\infty \exp\left\{-\gamma' \int_0^1 \int_{\mathbb{X}} \left(1-e^{-sg(x)}\right) \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} s^{-\alpha-1}(1-s)^{\theta+\alpha+n-1}\mathrm{d}sG_0(\mathrm{d}x)\right\} \\
&\qquad\qquad \times \frac{(g_n(\theta,\alpha)+b)^{k+a}}{\Gamma(k+a)} \gamma^{k+a-1} e^{-(g_n(\theta,\alpha)+b)\gamma}\mathrm{d}\gamma \\
&\quad= \left(1+\int_0^1 \int_{\mathbb{X}} \left(1-e^{-sg(x)}\right)\rho'(\mathrm{d}s)G_0(\mathrm{d}x)\right)^{-k-a}, \qquad (3.31)
\end{aligned}$$

where we have set

$$\rho'(\mathrm{d}s) := (g_n(\theta,\alpha)+b)^{-1} \cdot \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} s^{-\alpha-1}(1-s)^{\theta+\alpha+n-1}\mathrm{d}s.$$

We note that the expression in (3.31) is the Laplace functional of a negative binomial process (3.15) with parameters $(k+a, \rho'; G_0)$. Thus, the thesis follows.

## 3.C Proofs of Section 3.4

### 3.C.1 Proof of Proposition 3.6

We first concentrate on the proof of (3.17). We integrate the EFPF of a BB process (3.3) with respect to $M \sim \text{Poisson}(\lambda)$:

$$\begin{aligned}
\pi_n(m_1,\ldots,m_k) &= \sum_{M\geq k} \binom{M}{k} \left(\frac{-\alpha}{(\theta+\alpha)_n}\right)^k \left(\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^M \prod_{\ell=1}^k (1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell} \cdot e^{-\lambda}\frac{\lambda^M}{M!} \\
&= \frac{1}{k!}e^{-\lambda}\lambda^k \left(\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^k \left(\frac{-\alpha}{(\theta+\alpha)_n}\right)^k \\
&\qquad \times \prod_{\ell=1}^k (1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell} \sum_{M\geq k} \frac{1}{(M-k)!}\lambda^{M-k}\left(\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^{M-k} \\
&= \frac{1}{k!}e^{-\lambda}\lambda^k \left(\frac{(\theta+\alpha)_n}{(\theta)_n}\right)^k \left(\frac{-\alpha}{(\theta+\alpha)_n}\right)^k \\
&\qquad \times \prod_{\ell=1}^k (1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell} \exp\left\{\lambda\frac{(\theta+\alpha)_n}{(\theta)_n}\right\} \\
&= \frac{1}{k!}\left\{\frac{-\alpha\lambda}{(\theta)_n}\right\}^k \exp\left\{-\lambda\left(1-\frac{(\theta+\alpha)_n}{(\theta)_n}\right)\right\} \prod_{\ell=1}^k (1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}.
\end{aligned}$$

It is now easy to realize that

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

where $V_{n,k}$ is as in (3.17).

Now, we move to the proof of (3.18). As before, we consider the EFPF in product form of a BB process, as specified in (3.3), and we integrate out $M$, distributed as a negative binomial random variable with parameters $(n_0, \mu_0)$. Throughout the computation, we also set $p = n_0/(\mu_0 + n_0)$ so that the negative binomial is parametrized with respect to $n_0$ and the success probability $p$. The EFPF equals

$$\begin{aligned}\pi_n(m_1,\ldots,m_k) &= \sum_{M\geq k}\binom{M}{k}\Big(\frac{-\alpha}{(\theta+\alpha)_n}\Big)^k\Big(\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^M\prod_{\ell=1}^k(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}\\ &\qquad\times\binom{M+n_0-1}{M}p^{n_0}(1-p)^M\\ &= \prod_{\ell=1}^k(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}\cdot p^{n_0}\Big(\frac{-\alpha}{(\theta+\alpha)_n}\Big)^k\\ &\qquad\times\frac{1}{k!(n_0-1)!}\sum_{M\geq k}\frac{(M+n_0-1)!}{(M-k)!}\Big((1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^M,\end{aligned}$$

where we have simply rearranged all the terms. By a simple change of variable we get

$$\begin{aligned}\pi_n(m_1,\ldots,m_k) &= \prod_{\ell=1}^k(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}\cdot p^{n_0}\Big(\frac{-\alpha}{(\theta+\alpha)_n}\Big)^k\\ &\qquad\times\frac{1}{k!(n_0-1)!}\sum_{M\geq 0}\frac{(M+k+n_0-1)!}{M!}\Big((1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^{M+k}.\end{aligned}$$

Since $\alpha < 0$ and $\alpha + \theta > 0$, we have

$$0 < (1-p)\frac{(\theta+\alpha)_n}{(\theta)_n} < 1,$$

thus, it is easy to rearrange all the terms and recognize the probability mass function of a negative binomial random variable in the summation over $M$. More precisely, we have

$$\begin{aligned}\pi_n(m_1,\ldots,m_k) &= \prod_{\ell=1}^k(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}\cdot p^{n_0}\Big(\frac{-\alpha}{(\theta+\alpha)_n}\Big)^k\\ &\qquad\times\frac{(k+n_0-1)!}{k!(n_0-1)!}\Big((1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^k\sum_{M\geq 0}\binom{M+k+n_0-1}{M}\Big((1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^M\\ &= \prod_{\ell=1}^k(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}\cdot p^{n_0}\Big(\frac{-\alpha}{(\theta+\alpha)_n}\Big)^k\Big((1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^k\\ &\qquad\times\binom{k+n_0-1}{k}\Big(1-(1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^{-k-n_0}\\ &\qquad\times\sum_{M\geq 0}\binom{M+k+n_0-1}{M}\Big((1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^M\Big(1-(1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^{k+n_0}.\end{aligned}$$

The sum in the last expression is equal to 1 because this is the sum of the probability masses of a negative binomial distribution with dispersion parameter $k + n_0$ and success

probability $1-(1-p)(\theta+\alpha)_n/(\theta)_n$. Thus, we get

$$\begin{aligned}\pi_n(m_1,\dots,m_k) &= \prod_{\ell=1}^{k}(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}\cdot p^{n_0}\Big(\frac{-\alpha}{(\theta+\alpha)_n}\Big)^k\Big((1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^k \\ &\quad\times\binom{k+n_0-1}{k}\Big(1-(1-p)\frac{(\theta+\alpha)_n}{(\theta)_n}\Big)^{-k-n_0} \\ &= V_{n,k}\prod_{\ell=1}^{k}(1-\alpha)_{m_\ell-1}(\theta+\alpha)_{n-m_\ell}\end{aligned}$$

which is exactly in product form and $V_{n,k}$ is as in Equation (3.18).

### 3.C.2 Proof of Proposition 3.7

The distributions of $K_n$ follow by specializing Theorem 3.2 with the specifications of the weights $V_{n,k}$'s for the three models. Analogously, the posterior distributions of $K_m^{(n)}$ follow from Theorem 3.3.

### 3.C.3 Proof of Proposition 3.8

It follows from the general Theorem 3.3 by considering the case $\alpha<0$. The posterior distribution of $M\,|\,\boldsymbol{Z}$ easily follows from (3.9), by plugging-in the correct prior probability mass function $p_M$. Specifically, for $M\sim\text{Poisson}(\lambda)$, $p_M$ is a Poisson probability mass function; for $M\sim\text{NegBinomial}(n_0,\mu_0)$, $p_M$ is a negative binomial probability mass function.

### 3.C.4 Proofs of hierarchical representations (3.20) and (3.21)

Using the hierarchical representation of the BB model provided in Lemma 3.4, we first note that the Laplace functional of the BB model with parameters $(M,\alpha,\theta)$ can be written as

$$\mathcal{L}_{\tilde\mu}(g)=\mathsf{E}\left[e^{-\int_{\mathbb{X}}g(x)\tilde\mu(\mathrm{d}x)}\right]=\left[\int_{\mathbb{X}}\int_0^1 e^{-sg(x)}\frac{1}{B(-\alpha,\theta+\alpha)}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}sG_0(\mathrm{d}x)\right]^M, \tag{3.32}$$

for any measurable function $g:\mathbb{X}\to\mathbb{R}_+$.

Consequently, in order to derive representation (3.20), it is sufficient to integrate out $M\sim\text{Poisson}(\lambda)$ in the expression of the Laplace functional (3.32):

$$\begin{aligned}\mathcal{L}_{\tilde\mu}(g) &= e^{-\lambda}\sum_{M=0}^{\infty}\frac{1}{M!}\left[\lambda\int_{\mathbb{X}}\int_0^1 e^{-sg(x)}\frac{1}{B(-\alpha,\theta+\alpha)}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}sG_0(\mathrm{d}x)\right]^M \\ &= \exp\left\{-\lambda\left[1-\int_{\mathbb{X}}\int_0^1 e^{-sg(x)}\frac{1}{B(-\alpha,\theta+\alpha)}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}sG_0(\mathrm{d}x)\right]\right\} \\ &= \exp\left\{-\lambda\int_{\mathbb{X}}\int_0^1(1-e^{-sg(x)})\frac{1}{B(-\alpha,\theta+\alpha)}s^{-\alpha-1}(1-s)^{\theta+\alpha-1}\mathrm{d}sG_0(\mathrm{d}x)\right\}\end{aligned}$$

and this is the Laplace functional of a CRM with intensity measure as in (3.20). Similarly, for the negative binomial prior on $M$, we end up with the Laplace functional of a negative binomial process with parameters as in (3.21).

### 3.C.5 Proof of Corollary 3.3

The corollary follows from Theorem 3.4 by specializing the result for $\alpha < 0$ in the case of a Poisson or a negative binomial prior for $M$. We first concentrate on the case $M \sim \text{Poisson}(\lambda)$. Since in the posterior representation (3.10), $\mu'$ and $\mu^*$ are independent, we need only to evaluate the Laplace functional of $\mu'$. Observe that $\mu' \,|\, M' \stackrel{\text{d}}{=} \sum_{j=1}^{M'} q'_j \delta_{\tilde{X}'_j}$ can be seen as the random measure associated with a BB model with parameters $(M', \alpha, \theta+n)$, moreover $M' + k \stackrel{\text{d}}{=} M \,|\, \boldsymbol{Z}$. Therefore, by a suitable adaptation of Equation (3.32), the Laplace functional of $\mu'$, for an arbitrary measurable function $g : \mathbb{X} \to \mathbb{R}_+$, equals

$$\begin{aligned}\mathcal{L}_{\mu'}(g) &= \mathsf{E}\left[\mathsf{E}\left[e^{-\int_{\mathbb{X}} g(x)\mu'(\mathrm{d}x)} \,|\, M'\right]\right] \\ &= \sum_{t=0}^{\infty}\left[\int_{\mathbb{X}}\int_0^1 e^{-sg(x)}\frac{1}{B(-\alpha,\theta+n+\alpha)}s^{-\alpha-1}(1-s)^{\theta+n+\alpha-1}\mathrm{d}s G_0(\mathrm{d}x)\right]^t p_{M'}(t)\end{aligned} \tag{3.33}$$

where $p_{M'}$ denotes the probability mass function of $M'$. The posterior of $M'$ has been characterized in Proposition 3.8, where we showed that $M' \sim \text{Poisson}\left(\lambda(1-p_n(\theta,\alpha))\right.$, with $p_n(\theta,\alpha) = 1 - (\theta+\alpha)_n/(\theta)_n$. Thus, we get

$$\mathcal{L}_{\mu'}(g) = \exp\left\{-\int_{\mathbb{X}}\int_0^1 (1-e^{-sg(x)})\frac{\lambda}{B(-\alpha,\theta+n+\alpha)}\frac{(\theta+\alpha)_n}{(\theta)_n}s^{-\alpha-1}(1-s)^{\theta+n+\alpha-1}\mathrm{d}s G_0(\mathrm{d}x)\right\},$$

that is the Laplace functional of a CRM with intensity measure

$$\rho'(\mathrm{d}s) = \lambda \cdot \frac{\Gamma(\theta)}{\Gamma(-\alpha)\Gamma(\theta+\alpha)}s^{-\alpha-1}(1-s)^{n+\theta+\alpha-1}\mathrm{d}s.$$

When $M$ has a negative binomial distribution, then $p_{M'}$ in (3.33) coincides with a negative binomial distribution as specified in (3.19), i.e., for $t = 0, 1, \ldots$,

$$p_{M'}(t) = \binom{t+n_0+k-1}{t}(1-p_n^*(\theta,\alpha))^t(p_n^*(\theta,\alpha))^{n_0+k},$$

where we have set

$$p_n^*(\theta,\alpha) := 1 - (1-p_n(\theta,\alpha))\frac{\mu_0}{\mu_0+n_0}.$$

Thus, simple calculations show that the Laplace functional of $\mu'$ boils down to

$$\mathcal{L}_{\mu'}(g) = \left\{1 + \int_0^1\int_{\mathbb{X}}(1-e^{-sg(x)})\frac{1}{B(-\alpha,\theta+n+\alpha)}\frac{1-p_n^*(\theta,\alpha)}{p_n^*(\theta,\alpha)}s^{-\alpha-1}(1-s)^{\theta+n+\alpha-1}\mathrm{d}s G_0(\mathrm{d}x)\right\}^{-(n_0+k)}.$$

We recognize that the previous Laplace functional is the one of a negative binomial process $\text{NB}(n_0+k, \rho'; G_0)$, with

$$\rho'(\mathrm{d}s) = \frac{\Gamma(\theta)}{\Gamma(-\alpha)\Gamma(\theta+\alpha)} \cdot \frac{\frac{\mu_0}{\mu_0+n_0}}{1-\frac{\mu_0}{\mu_0+n_0}\frac{(\theta+\alpha)_n}{(\theta)_n}}s^{-\alpha-1}(1-s)^{n+\theta+\alpha-1}\mathrm{d}s.$$

Thus, with simple rearrangements of the terms one may realize that the intensity $\rho'$ is the one specified in the statement of the corollary, i.e.,

$$\rho'(\mathrm{d}s) = \frac{\Gamma(\theta)}{\Gamma(-\alpha)\Gamma(\theta+\alpha)} \cdot \frac{1}{n_0/\mu_0 + p_n(\theta,\alpha)}s^{-\alpha-1}(1-s)^{n+\theta+\alpha-1}\mathrm{d}s.$$

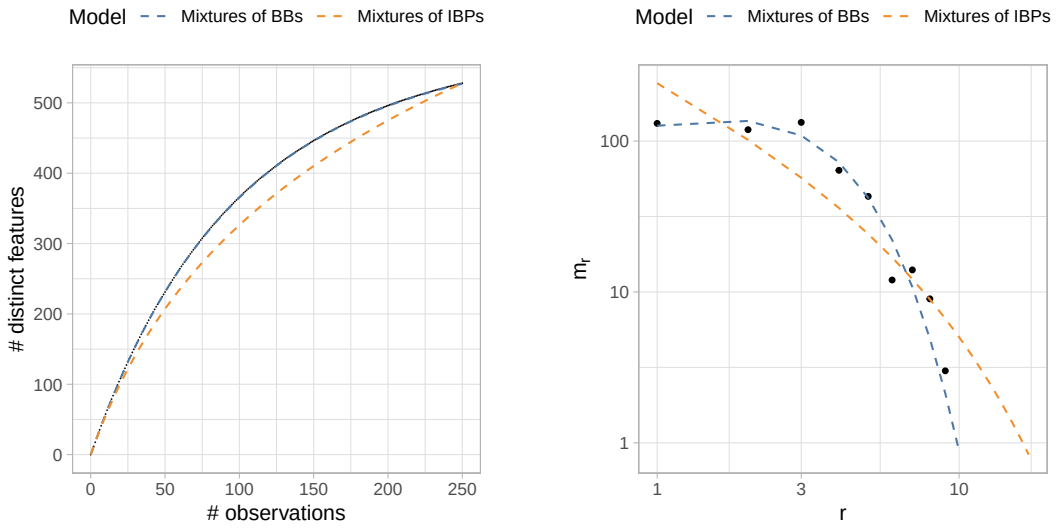



Figure 3.D.1: Case $n = 250$. Left panel: the empirical accumulation curve (black dots) and the rarefaction curve of the models (blue and orange dashed lines). Right panel: the observed values of $K_{n,r}$ (black dots) compared with the expected curve $\mathsf{E}(K_{n,r})$ of the models (blue and orange 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).

## 3.D Simulation studies

### 3.D.1 Simulation study A

We consider a true data generating process with $H = 600$ total features, having occurrence probabilities $\pi_k$, $k = 1, \ldots, H$, set as follows: 200 features have $\pi_k = 0.005$, 200 features have $\pi_k = 0.01$ and 200 features have $\pi_k = 0.015$. We generate a dataset with increasing dimensions for the training set $n \in \{50, 250, 1250\}$. In Figure 3.D.1, we apply the visual approach to model-checking for the case $n = 250$. As it is evident from the plots, the mixtures of IBPs are definitely not a good model for such data. Conversely, we argue that the mixtures of BBs can be assumed to be correctly specified here. This preference for the mixtures of BBs is further supported by the comparison of deviances: $D(\hat{\boldsymbol{\theta}}) = 16406.7$ for the BB model versus $D(\hat{\boldsymbol{\theta}}) = 16603.4$ for the IBP model. Therefore, we select the mixtures of BBs for the inference and prediction. For completeness, Figure 3.D.2 reports the 95% credible intervals of $K_1, \ldots, K_n$ and of $K_{n,r}$, $r \geq 1$, for the Poisson and two examples of negative binomial mixture of BBs. It is apparent that the width of the credible intervals increases as the prior variance on $M$ increases.

Focusing on the prediction of the number of unseen features in an additional sample of increasing size, we compare the Poisson mixture of BBs, the negative binomial mixture of BBs, the frequentist estimator (Chao) in Chao et al. (2014) and a variation of the Good–Toulmin estimators (GT) in Chakraborty et al. (2019). For the negative binomial mixture of BBs, we analyse two distinct choices for the prior variance of $M$, specifically $\text{Var}(M) = \mu_0 \times c$, with $c \in \{10, 100\}$. We report the comparison in Figure 3.D.3, for $n \in \{50, 250\}$. In the case $n = 50$, the point-wise estimates produced by our mixtures of BBs are rather good, definitely outperforming the ones obtained with the frequentist

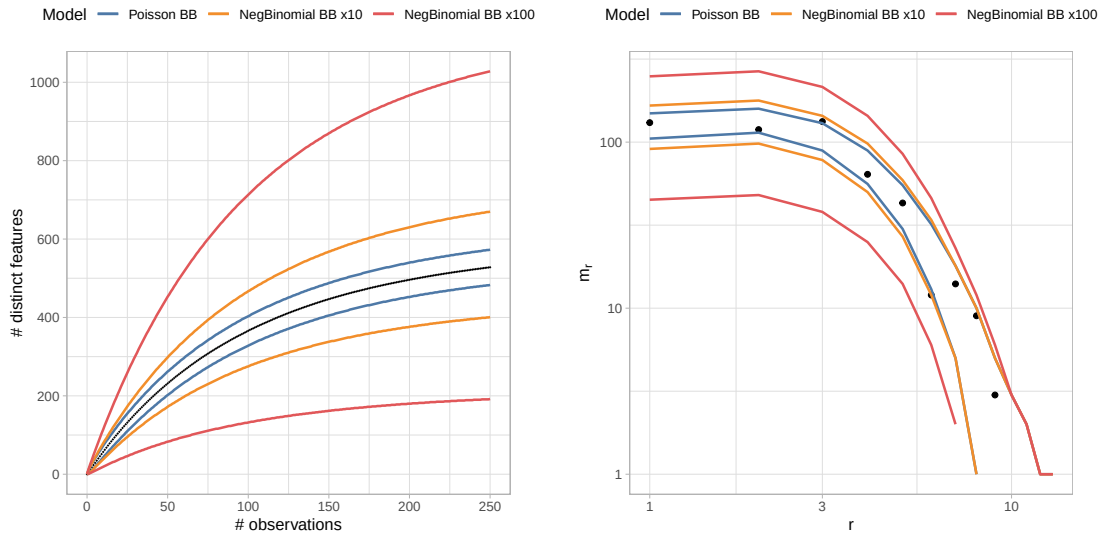



Figure 3.D.2: Case $n = 250$. Left panel: the empirical accumulation curve (black dots) and the credible intervals (delimited by colored lines) of $K_1, \ldots, K_n$ for the Poisson mixture of BBs and two examples of negative binomial mixture of BBs. Right panel: the observed values of $K_{n,r}$ (black dots) and the credible intervals (delimited by colored lines) of $K_{n,r}$ for the same mixtures of BBs. The right plot is in log-log scale.

estimator in Chao et al. (2014), which underestimates the extrapolation curve. While the Good–Toulmin estimates seem pretty reliable in the case $n = 50$, such an estimator shows bad predictive performance and well-known stability issues for $n = 250$. Moreover, the mixtures of BBs allow us to quantify the uncertainty around the point estimates through credible intervals: this is a remarkable advantage of our Bayesian framework. Observe that the credible bands contain the observed curve in the test set; for $n = 50$ it is evident how the negative binomial mixtures account for larger dispersion than the Poisson mixture, having wider credible intervals. For completeness, the sanity check in the larger sample $n = 1250$, where $k = 599$ features are observed, is satisfactory: the credible interval of $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, for $m = 300$, is $[599, 601]$, for all the considered mixtures of BBs. Finally, we remark that, while both the frequentist estimators are exclusively designed for the specific extrapolation problem, our model-based approach through mixtures of BBs offers a coherent and self-contained framework for a number of inference problems, where the prediction of the number of unseen features is just one available example.

Additionally, mixtures of BBs also allow estimation of the richness. We refer to Figure 3.D.4 for an insightful illustration on the richness estimation via the Poisson mixture and two negative binomial mixtures with $\text{Var}(M) = \mu_0 \times c$, for $c \in \{10, 100\}$. For the purpose of discussion, we also consider the standard BB model: interpret it as a mixture of BBs where the prior on $M$ is a point mass on the value $M$. Specifically, Figure 3.D.4 shows, for increasing sample sizes $n \in \{50, 250, 1250\}$, the posterior mean of the species richness $M$, for different choices of the parameters of the prior distribution on $M$. In particular, for each model, we estimate the parameters following the usual empirical Bayes procedure described in Section 3.5.1 (referred to as EB in the figure); moreover, we fix different choices for the prior mean of $M$, i.e., $\mathsf{E}(M) \in \{200, 400, 800\}$, and we estimate the parameters $\alpha$

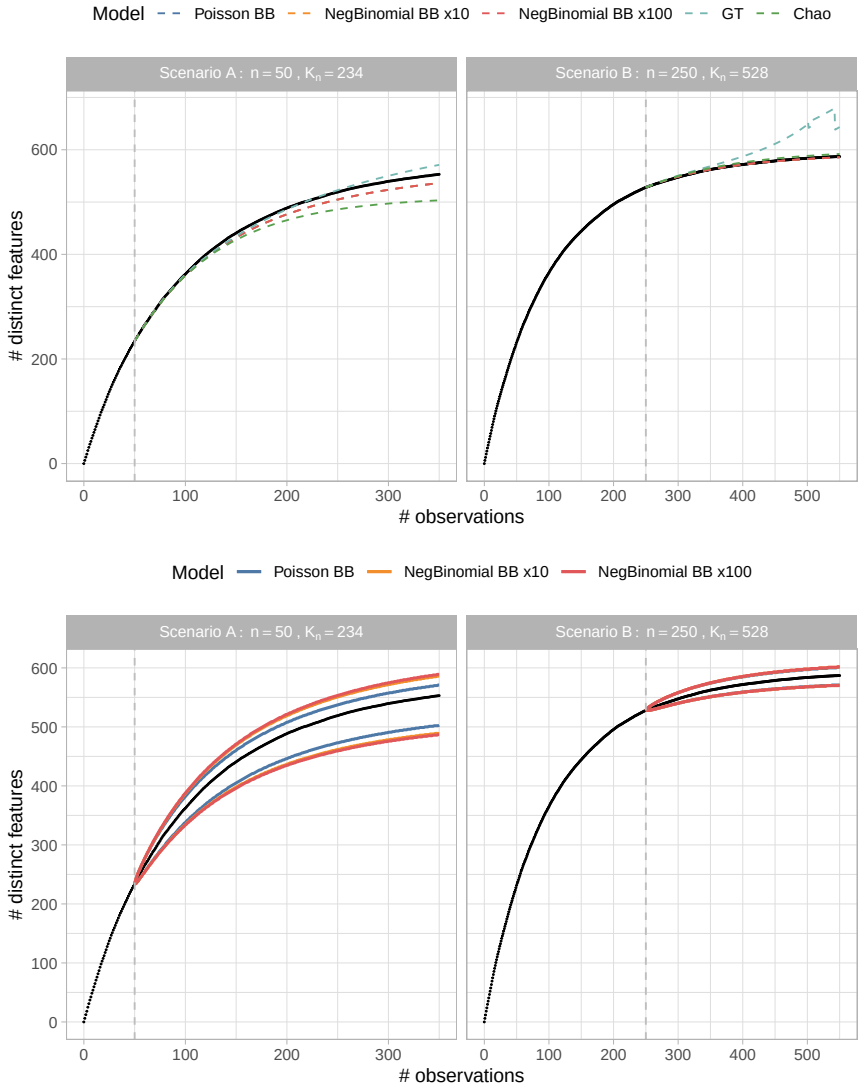



Figure 3.D.3: Top row: the accumulation curve $K_m^{(n)} + k \,|\, \boldsymbol{Z}$ under the mixtures of BBs, the Good–Toulmin estimator and the Chao estimator for $n = 50$ (left) and $n = 250$ (right). Bottom row: the 95% credible intervals of $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, for $n = 50$ (left) and $n = 250$ (right), for the Poisson and the negative binomial mixtures of BBs. The grey vertical lines indicate the training set.

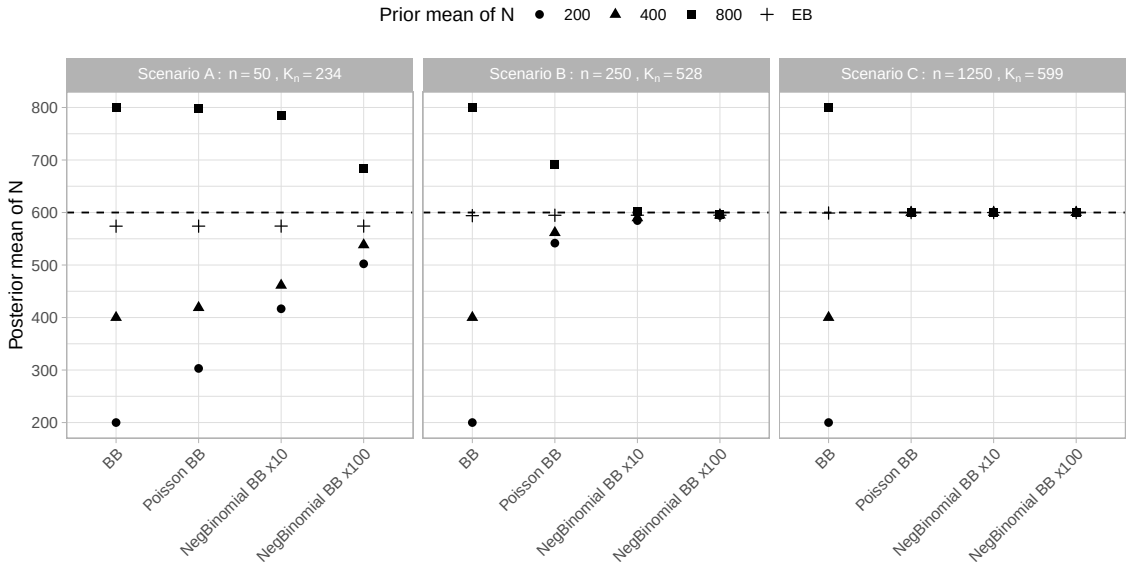



Figure 3.D.4: Posterior mean of the richness $M$, for increasing sample sizes $n \in \{50, 250, 1250\}$, and different values of the parameters of the models. The dashed lines indicate the true richness $H = 600$.

and $\theta$ as discussed in Section 3.5.1. We observe that, as the size $n$ of the training set increases, the richness estimates shrink towards the true value $H = 600$, under proper mixtures of BBs. We also highlight that when prior guesses on $M$ are far from the truth, e.g., $\mathsf{E}(M) \in \{200, 400, 800\}$, negative binomial mixtures of BBs with larger variances produce richness estimates that are closer to the true value $H$ than the ones produced by the Poisson mixture, due to the greater flexibility of the negative binomial prior, which gives less weight to prior guesses. In contrast, under the standard BB model, the prior guess for $M$ deterministically fixes the richness, making it impossible to correct a poor prior guess through posterior inference.

Finally, in Figure 3.D.5, we show the posterior distributions of the richness $M$, reporting (i) the empirical Bayes approach for parameters elicitation and (ii) the prior mean of $M$ equal to 400. First, we comment on the inference produced by the proper mixtures of BBs. In case (i), we observe that such posterior distributions give high probability mass to the true richness $H = 600$; for smaller sample size $n = 50$, we can highlight the higher dispersion of the posterior distribution of $M$ for the negative binomial mixtures. In case (ii), when the prior guess of $M$ is bad, e.g., $\mathsf{E}(M) = 400$, the inference gets worse as the sample size decreases, e.g., $n = 50$. This is clearly expected since the posterior of $M$ gives more weight to the contribution of the prior, which brings an incorrect guess. However, it is interesting to note that this posterior under the negative binomial mixture with larger prior variance for $M$, i.e., $\mathrm{Var}(M) = 100 \times \mu_0$, is significantly moving towards the true richness even for $n = 50$. Second, we discuss the undesired behavior of the standard BB model. In case (i), although the posterior of $M$ concentrates on a value relatively close to the true one $H = 600$, the lack of uncertainty in the posterior makes $H$ not plausible under the standard BB model. As a result, the inference is highly unreliable. The same

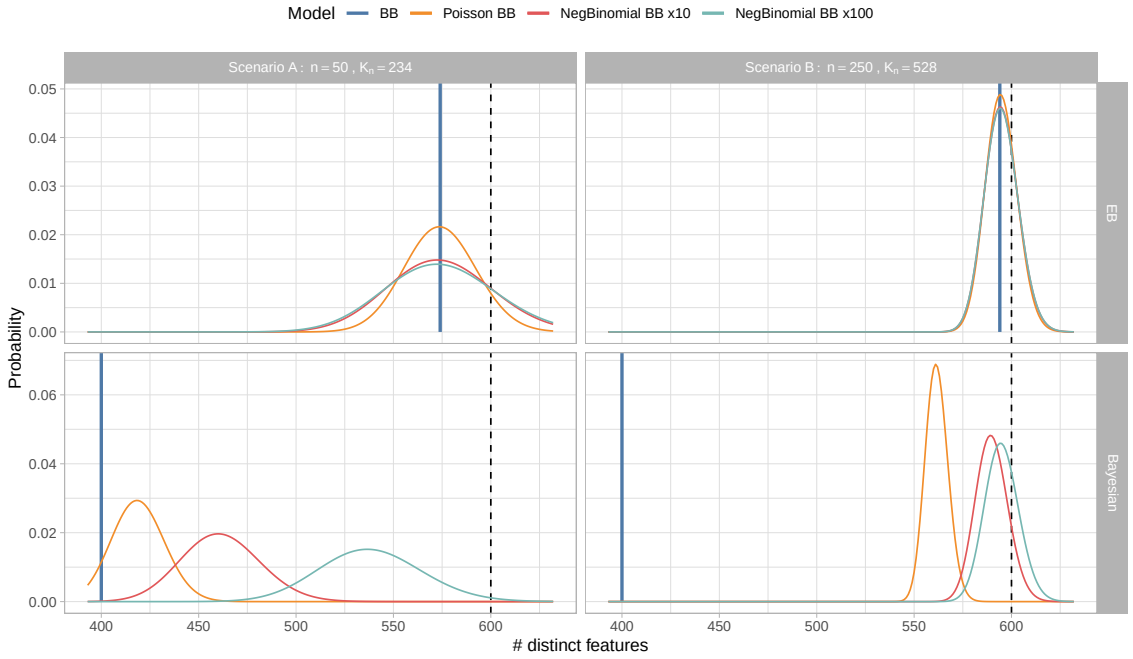



Figure 3.D.5: Posterior distribution of the richness $M$ for increasing sample sizes $n \in \{50, 250\}$. Top row: parameters elicited according to the empirical Bayes approach; bottom row: $\mathsf{E}(M) = 400$, with $\alpha$ and $\theta$ chosen as in Section 3.5.1. The vertical black lines indicate the true richness $H = 600$.

issue emerges in case (ii), when the prior guess of $M$ is bad (equal to 400). Again, note that under the standard BB model, the posterior of $M$ remains concentrated at 400 and cannot learn from data as the sample size increases.

### 3.D.2 Simulation study B

We assume a generating mechanism such that the feature occurrence probabilities $\pi_k$'s are set as $\pi_k = 1/k$, for $k = 1, \ldots, H$, with $H = 10^6$ total features. The choice of a large $H$ is intended to mimic a scenario where the number of features is infinite. We generate a dataset of observations and we consider increasing dimensions of the training set $n \in \{10, 50, 250\}$. In Figure 3.D.6 we report the results of the suggested visual model-checking procedure for the case $n = 50$. Differently from simulation study A, the mixtures of BBs are not properly fitting the data. Instead, we argue that the mixtures of IBPs can be assumed to be correctly specified. A more quantitative assessment of model fit can be obtained by comparing the deviances, with $D(\hat{\boldsymbol{\theta}}) = 5398.7$ for the BB model versus $D(\hat{\boldsymbol{\theta}}) = 5094.5$ for the IBP model. These results are consistent with the conclusions drawn from the visual inspection of Figure 3.D.6. For completeness, Figure 3.D.7 reports the 95% credible intervals of $K_1, \ldots, K_n$ and of $K_{n,r}$, $r \geq 1$, for two examples of gamma mixture of IBPs. We note that the width of the credible intervals increases as the prior variance on $\gamma$ increases.

We address the prediction of the number of unseen features in an additional sample of increasing size, comparing the gamma mixture of IBPs and a variation of the Good–Toulmin estimators (GT) in Chakraborty et al. (2019). We do not consider the estimator in Chao et al. (2014) here, since it is specifically designed for situations where the assumption

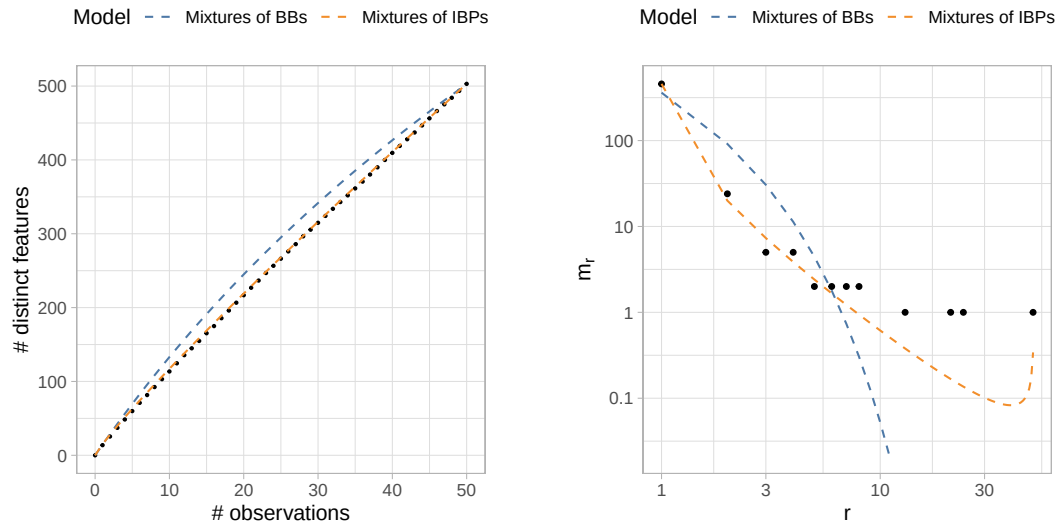



Figure 3.D.6: Case $n = 50$. Left panel: the empirical accumulation curve (black dots) and the rarefaction curve of the models (blue and orange dashed lines). Right panel: the observed values of $K_{n,r}$ (black dots) compared with the expected curve $\mathsf{E}(K_{n,r})$ of the models (blue and orange 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).

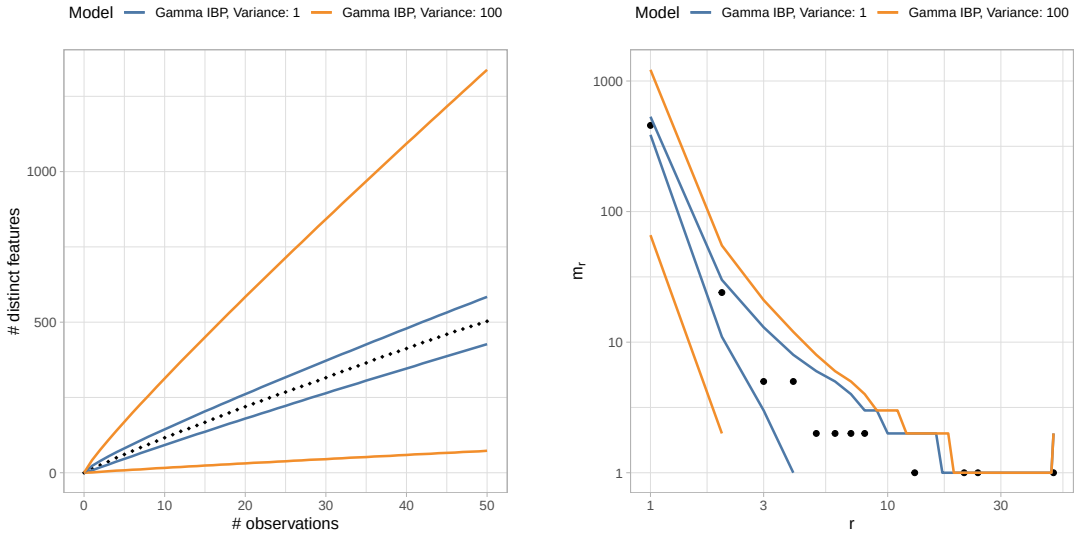



Figure 3.D.7: Case $n = 50$. Left panel: the empirical accumulation curve (black dots) and the credible intervals (delimited by colored lines) of $K_1, \ldots, K_n$ for two examples of gamma mixture of IBPs. Right panel: the observed values of $K_{n,r}$ (black dots) and the credible intervals (delimited by colored lines) of $K_{n,r}$ for the same mixtures of IBPs. The right plot is in log-log scale.

of finite richness is plausible. For the gamma mixture of IBPs, we consider two distinct choices for the prior variance of $\gamma$, i.e., $\text{Var}(\gamma) \in \{1, 100\}$. For the purpose of discussion, we also analyze the standard IBP model. We report the comparison in Figure 3.D.8, for $n \in \{10, 50, 250\}$. Starting with the Good–Toulmin estimates, it is evident how the prediction curve catches the observed curve just for very limited horizons $m$. On the other hand, all the gamma mixtures are able to capture the growth rate of $K_m^{(n)}$ for increasing values of $m$. Remarkably, a larger prior variance of $\gamma$, e.g., $\text{Var}(\gamma) = 100$, results in wider credible intervals for the prediction, allowing for a more conservative uncertainty quantification. Indeed, under the standard IBP model, the observed extrapolation curve is not always contained in the credible bands, leading to a less reliable prediction of the variability of the phenomenon. Remind that the IBP model corresponds to the limiting case $\text{Var}(\gamma) \to 0$; this clearly justifies the need of the mixtures of IBPs.

### 3.D.3 Simulation study C

Here, we report a simulation study where the data are generated from the BB model, with total number of features equal to $H = 500$. Specifically, the feature occurrence probabilities are drawn as $\pi_k \overset{\text{iid}}{\sim} \text{Beta}(1, 100)$, $k = 1, \ldots, H$. We generate a dataset of observations and we consider increasing dimensions of the training set $n \in \{200, 1000, 5000\}$. The sample sizes are larger than the other simulation studies since the growth of the accumulation curve is rather slow.

Figure 3.D.9 shows the visual model-checking for the case $n = 1000$. As expected, the mixtures of IBPs cannot explain the observed data. Conversely, it is reasonable to assume the mixtures of BBs to be correctly specified for this dataset. This claim is further supported by the comparison of deviances, with the BB model yielding $D(\hat{\boldsymbol{\theta}}) = 53208.8$ compared to $D(\hat{\boldsymbol{\theta}}) = 53392.2$ for the IBP model. Based on both visual and quantitative evidence, we focus on the mixtures of BBs for the inference and prediction. As in the previous simulation studies, Figure 3.D.10 reports the 95% credible intervals of $K_1, \ldots, K_n$ and of $K_{n,r}$, $r \geq 1$, for the Poisson and two examples of negative binomial mixture of BBs.

Similarly to simulation study A, we compare the Poisson mixture of BBs, the negative binomial mixture of BBs, the frequentist estimator (Chao) in Chao et al. (2014) and a variation of the Good–Toulmin estimators (GT) in Chakraborty et al. (2019), in terms of prediction of the number of unseen features in an additional sample of increasing size. For the negative binomial mixture of BBs, we analyse two distinct choices for the prior variance of $M$, specifically $\text{Var}(M) = \mu_0 \times c$, with $c \in \{10, 100\}$. We report the comparison in Figure 3.D.11, for $n \in \{200, 1000\}$. For larger sample sizes, e.g., $n = 1000$, all the models seem to produce reliable predictions, at least for the analyzed horizon $m = 1, \ldots, 500$. Significant differences are observed for the smaller sample size $n = 200$, where the pointwise estimates (the expected values of $K_m^{(n)} + k$, for $m = 1, \ldots, 500$) produced by our mixtures of BBs accurately predict the observed curve in the test set, with the credible intervals nicely quantifying the uncertainty around such estimates. Indeed, the observed curve is always contained in the credible bands. On the other hand, the Good–Toulmin estimates show their well-known stability issues, while the estimator in Chao et al. (2014) clearly underestimates the observed curve.

For the analyzed dataset, the model-checking in Figure 3.D.9 suggests the correct spec-

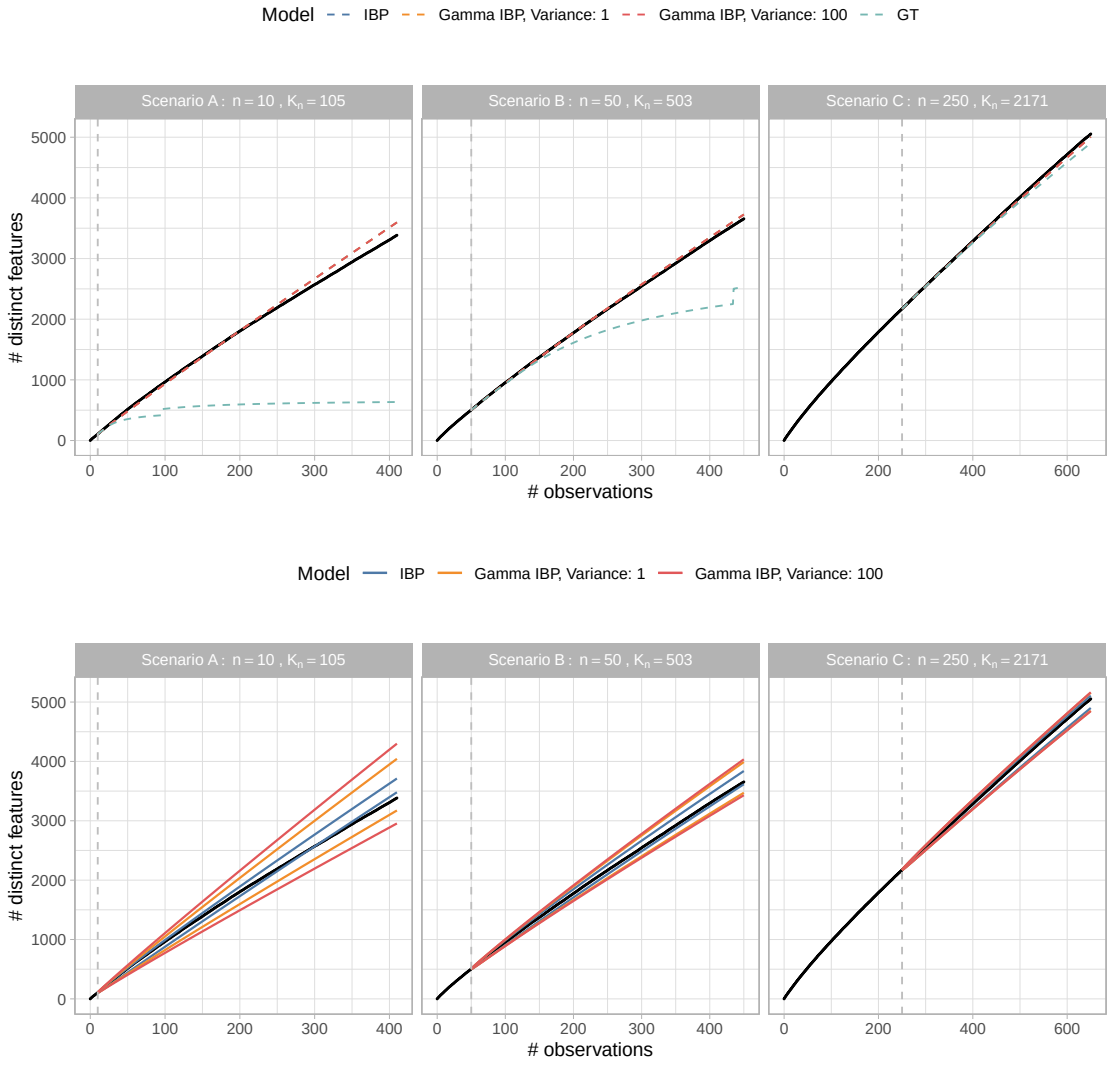



Figure 3.D.8: Top row: the accumulation curve $K_m^{(n)} + k \,|\, \boldsymbol{Z}$ under the gamma mixtures of IBPs and the Good–Toulmin estimator for $n = 10$ (left), $n = 50$ (center) and $n = 250$ (right). Bottom row: the 95% credible intervals of $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, for $n = 10$ (left), $n = 50$ (center) and $n = 250$ (right), for the gamma mixtures of IBPs.

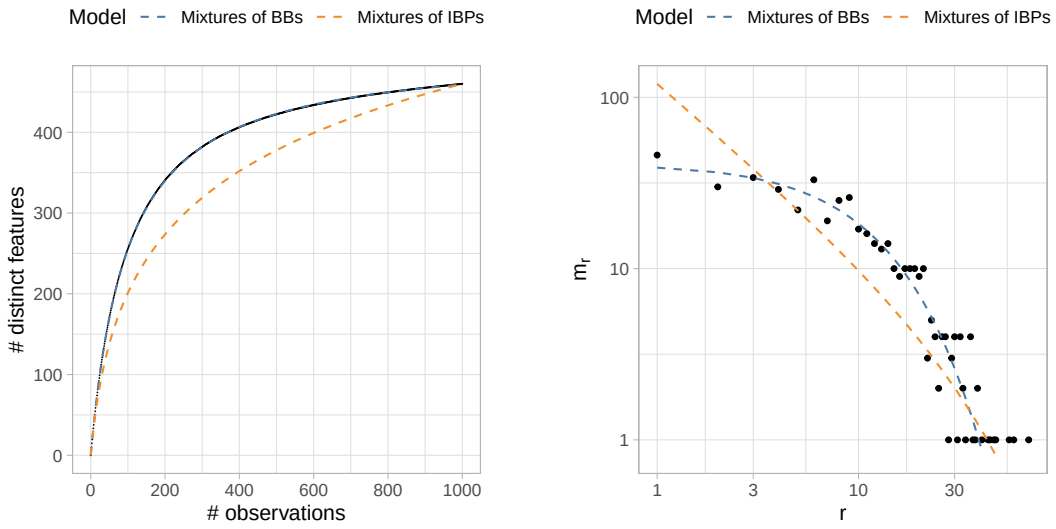



Figure 3.D.9: Case $n = 1000$. Left panel: the empirical accumulation curve (black dots) and the rarefaction curve of the models (blue and orange dashed lines). Right panel: the observed values of $K_{n,r}$ (black dots) compared with the expected curve $\mathsf{E}(K_{n,r})$ of the models (blue and orange 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).

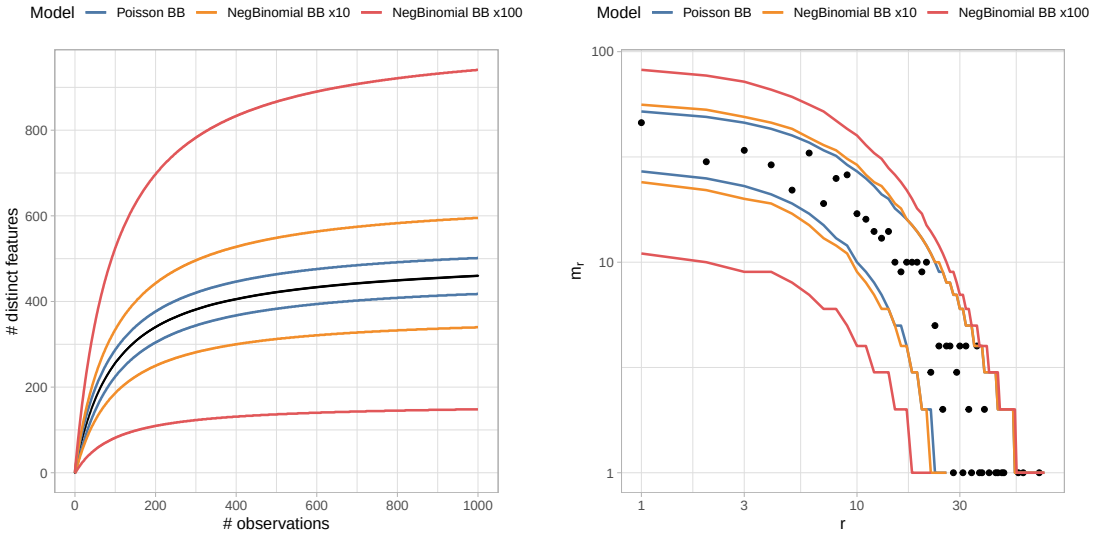



Figure 3.D.10: Case $n = 1000$. Left panel: the empirical accumulation curve (black dots) and the credible intervals (delimited by colored lines) of $K_1, \ldots, K_n$ for the Poisson mixture of BBs and two examples of negative binomial mixture of BBs. Right panel: the observed values of $K_{n,r}$ (black dots) and the credible intervals (delimited by colored lines) of $K_{n,r}$ for the same mixtures of BBs. The right plot is in log-log scale.

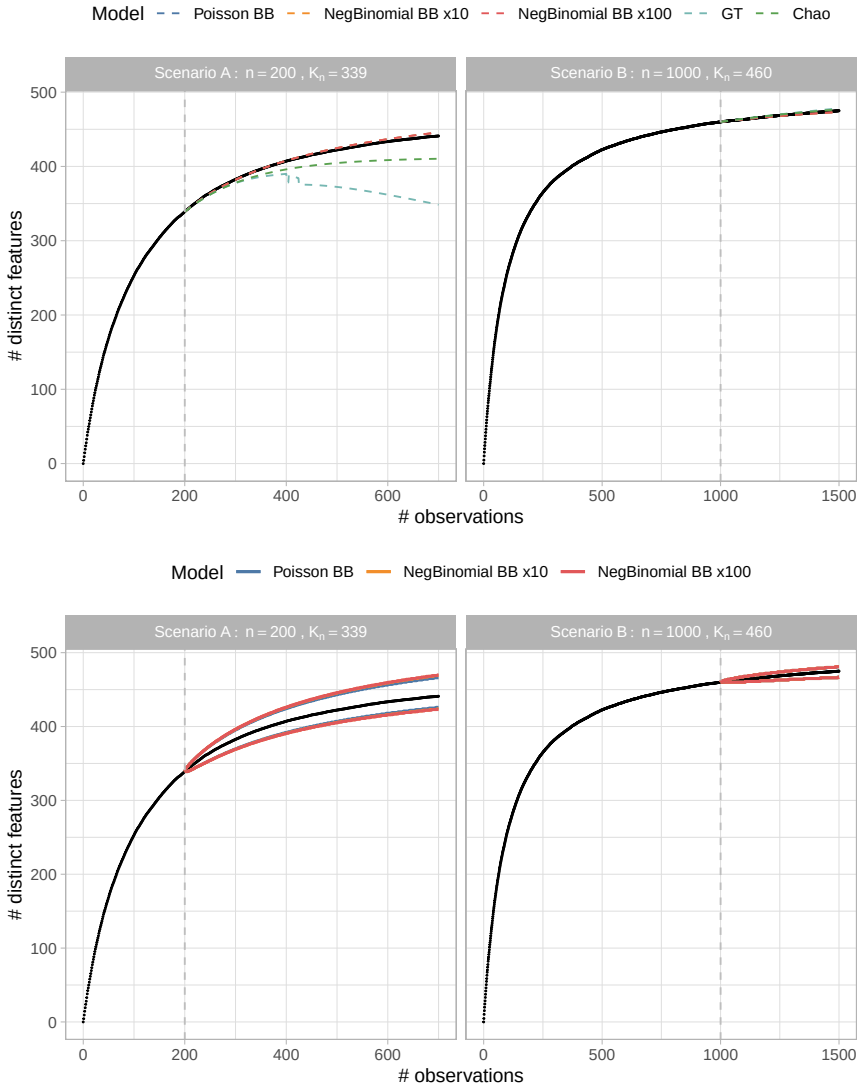



Figure 3.D.11: Top row: the accumulation curve $K_m^{(n)} + k \,|\, \boldsymbol{Z}$ under the mixtures of BBs, the Good–Toulmin estimator and the Chao estimator for $n = 200$ (left) and $n = 1000$ (right). Bottom row: the 95% credible intervals of $K_m^{(n)} + k \,|\, \boldsymbol{Z}$, for $n = 200$ (left) and $n = 1000$ (right), for the Poisson and the negative binomial mixtures of BBs. The grey vertical lines indicate the training set.

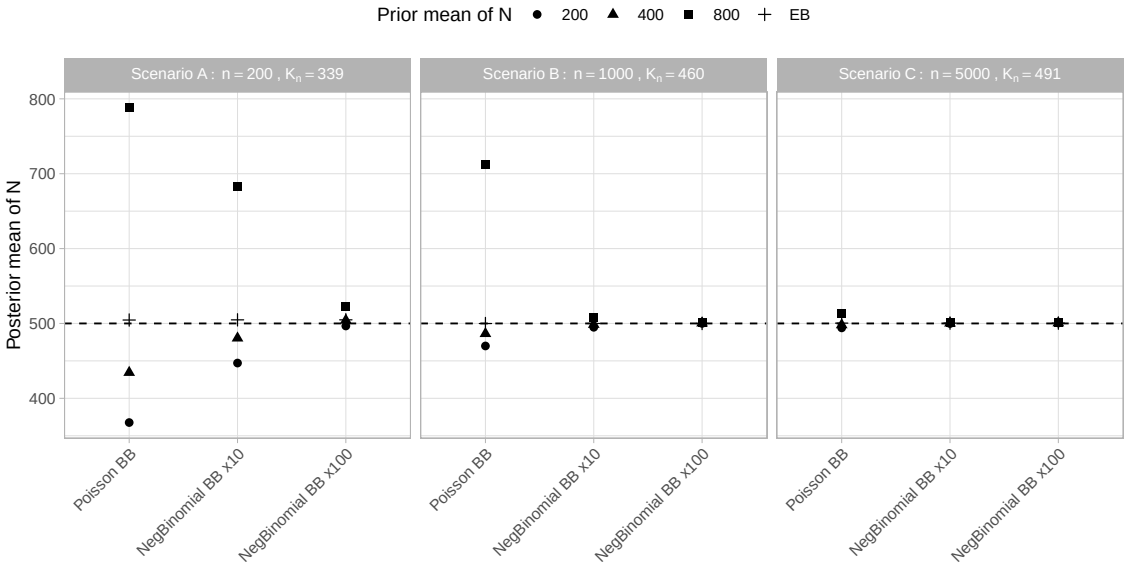



Figure 3.D.12: Posterior mean of the richness $M$, for increasing sample sizes $n \in \{200, 1000, 5000\}$, and different values of the parameters of the models. The dashed lines indicate the true richness $H = 500$.

ification of the mixtures of BBs. Consequently, it is reasonable to assume the finiteness of the richness and address its estimation. Figure 3.D.12 shows an expected, still remarkable, illustration on the richness estimation via the Poisson mixture and two negative binomial mixtures with $\text{Var}(M) = \mu_0 \times c$, with $c \in \{10, 100\}$. Specifically, for increasing sample sizes $n \in \{200, 1000, 5000\}$, it reports the posterior mean of the species richness $M$, for different choices of the parameters of the prior distribution on $M$. In particular, for each model, we estimate the parameters following the usual empirical Bayes procedure (referred to as EB in the figure); moreover, we fix different choices for the prior mean of $M$, i.e., $\mathsf{E}(M) \in \{200, 400, 800\}$, and we estimate the parameters $\alpha$ and $\theta$ as discussed in Section 3.5.1. We observe that, as the size $n$ of the training set increases, the richness estimates shrink towards the true value $H = 500$. We also highlight that when prior guesses on $M$ are substantially wrong, e.g., $\mathsf{E}(M) \in \{200, 400, 800\}$, negative binomial mixtures of BBs with larger variances produce richness estimates that are closer to the true value $H$ than the ones produced by the Poisson mixture, due to the greater flexibility of the negative binomial prior, which is able to give less weight to wrong prior guesses. In contrast, under the standard BB model, the prior guess for $M$ deterministically fixes the richness, and thus posterior inference on $M$ is impossible.

Finally, in Figure 3.D.13, we show the posterior distributions of the richness $M$, reporting (i) the empirical Bayes approach for parameters elicitation and (ii) the prior mean of $M$ equal to 400. Commenting on the inference produced in case (i), we observe that such posterior distributions give high probability mass to the true richness $H = 500$; for smaller sample size $n = 200$, we can remark the higher dispersion of the posterior distribution of $M$ for the negative binomial mixtures. In case (ii), when the prior guess of $M$ is wrong, e.g., $\mathsf{E}(M) = 400$, the inference gets worse as the sample size decreases, e.g., $n = 200$. This

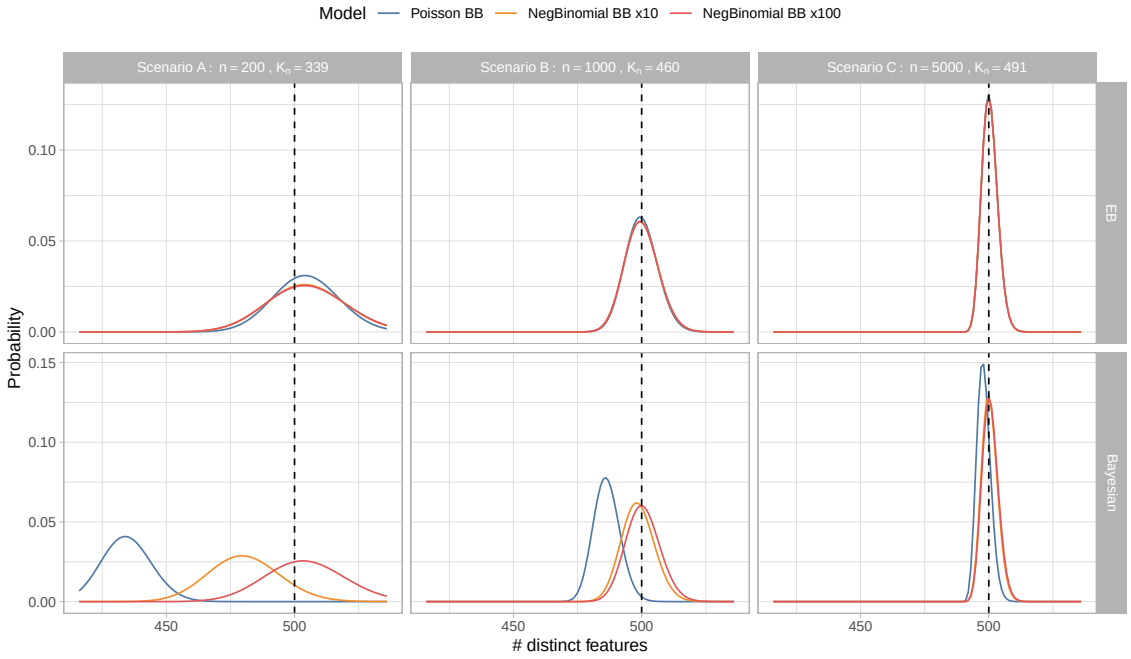



Figure 3.D.13: Posterior distribution of the richness $M$ for increasing sample sizes $n \in \{200, 1000, 5000\}$. Top row: parameters elicited according to the empirical Bayes approach; bottom row: $\mathsf{E}(M) = 400$, with $\alpha$ and $\theta$ chosen as in Section 3.5.1. The vertical black lines indicate the true richness $H = 500$.

is clearly expected since the posterior of $M$ gives more weight to the contribution of the prior, which brings an incorrect guess. However, It is interesting to note how the negative binomial mixtures are perfectly catching the true richness even for $n = 200$.

## 3.E Ecological applications: additional details and fully Bayesian approach

### 3.E.1 Additional analyses

In the present section we provide additional details and analyses for the ecological applications discussed in Section 3.6 of the main text. For each ecological dataset, we present: (i) the observed taxon accumulation curve, (ii) the 95% credible intervals of $K_1, \ldots, K_n$ and of $K_{n,r}$, $r \geq 1$, based on some examples from the class of mixtures favored by the model selection procedures, (iii) the results of a data-holdout experiment, in which each model is trained on half of the observed data and evaluated in terms of its predictive performance on the held-out portion.

**Vascular plants in Danish forest**

Figure 3.E.1 shows the observed taxon accumulation for the vascular plants in Mazziotta et al. (2016a); the plot clearly indicates that the asymptote of the curve has not yet been reached, suggesting that the species richness will exceed the observed number of species in the sample.

As detailed in the main text, the model-checking procedures favor mixtures of IBPs for

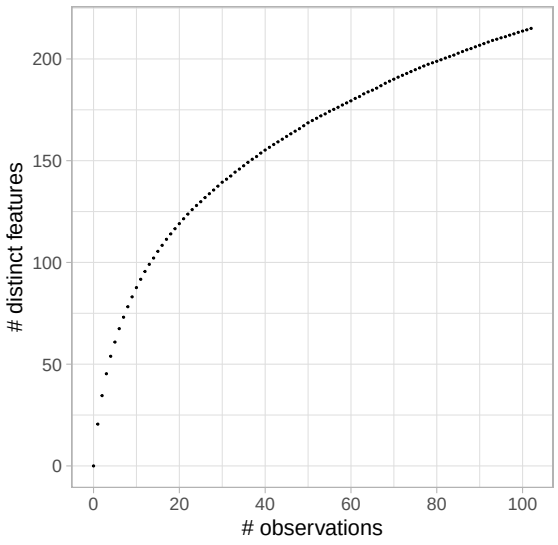



Figure 3.E.1: Taxon accumulation curve for the vascular plants in Mazziotta et al. (2016a).

this dataset. Accordingly, Figure 3.E.2 reports the 95% credible intervals of $K_1, \ldots, K_n$ and of $K_{n,r}$, $r \geq 1$, for two examples of gamma mixture of IBPs.

For the data-holdout experiment, we randomly partition the observed data, selecting half as the training set, whose size equals $n/2 = 51$, with $n = 102$. Then, we fit the Poisson and two examples of negative binomial mixture of BBs, along with two examples of gamma mixture of IBPs, on the training data. In Figure 3.E.3, we compare the predictions of the extrapolation curve from all fitted models against the observed extrapolation curve in the held-out test set. The plots clearly show that the mixtures of IBPs effectively predict the number of unseen species in the training data which are displayed in the test data. It is also apparent that the mixtures of BBs struggle to capture the true growth of the observed extrapolation curve.

### Trees in Barro Colorado Island

Figure 3.E.4 illustrates the observed taxon accumulation for the Barro Colorado Island dataset, freely available in the VEGAN package in R. The curve does not exhibit any clear asymptote, indicating that the species richness likely exceeds the observed number of species in the sample.

As discussed in the main text, the model-checking procedures favor mixtures of BBs for this dataset. Accordingly, Figure 3.E.5 reports the 95% credible intervals of $K_1, \ldots, K_n$ and of $K_{n,r}$, $r \geq 1$, for the Poisson and two examples of negative binomial mixture of BBs.

For the data-holdout experiment, we randomly split the dataset, using half of the data as the training set, whose size equals $n/2 = 50$, with $n = 100$. Then, we fit the Poisson and two examples of negative binomial mixture of BBs, along with two examples of gamma mixture of IBPs, on the training data. In Figure 3.E.6, we compare the predictions of the extrapolation curve from all fitted models against the observed extrapolation curve in the held-out test set. The plots clearly show that the mixtures of BBs successfully predict the number of previously unseen species observed in the test data, while the mixtures of IBPs

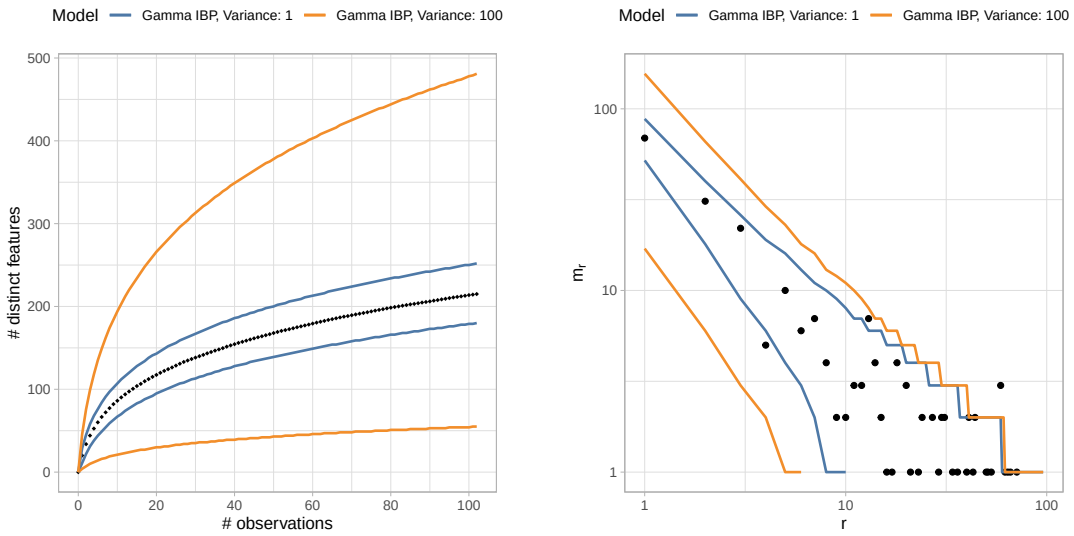



Figure 3.E.2: Left panel: the empirical accumulation curve (black dots) and the credible intervals (delimited by colored lines) of $K_1, \ldots, K_n$ for two examples of gamma mixture of IBPs. Right panel: the observed values of $K_{n,r}$ (black dots) and the credible intervals (delimited by colored lines) of $K_{n,r}$ for the same mixtures of IBPs. The right plot is in log-log scale.

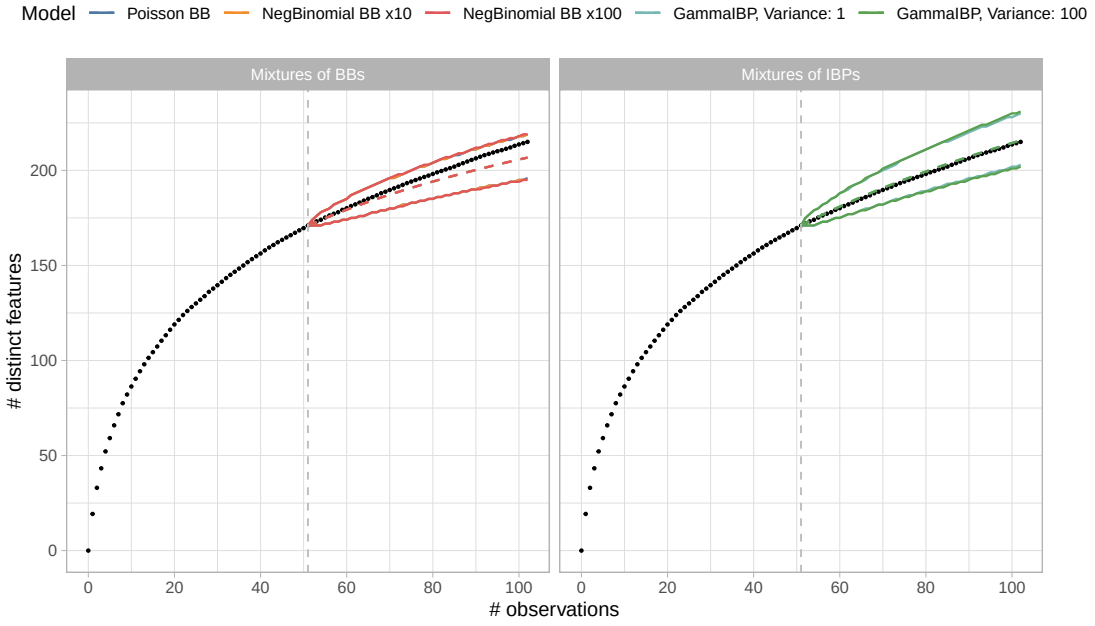



Figure 3.E.3: Data-holdout analysis with training set size equal to $n/2$, where $n = 102$: expected values (dashed lines) and 95% credible intervals (delimited by solid lines) of $K_m^{(n/2)} + k \,|\, \boldsymbol{Z}^{(n/2)}$, for the mixtures of BBs (left) and the mixtures of IBPs (right). The grey vertical lines indicate the training set.

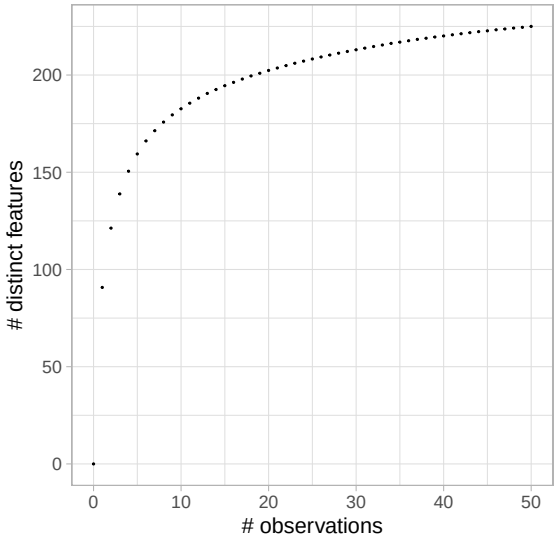



Figure 3.E.4: Taxon accumulation curve for the trees in the Barro Colorado Island data.

fail to capture this extrapolation behavior.

### 3.E.2 Fully Bayesian approach: details

In Section 3.5.1 we describe a recommended procedure for eliciting the parameters of the models via an empirical Bayes approach. We consider such a procedure for all the discussed simulation studies and real data examples. However, one might prefer to adopt a fully Bayesian approach instead of the proposed empirical Bayes one. Specifically, the interest might be in assuming prior distributions for the parameters $\alpha$ and $\theta$, for all of the mixtures. Within the mixtures of BBs, we argue that the negative binomial mixture has to be regarded as the fully Bayesian version of the Poisson mixture, since the former is obtained by placing a gamma prior on the $\lambda$ parameter of the latter.

In this section, we first discuss the prior elicitation for the parameters $\alpha$ and $\theta$; then, we illustrate the comparison with the empirical Bayes approach, in terms of inference and prediction obtained in the two real data scenarios. As far as prior elicitation is concerned, since the constraint $\theta > -\alpha$ for all of the models, it is convenient to perform a change of variable, introducing $s = \theta + \alpha$ and removing $\theta$, so that $s > 0$. Then, prior distributions on $\alpha$ and $s$ are specified. Specifically, for the mixtures of IBPs, characterized by $\alpha \in [0, 1)$, we consider

$$\alpha \sim \text{Beta}\left(a_\alpha, b_\alpha\right), \; s \sim \text{Gamma}\left(a_s, b_s\right),$$

with $a_\alpha, b_\alpha > 0$ and $a_s, b_s > 0$. Denoting with $\hat{\alpha}$ and $\hat{\theta}$ the empirical Bayes estimates obtained as in Section 3.5.1, we select the hyperparameters of the priors by imposing $\mathsf{E}(\alpha) = \hat{\alpha}$ and $\mathsf{E}(s) = \hat{s}$, where $\hat{s} = \hat{\alpha} + \hat{\theta}$. It follows that $a_\alpha/(a_\alpha + b_\alpha) = \hat{\alpha}$ and $a_s/b_s = \hat{s}$. Two additional equations, one involving $a_\alpha, b_\alpha$ and one involving $a_s, b_s$, are necessary to fix the four hyperparameters. For example, such equations may control the prior variances of $\alpha$ and $s$, so that the practitioner can select the degree of regularization induced by the priors. Draws from the posterior distribution of $(\alpha, s)$ can be easily obtained via any Metropolis-Hastings algorithm. Specifically, we apply the change of variables $\alpha' = \log(\alpha/(1-\alpha))$,

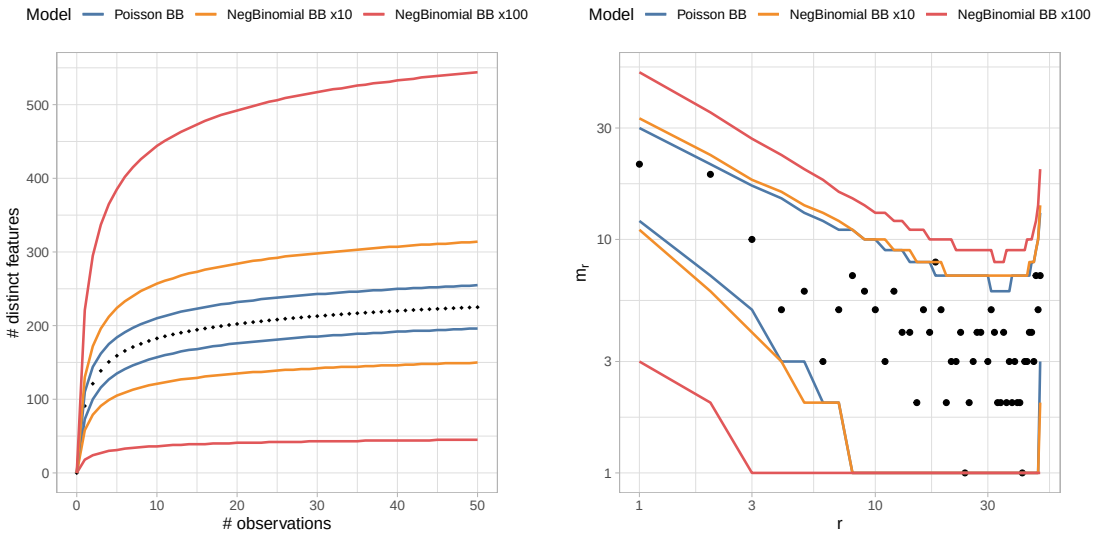



Figure 3.E.5: Left panel: the empirical accumulation curve (black dots) and the credible intervals (delimited by colored lines) of $K_1, \ldots, K_n$ for the Poisson mixture of BBs and two examples of negative binomial mixture of BBs. Right panel: the observed values of $K_{n,r}$ (black dots) and the credible intervals (delimited by colored lines) of $K_{n,r}$ for the same mixtures of BBs. The right plot is in log-log scale.

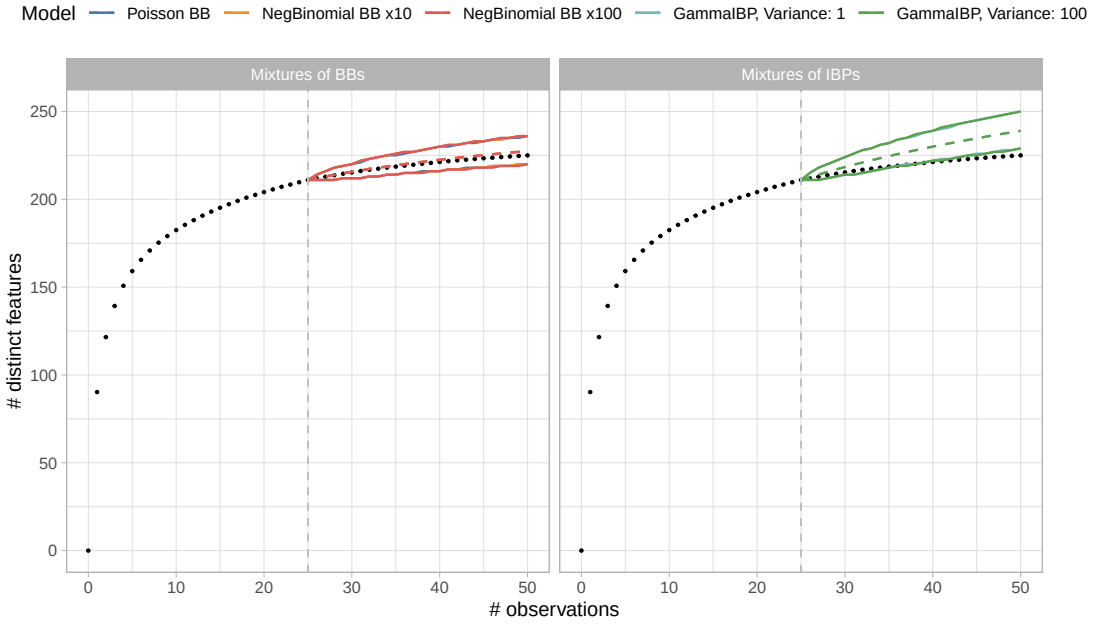



Figure 3.E.6: Data-holdout analysis with training set size equal to $n/2$, where $n = 100$: expected values (dashed lines) and 95% credible intervals (delimited by solid lines) of $K_m^{(n/2)} + k \,|\, \boldsymbol{Z}^{(n/2)}$, for the mixtures of BBs (left) and the mixtures of IBPs (right). The grey vertical lines indicate the training set.

$s' = \log(s)$, so that $(\alpha', s') \in \mathbb{R}^2$, and we update $(\alpha', s')$ via a pre-conditioned MALA (Metropolis Adjusted Langevin Algorithm), since the gradient of the log-full-conditional density of $(\alpha', s')$ is available in closed form.

For the mixtures of BBs, having $\alpha < 0$, we consider

$$-\alpha \sim \text{Gamma}\,(a_\alpha, b_\alpha)\,,\ s \sim \text{Gamma}\,(a_s, b_s)\,,$$

with $a_\alpha, b_\alpha > 0$ and $a_s, b_s > 0$. We choose the hyperparameters of the priors such that $\mathsf{E}(\alpha) = \hat{\alpha}$ and $\mathsf{E}(s) = \hat{s}$. This imposes $a_\alpha / b_\alpha = -\hat{\alpha}$ and $a_s / b_s = \hat{s}$. Similarly to the mixtures of IBPs, one may set the degree of regularization induced by the priors by choosing the prior variances of $\alpha$ and $s$. Sampling from the posterior distribution of $(\alpha, s)$ is straightforward through any Metropolis-Hastings algorithm. We resort again to a pre-conditioned MALA for the transformed parameters $(\alpha', s')$, where $\alpha' = \log(-\alpha)$, $s' = \log(s)$, and we stress that the gradient of the log-full-conditional density of $(\alpha', s')$ is available in closed form.

Here, we report the comparison among the fully Bayesian approach and the empirical Bayes approach described in Section 3.5.1, in terms of inference and/or prediction, concerning the two real data scenarios.

**Vascular plants in Danish forest**

Analyzing the vascular plants data of Mazziotta et al. (2016a) in Section 3.6.1, we claimed the correct specification of the mixtures of IBPs. We consequently showed the prediction obtained via the gamma mixture of IBPs, using the empirical Bayes approach to select the parameters. In particular, the empirical Bayes estimates for $\alpha$ and $\theta$ result as follows: $\hat{\alpha} = 0.17$, $\hat{\theta} = 1.7$. For the fully Bayesian approach, we then select the hyperparameters by imposing: $\mathsf{E}(\alpha) = \hat{\alpha} = 0.17$, $\text{Var}(\alpha) = 10^{-2}$, and $\mathsf{E}(s) = \hat{s} = 1.87$, $\text{Var}(s) = 187$, so that a moderate amount of regularization is introduced. We run the MCMC algorithm for $5 \cdot 10^4$ iterations, discarding the first $5 \cdot 10^3$ and keeping one every two iterations. In Figure 3.E.7, we report the comparison between the fully Bayesian approach and the empirical Bayes approach described in Section 3.5.1, in terms of the extrapolation curve. We observe that, for both the choices of the prior variance of $\gamma$, the expected number of unseen species which will be collected in additional samples of increasing sizes $m$ is almost identical for the two approaches. This is desirable since it confirms that the empirical Bayes approach leads to equivalent prediction than the fully Bayesian procedure. On the other hand, the credible intervals produced by the fully Bayesian approach are larger than the ones estimated with the empirical Bayes method, as naturally expected.

**Trees in Barro Colorado Island**

Differently than for the vascular plants of Mazziotta et al. (2016a), the tree data investigated in Section 3.6.2 are suitably modelled via the mixtures of BBs (refer to Figure 3.4). In the main text, we report the inference and prediction produced by the Poisson and negative binomial mixtures of BBs, with parameters selected via the empirical Bayes approach discussed in Section 3.5.1. In particular, we get $\hat{\alpha} = -0.29$ and $\hat{\theta} = 0.95$ as empirical Bayes estimates for the parameters. In the fully Bayesian approach, we select the hyperparameters so that $\mathsf{E}(\alpha) = \hat{\alpha} = -0.29$, $\text{Var}(\alpha) \approx 300$, and $\mathsf{E}(s) = \hat{s} = 0.66$, $\text{Var}(s) \approx 650$. We run the MCMC algorithm for $5 \cdot 10^4$ iterations, discarding the first $5 \cdot 10^3$ and keeping

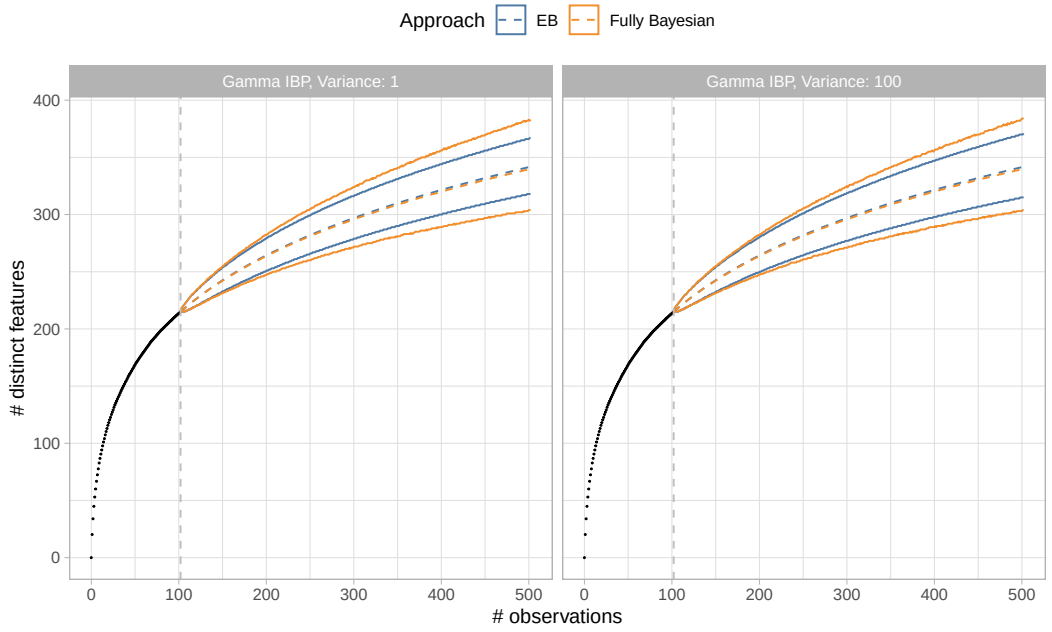



Figure 3.E.7: Comparison between the empirical Bayes approach (EB) and the fully Bayesian approach in terms of expected values and the 95% credible intervals for $K_m^{(n)} + k$, with $k = 215$, for the gamma mixtures of IBPs. The extrapolation horizon is $m = 1, \ldots, 400$. The left panel shows the case $\text{Var}(\gamma) = 1$, the right panel considers $\text{Var}(\gamma) = 100$.

one every two iterations. In Figure 3.E.8, we report the comparison between the fully Bayesian approach and the empirical Bayes approach, in terms of the extrapolation curve. The expected values for the extrapolation curve are similar between the two approaches, with larger credible intervals in the fully Bayesian case. This behavior is expected and desired. Moreover, Figure 3.E.9 shows the posterior distribution of $M$, the species richness, under the different approaches. It is evident that the fully Bayesian approach leads to much more dispersed posterior distributions for $M$. In particular, we remind that the empirical Bayes procedure estimates an expected species richness of 296.17, with credible interval $[278, 316]$, for both the choices of the prior variance of $M$. The fully Bayesian approach produces an expected species richness of 306.24 with credible interval $[262, 375]$, for the case $\text{Var}(M) = 10 \times \mu_0$, while it leads to an expected species richness of 316.44 with credible interval $[261, 425]$, for the case $\text{Var}(M) = 100 \times \mu_0$.

## 3.F Discussion on computational complexity

In this section, we present a brief discussion on some computational aspects related to the proposed methodology. In general, the computational procedure of our methodology is highly efficient, due to the availability of closed form expressions for the EFPFs. More specifically, the computational effort depends on the chosen parameter elicitation strategy, namely: (i) the empirical Bayes approach described in Section 3.5.1; (ii) the fully Bayesian approach presented in the real data analysis of Section 3.E.2.

Concerning (i), i.e., the empirical Bayes approach (Section 3.5.1), computation is limited

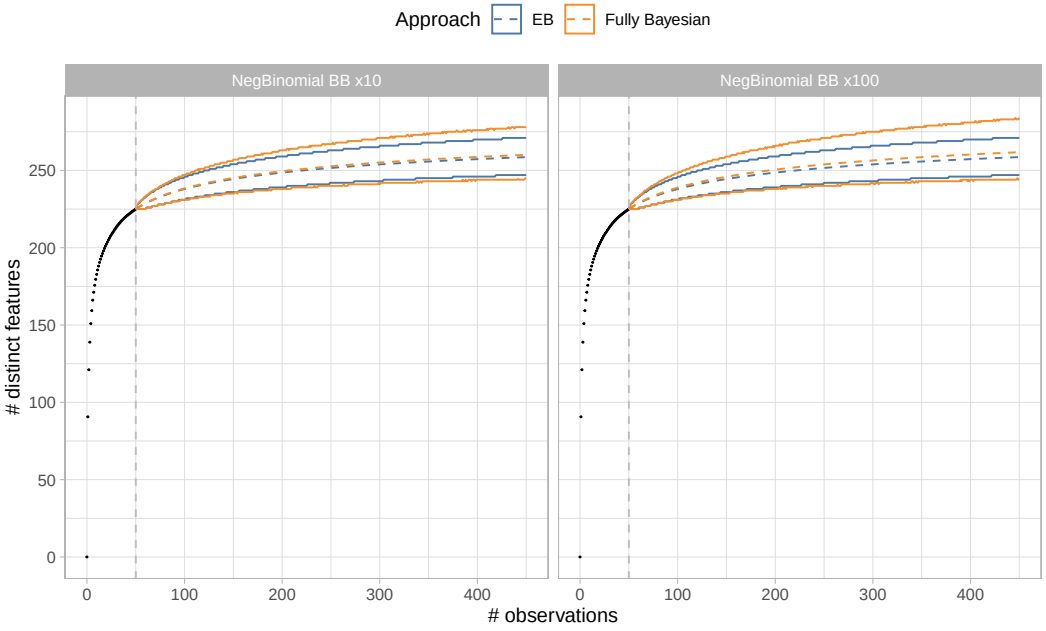



Figure 3.E.8: Comparison between the empirical Bayes approach (EB) and the fully Bayesian approach in terms of expected values and the 95% credible intervals for $K_m^{(n)} + k$, with $k = 225$, for the negative binomial mixtures of BBs. The extrapolation horizon is $m = 1, \ldots, 400$. The left panel shows the case $\text{Var}(M) = 10 \times \mu_0$, the right panel considers $\text{Var}(M) = 100 \times \mu_0$.

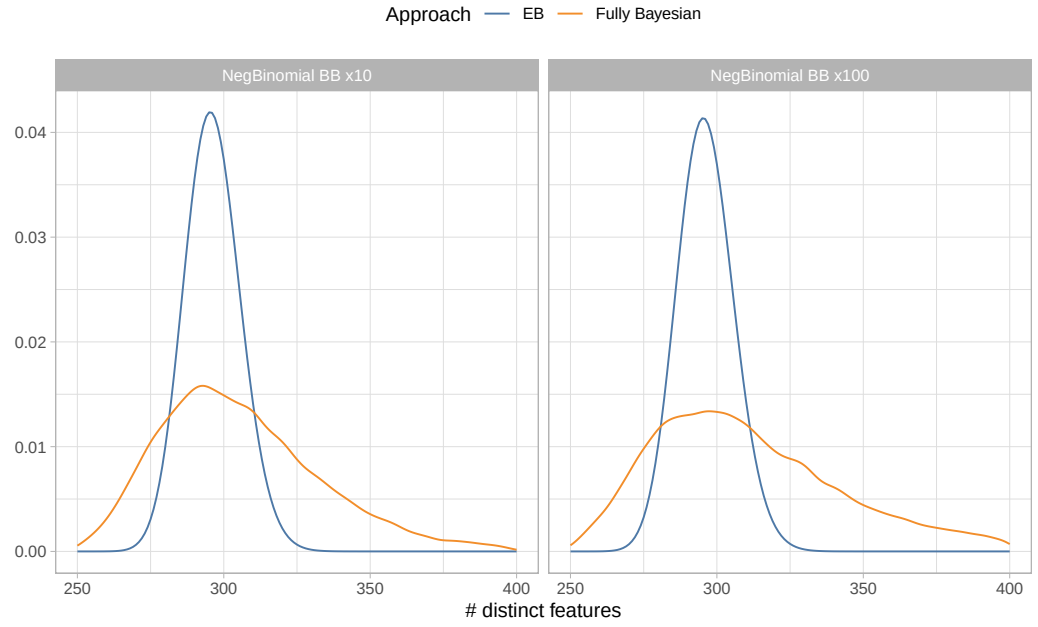



Figure 3.E.9: Comparison between the empirical Bayes approach (EB) and the fully Bayesian approach in terms of the posterior distribution of the species richness $M$, for the negative binomial mixtures of BBs. The left panel shows the case $\text{Var}(M) = 10 \times \mu_0$, the right panel considers $\text{Var}(M) = 100 \times \mu_0$.

to optimizing the parameters of the mixtures of BBs or the mixtures of IBPs. Specifically, for the mixtures of BBs, the parameters are estimated by maximizing the EFPF of the BB model as given in Equations (3.1)–(3.3). The computational complexity of each evaluation of the EFPF of the BB model, as a function of the sample size $n$ and the number of observed features $k$, scales as $k$, since the complexity of the term in (3.3) is independent of both $n$ and $k$. Similarly, for the mixtures of IBPs, parameter estimation involves maximizing the EFPF of the IBP model from Equations (3.1)–(3.2). Here, each evaluation scales as $n + k$, where the term in Equation (3.2) contributes to the dependence on $n$. For both optimization problems, we use the `nlminb` function from `R`'s `stats` package. This results in extremely fast computations across all settings.

Concerning (ii), i.e., the fully Bayesian approach (Section 3.E.2), MCMC algorithms are employed for the updating of the parameters $\alpha$ and $\theta$. In particular, for the mixtures of BBs, we apply a pre-conditioned MALA to the transformed parameters $(\alpha', s')$, with $\alpha' = \log(-\alpha)$ and $s' = \log(s)$. The gradient of the log-full-conditional density of $(\alpha', s')$ is available in closed form. Similarly, for the gamma mixture of IBPs, we apply the change of variables $\alpha' = \log(\alpha/(1-\alpha))$ and $s' = \log(s)$, again resorting to a pre-conditioned MALA, where the gradient of the log-full-conditional density of $(\alpha', s')$ is available in closed form. Thus, in all cases, the MCMC algorithms involve only two transformed parameters and benefits from analytic gradients, ensuring both speed and stability.

To quantify the efficiency of the MCMC implementations, we provide below the computational details which concern the real data examples of Section 3.E.2. In particular, for the gamma mixture of IBPs on the *Vascular plants in Danish forest*, under the choice $\mathrm{Var}(\gamma) = 100$, we run $5 \cdot 10^4$ iterations, discarding the first $5 \cdot 10^3$ and keeping one every two iterations. This yields effective sample sizes of 8071.03 for $\alpha$ and 11396.61 for $\theta$. The total runtime is 36.71 seconds on a Windows 11 system with an Intel Core i7-1165G7 CPU @ 2.80 GHz and 8 GB RAM. Finally, for the negative binomial mixture of BBs on the *Trees in Barro Colorado Island*, using $\mathrm{Var}(M) = 10 \times \mu_0$ and the same MCMC settings, we obtain effective sample sizes of 11456.89 for $\alpha$ and 11453.94 for $\theta$. The total runtime is 30.23 seconds using the same machine.

# 4 Bayesian calculus and predictive characterizations of extended feature allocation models

This chapter introduces a unified Bayesian framework for studying extended feature allocation models. Here, "extended" refers to models that incorporate feature interactions, such as repulsiveness or attractiveness, in contrast to traditional feature allocation approaches that assume independent and identically distributed (i.i.d.) features. Moreover, the proposed class of models also allows for arbitrary dependencies across feature weights. Our extended framework offers several key advantages. First, it provides substantially greater modeling flexibility despite only marginally increasing the analytical complexity compared to available models. Second, incorporating repulsiveness in feature priors has important practical applications, mainly when features are latent rather than directly observed, such as in image segmentation and latent feature modeling. The growing literature on repulsive mixture models demonstrates that repulsive priors typically lead to more interpretable posterior inference (see, e.g., Petralia et al., 2012; Xie and Xu, 2019; Cremaschi et al., 2024; Beraha et al., 2025, and the references therein). Third, our framework enables novel applications beyond traditional domains, as we demonstrate through applications in spatial statistics.

We derive explicit closed-form expressions for the marginal, posterior, and predictive distribution under extended feature allocation models, which hold without making assumptions on the prior distribution, therefore generalizing the treatment in James (2017); Broderick et al. (2018) and Chapter 3. Based on the expression of the predictive distribution, we establish two sufficientness postulates for the extended feature models, which characterize priors that lead to predictions for the unseen features depending either only on the sample size or on both the sample size and the number of distinct features. In particular, we show that the former case characterizes the class of CRMs, while the latter pertains to a much broader class of random measures built from mixed Poisson and mixed binomial point processes. See Section 4.1 for the definition of these processes. We specialize the general Bayesian analysis for some notable classes of priors, including Poisson, mixed Poisson, and mixed binomial processes. Our treatment and proofs heavily rely on point process theory and Palm calculus (Kallenberg, 2017; Baccelli et al., 2020), providing new tools for studying feature allocation models. Within Bayesian nonparametrics, Palm calculus plays a central role in Poisson partition calculus, originally pioneered by Lo (1984) and subsequently extended by James (2002, 2005, 2006, 2017). As a byproduct of our contribution, we also obtain a novel characterization of the Poisson process.

Finally, we discuss the practical application of our framework in spatial statistics. We demonstrate that an extended feature model based on a determinantal point process (Hough et al., 2006; Lavancier et al., 2015) can be used to estimate the size of a forest from

partial observations of tree locations, as well as to locate the unseen trees.

The chapter is organized as follows. In Section 4.1, we define the framework of extended feature models and review key concepts from point process theory necessary to our treatment. In Section 4.2, we develop a general Bayesian theory for extended feature allocation models, without any assumption on the prior distribution. In Section 4.3, we provide the two sufficientness postulates, and we discuss strategies to enrich the models' predictive distributions. Section 4.4 applies the general results from Section 4.2 to specific classes of priors, with emphasis on a determinantal point process-based model. Section 4.5 demonstrates the key advantage of this latter model in both simulated and real-world spatial statistics scenarios. The chapter ends with a discussion. Proofs, additional results, and further details on the illustrations are collected in the Appendix.

## 4.1 Point process formulation of extended feature allocation models

### 4.1.1 Background on point processes

Point processes play a central role in the general definition of extended feature allocation models. Thus, here we remind the main notions of point process theory needed in the following sections, among which Palm calculus. A rigorous treatment can be found in Daley and Vere-Jones (2008) or Baccelli et al. (2020). See also Kallenberg (1984) for some further intuition.

First, let us recall the formal definition of a point process. Indicate by $\mathbb{X}$ a Polish space equipped with the corresponding Borel $\sigma$-algebra $\mathscr{X}$. We say that a measure $\nu$ on $(\mathbb{X}, \mathscr{X})$ is *locally finite* if $\nu(B) < \infty$ for any relative compact set $B \in \mathscr{X}$. Denote by $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ the space of locally finite counting measures on $(\mathbb{X}, \mathscr{X})$, equipped with the corresponding Borel $\sigma$-algebra $\mathcal{M}_{\mathbb{X}}$. A point process $\Phi$ on the space $\mathbb{X}$ is a measurable map from an underlying probability space $(\Omega, \mathscr{A}, \mathsf{P})$ taking values in $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ so that $\Phi(B)$ is a nonnegative integer almost surely (a.s.), for any relatively compact Borel set $B$. Let $\mathbf{P}_\Phi := \mathsf{P} \circ \Phi^{-1}$ denote the probability distribution of $\Phi$, which is uniquely characterized by the Laplace functional $\mathcal{L}_\Phi(f) := \mathsf{E}[\exp\{-\int_{\mathbb{X}} f(x)\Phi(\mathrm{d}x)\}]$, for any measurable function $f : \mathbb{X} \to \mathbb{R}_+$.

Any point process $\Phi$ can be represented as $\Phi = \sum_{j\geq 1} \delta_{\tilde{X}_j}$, where $(\tilde{X}_j)_{j\geq 1}$ is a sequence of random variables taking values in $\mathbb{X}$, and $\delta_x$ denotes the Dirac delta mass at $x$. In the sequel, we will deal with *simple* point processes, for which the atoms of $\Phi$ are a.s. distinct, i.e., $\mathsf{P}(\tilde{X}_i = \tilde{X}_j) = 0$ for all $i \neq j$. The mean measure $M_\Phi$ associated with a point process $\Phi$ is defined as $M_\Phi(B) = \mathsf{E}[\Phi(B)]$ for any Borel set $B \in \mathscr{X}$. The *k-th order factorial* moment measure $M_\Phi^{(k)}$ of $\Phi$ is the mean measure of the $k$-th factorial power of $\Phi$, i.e., of the point process $\Phi^{(k)}$ defined as:

$$\Phi^{(k)} := \sum_{(j_1,\dots,j_k)}^{\neq} \delta_{(\tilde{X}_{j_1},\dots,\tilde{X}_{j_k})},$$

where the symbol $\neq$ over the summation means that the sum is extended over all pairwise distinct indexes. Therefore, $M_\Phi^{(k)}(B) = \mathsf{E}[\Phi^{(k)}(B)]$, where $B$ is a Borel set in $\mathbb{X}^k$.

We now introduce two of the most relevant classes of point processes in the chapter.

**Example 4.1** (Poisson and mixed Poisson processes)**.** Let $\nu$ be a locally finite measure on $(\mathbb{X}, \mathscr{X})$. We say that $\Phi$ is a Poisson process on $\mathbb{X}$ with intensity measure $\nu$ if, for any $B \in \mathscr{X}$, $\Phi(B)$ is a Poisson random variable with mean $\nu(B)$, and for every collection of disjoint Borel sets $B_1, \ldots, B_k$ the random variables $\Phi(B_1), \ldots, \Phi(B_k)$ are independent, for any $k \geq 1$. In the sequel, we will write $\Phi \sim \mathrm{PP}(\nu)$. In this example, the mean measure $M_\Phi$ equals $\nu$ and the $k$-th factorial moment measure satisfies $M_\Phi^{(k)} = \nu^k$.

Second, to define a mixed Poisson process, let $\gamma$ be a positive finite random variable with law $f_\gamma$ on $\mathbb{R}_+$, we say that $\Phi$ is a mixed Poisson process directed by $f_\gamma$ if $\Phi \,|\, \gamma \sim \mathrm{PP}(\gamma\nu)$ and $\gamma \sim f_\gamma$. We will write $\Phi \sim \mathrm{MP}(\nu, f_\gamma)$. It is easy to see that the $k$-th factorial moment measure satisfies $M_\Phi^{(k)} = \mathsf{E}[\gamma^k]\nu^k$.

**Example 4.2** (Binomial and mixed binomial processes)**.** Let $\nu$ be a probability measure on $(\mathbb{X}, \mathscr{X})$, and consider a sequence of i.i.d. random variables $(\tilde{X}_j)_{j\geq 1}$ with common distribution $\nu$. Then, for any integer $m \geq 1$, the point process $\Phi_m := \sum_{j=1}^m \delta_{\tilde{X}_j}$ is termed a binomial point process, denoted as $\Phi_m \sim \mathrm{BP}_m(\nu)$.

Let $M$ be a random variable taking values in the set of nonnegative integers with probability mass function $q_M$ (possibly degenerate), independent of $(\tilde{X}_j)_{j\geq 1}$. We say that the point process $\Phi = \sum_{j=1}^M \delta_{\tilde{X}_j}$ is a mixed binomial process and we write $\Phi \sim \mathrm{MB}(\nu, q_M)$. It is easy to see that the $k$-th factorial moment measure equals $M_\Phi^{(k)} = \mathsf{E}[M^{(k)}]\nu^k$, where $\mathsf{E}[M^{(k)}]$ is the $k$-th factorial moment of $M$, i.e., $\mathsf{E}[M^{(k)}] = \mathsf{E}[M(M-1)\cdots(M-k+1)]$. With a slight abuse of notation, we will also write $\Phi \sim \mathrm{MB}(\nu, q_M)$ even when $\nu$ is a generic finite measure, with the understanding that this is interpreted after normalization.

Our mathematical treatment of extended feature allocations is based on Palm calculus. Hence, we now review key definitions and concepts, following the notation of Baccelli et al. (2020). The Campbell measure $\mathscr{C}_\Phi$ of $\Phi$ is a measure on $\mathscr{X} \times \mathcal{M}_\mathbb{X}$ defined as $\mathscr{C}_\Phi(C \times L) := \mathsf{E}[\Phi(C)\mathbb{1}(\Phi \in L)]$ for $C \in \mathscr{X}$ and $L \in \mathcal{M}_\mathbb{X}$. Provided that the mean measure $M_\Phi$ is $\sigma$-finite, the Palm kernel $\{\mathbf{P}_\Phi^x\}_{x\in\mathbb{X}}$ of $\Phi$ is defined as the (a.s.) unique disintegration probability kernel of $\mathscr{C}_\Phi$ with respect to $M_\Phi$, i.e.,

$$\mathscr{C}_\Phi(C \times L) = \int_C \mathbf{P}_\Phi^x(L) M_\Phi(\mathrm{d}x).$$

Formally, the Palm kernel is an extension of the concept of regular conditional distribution to the case of point processes. Informally, $\mathbf{P}_\Phi^x$ can be understood as the probability distribution of $\Phi$ given that $\Phi$ has an atom at $x$. A point process $\Phi_x$ with distribution $\mathbf{P}_\Phi^x$ is called the *Palm version of* $\Phi$ *at* $x \in \mathbb{X}$, and of course $\mathsf{P}(\Phi_x(\{x\}) \geq 1) = 1$, i.e., $x$ is a trivial atom of $\Phi_x$. As a consequence, one can define the point process $\Phi_x^! := \Phi_x - \delta_x$, which is called the *reduced Palm version of* $\Phi$ *at* $x \in \mathbb{X}$, whose associated reduced Palm kernel is indicated by $\mathbf{P}_\Phi^{!x}$. In a similar fashion, under the assumption that the $k$-th factorial moment measure $M_\Phi^{(k)}$ is $\sigma$-finite, it is possible to construct the family of $k$-th order Palm distributions $\{\mathbf{P}_\Phi^{\boldsymbol{x}}\}_{\boldsymbol{x}\in\mathbb{X}^k}$, and the generic probability measure $\mathbf{P}_\Phi^{\boldsymbol{x}}$ can be interpreted as the distribution of $\Phi$ given that $\boldsymbol{x} = (x_1, \ldots, x_k)$ are atoms of $\Phi$. Again, by removing the trivial atoms $(x_1, \ldots, x_k)$, we obtain the reduced Palm distributions $\mathbf{P}_\Phi^{!\boldsymbol{x}}$, namely the probability law of

$$\Phi_{\boldsymbol{x}}^! := \Phi_{\boldsymbol{x}} - \sum_{j=1}^k \delta_{x_j}.$$

For the remainder of the chapter, we denote by $\mathbb{X}$ the space of all possible feature labels. The point processes central to our statistical modeling, introduced in the next section, are defined on the product space $\mathbb{X} \times (0, 1]$.

### 4.1.2 Extended feature allocation models

Let $(\tilde{X}_j)_{j\geq 1}$ be a sequence of $\mathbb{X}$-valued random variables, hereafter interpreted as "feature labels". An extended feature allocation model is a stochastic model for observations $(Z_i)_{i\geq 1}$ such that each $Z_i$ is characterized by its expressed features, or equivalently by the sequence of pairs $((\tilde{X}_j, \tilde{A}_{ij}))_{j\geq 1}$, where $\tilde{A}_{ij} = 1$ if the feature $\tilde{X}_j$ belongs to the $i$-th individual, $\tilde{A}_{ij} = 0$ otherwise. Equivalently, we can represent each $Z_i$ as a point process on $\mathbb{X}$, namely

$$Z_i = \sum_{j\geq 1} \tilde{A}_{ij} \delta_{\tilde{X}_j}. \tag{4.1}$$

We assume that, for any $j \geq 1$, the $\tilde{A}_{ij}$'s are conditionally i.i.d. Bernoulli variables with a strictly positive random probability $\tilde{q}_j$, which entails both exchangeability of the $Z_i$'s and regularity of the feature allocation (Broderick et al., 2013). Namely, no feature appears in only a single observation within $(Z_i)_{i\geq 1}$, or, equivalently, if a label is observed in some individual, then it is observed in infinitely many individuals with probability one.

The feature labels $\tilde{X}_j$'s and parameters $\tilde{q}_j$'s, i.e., the probability of displaying feature $\tilde{X}_j$, are collected in the simple point process on $\mathbb{X} \times (0, 1]$ defined by $\Psi := \sum_{j\geq 1} \delta_{(\tilde{X}_j, \tilde{q}_j)}$ or in its functional

$$\tilde{\mu}(B) = \int_{\mathbb{X}\times(0,1]} s \mathbb{1}_B(x) \Psi(\mathrm{d}x\, \mathrm{d}s), \quad B \in \mathscr{X}, \tag{4.2}$$

where $\mathbb{1}_B$ denotes the indicator function of the Borel set $B$. We say that $Z_i$ in (4.1) is a Bernoulli process with parameter $\tilde{\mu}$, indicated as $Z_i \sim \mathrm{BP}(\tilde{\mu})$. According to our point process formulation, we observe that $Z_i$ in (4.1) is obtained by first thinning $\Psi$ with retention function $p(x, s) = s$ and then removing the second component. Summing up, we are dealing with the following statistical model

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \mathrm{BP}(\tilde{\mu}), \\ \tilde{\mu} &\sim \mathscr{Q}, \end{aligned} \tag{4.3}$$

where $\mathscr{Q}$ is the law of $\tilde{\mu}$. In other terms, $(Z_i)_{i\geq 1}$ is a sequence of exchangeable observations with de Finetti measure $\mathscr{Q}$.

We call model (4.3) an *extended feature allocation model*. Here, extended refers to the generality of the prior $\mathscr{Q}$, compared to the previous proposals in the literature. Indeed, previous models focused exclusively on i.i.d. feature labels. Moreover, the distributions for the jumps $(\tilde{q}_j)_{j\geq 1}$ previously considered were based either on the Poisson process or suitable transformations. By contrast, we allow for arbitrary interactions among the points of $\Psi$. For instance, the feature labels could exhibit a repulsive or attractive behavior, or the weights could depend on all the feature labels.

### 4.1.3 Related models

Extended feature allocation models can be used as standard feature allocation models (Broderick et al., 2013), where the main objective is to model the sharing of features

across sampled individuals. Indeed, consider a sample $\boldsymbol{Z}$ from (4.3), and suppose that it contains $K_n = k$ features with labels $X_1, \dots, X_k$. One can recover the associated feature allocation $F_n = (B_{n,1}, \dots, B_{n,K_n})$, where the set $B_{n,\ell}$ contains the indexes of the individuals exhibiting the $\ell$-th feature, with label $X_\ell$, as $\ell = 1, \dots, k$. Feature allocation models describe probability distributions on the object $F_n$. Remarkably, the class of extended feature allocation models encompasses the class of feature frequency models, and therefore induces the entire family of regular feature allocation models admitting an EFPF.

Another class of related model is the one of *trait processes* (Campbell et al., 2018), reviewed in Section 2.4, which generalize model (4.3) by assuming $Z_i = \sum_{j \geq 1} \tilde{A}_{ij} \delta_{\tilde{X}_j}$ where $\tilde{A}_{ij}$ are integer or real-valued random variables. Consequently, the weights of the random measure $\tilde{\mu}$ might not need to be restricted on $(0, 1]$. Examples of trait processes include the beta negative binomial process where the $\tilde{A}_{ij}$'s follow a negative binomial distribution and $\tilde{\mu}$ is distributed as a beta process. See Broderick et al. (2015) and Heaukulani and Roy (2016) for a detailed treatment. We finally mention that there exists a generalization of the EFPF to trait allocation models, called exchangeable trait probability function (Campbell et al., 2018).

## 4.2 Bayesian analysis of extended feature models

In the present section, we provide the full Bayesian analysis of the class of extended feature allocation models. We focus on the statistical model (4.3) to describe the marginal distribution of a sample, the predictive structure of the next observation, and the posterior representation of the underlying measure $\tilde{\mu}$. Our results are instrumental in proving the sufficientness postulates of Section 4.3, but they are also interesting *per se*, as showcased in the application (cf. Section 4.5).

Throughout the rest of the chapter, we assume that the $k$-th factorial moment measures of $\Psi$, denoted as $M_\Psi^{(k)}$, are $\sigma$-finite for all $k \geq 1$. This is a very mild assumption (for instance, it is verified by all feature allocation models previously considered in the literature) and is essential for the existence of the $k$-th order Palm distributions. Furthermore, under this assumption, the measure disintegration theorem (Kallenberg, 2021, Theorem 3.4) entails

$$M_\Psi^{(k)}(\mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{s}) = \rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x}) \tilde{m}_\xi^{(k)}(\mathrm{d}\boldsymbol{x}), \tag{4.4}$$

where $\tilde{m}_\xi^{(k)}(\,\cdot\,)$ is a $\sigma$-finite measure *equivalent* to $M_\Psi^{(k)}(\,\cdot \times (0,1]^k)$, i.e., both measures are absolutely continuous with respect to each other, and $\rho^{(k)}$ is a kernel from $\mathbb{X}^k$ to $(0,1]^k$.

Remarkably, another common assumption for feature allocation models is that any subject displays a.s. a finite number of features, meaning $Z_i(\mathbb{X}) < \infty$ a.s.. We provide some considerations on this hereafter.

**Remark 4.1.** *The property $Z_i(\mathbb{X}) < \infty$ a.s. is always guaranteed when $\Psi$ has a.s. a finite number of points, for example, when $\Psi$ is a mixed binomial process. Some considerations are needed for priors which entail an infinite number of points for $\Psi$. Since $Z_i$ is obtained from thinning $\Psi$ and discarding the second component, it follows that the mean measure of any $Z_i$ is $M_Z(B) = \int_{(0,1]} s M_\Psi(B \times \mathrm{d}s)$, $B \in \mathscr{X}$. Therefore, a sufficient condition for $Z_i(\mathbb{X}) < \infty$ a.s. stems from $\mathsf{E}[Z_i(\mathbb{X})] < \infty$, which corresponds to $\int_{\mathbb{X} \times (0,1]} s M_\Psi(\mathrm{d}x\, \mathrm{d}s) < \infty$.*

*However, this is not a necessary condition in general. It can be proved that a necessary condition on the mean measure of $\Psi$ to have $Z_i(\mathbb{X}) < \infty$ a.s. is the following:*

$$\int_{\mathbb{X}\times(0,1]} (1-s) M_\Psi(\mathrm{d}x\,\mathrm{d}s) = \infty.$$

*See Proposition 4.3 for the formal statement and proof. Notably, if $\Psi$ is a Poisson process with (infinite) mean measure $\nu$ or a mixed Poisson $\mathrm{MP}(\nu, f_\gamma)$, then the condition $\int_{\mathbb{X}\times(0,1]} s\nu(\mathrm{d}x\,\mathrm{d}s) < \infty$ is both necessary and sufficient for $Z_i(\mathbb{X}) < \infty$ a.s., as stated in Proposition 4.4. We defer to Proposition 4.5 further implications of the assumption $Z_i(\mathbb{X}) < \infty$ a.s. under Poisson processes.*

### 4.2.1 General formulas: marginal, posterior and predictive distributions

We start by describing the marginal distribution of a sample $\boldsymbol{Z} := (Z_1, \ldots, Z_n)$ from an extended feature model as in (4.3). The proofs are based on an application of Palm calculus (Baccelli et al., 2020) and deferred to Section 4.D.

**Theorem 4.1.** *Let $\boldsymbol{Z}$ be a sample from the statistical model (4.3), where $\tilde{\mu}$ is the functional of a point process $\Psi$ defined via (4.2). The probability that $\boldsymbol{Z}$ displays $k$ features labeled $\boldsymbol{x} = (X_1, \ldots, X_k)$ with corresponding vector of frequency counts $\boldsymbol{m} := (m_1, \ldots, m_k)$ is*

$$\int_{(0,1]^k} \mathsf{E}\left\{ e^{\int_{\mathbb{X}\times(0,1]} n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(\mathrm{d}z\,\mathrm{d}t)} \right\} \prod_{\ell=1}^k s_\ell^{m_\ell}(1-s_\ell)^{n-m_\ell} \rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x}) \cdot \tilde{m}^{(k)}_\xi(\mathrm{d}\boldsymbol{x}),$$

*where $\rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x})$ and $\tilde{m}^{(k)}_\xi(\mathrm{d}\boldsymbol{x})$ are defined as in (4.4), and, for $\boldsymbol{x} \in \mathbb{X}^k$ and $\boldsymbol{s} \in (0,1]^k$, $\Psi^!_{\boldsymbol{x},\boldsymbol{s}}$ is the reduced Palm version of $\Psi$ at $((x_1, s_1), \ldots, (x_k, s_k))$.*

We remark that the law of the induced feature allocation $F_n$ is obtained by marginalizing out the labels $\boldsymbol{x}$ and dividing by $k!$. Clearly, such a law admits EFPF for any extended feature model considered. This means that the induced law on $F_n$ can always be expressed as a symmetric function of the frequency counts $m_1, \ldots, m_k$. Moreover, all these extended feature models are regular by construction.

We proceed by describing the general posterior distribution of the point process $\Psi$, or equivalently the random measure $\tilde{\mu}$, given a sample of size $n$ from (4.3).

**Theorem 4.2.** *Let $\boldsymbol{Z}$ be a sample from the statistical model (4.3), where $\tilde{\mu}$ is the functional of a point process $\Psi$ defined via (4.2). Further, suppose that $\boldsymbol{Z}$ displays $k$ features having labels $\boldsymbol{x} = (X_1, \ldots, X_k)$ with corresponding vector of frequency counts $\boldsymbol{m} := (m_1, \ldots, m_k)$. Then, the posterior distribution of $\tilde{\mu}$, conditionally on $\boldsymbol{Z}$, satisfies the distributional equality*

$$\tilde{\mu} \,|\, \boldsymbol{Z} \overset{d}{=} \sum_{\ell=1}^k q_\ell \delta_{X_\ell} + \mu', \tag{4.5}$$

*where*

*(i) $\boldsymbol{q} := (q_1, \ldots, q_k)$ is a vector of positive random variables with joint distribution*

$$f_{\boldsymbol{q}}(\mathrm{d}\boldsymbol{s}) \propto \mathsf{E}\left\{ e^{\int_{\mathbb{X}\times(0,1]} n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(\mathrm{d}z\,\mathrm{d}t)} \right\} \prod_{\ell=1}^k s_\ell^{m_\ell}(1-s_\ell)^{n-m_\ell} \rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x});$$

*(ii) conditionally on* $\boldsymbol{q}$*, the random measure* $\mu'$ *admits the representation* $\mu' = \sum_{j\geq 1} q'_j \delta_{\tilde{X}'_j}$*, where the distribution of* $\Psi' := \sum_{j\geq 1} \delta_{(\tilde{X}'_j, q'_j)}$ *is absolutely continuous with respect to the distribution of* $\Psi^!_{\boldsymbol{x},\boldsymbol{q}}$*, with density*

$$f'_\Psi(\nu) \propto e^{\int_{\mathbb{X}\times(0,1]} n \log(1-t)\nu(\mathrm{d}z\,\mathrm{d}t)}. \tag{4.6}$$

From Theorem 4.2, the posterior distribution of $\tilde{\mu}$ is decomposed in a part that involves previously observed labels out of the sample and a component that involves hitherto unseen features, i.e., $\mu'$. Note that $\mu'$ in point (ii) is an a.s. discrete random measure whose Laplace functional is available and equals

$$\mathsf{E}\left\{e^{-\int_{\mathbb{X}} f(z)\mu'(\mathrm{d}z)} \mid \boldsymbol{q}\right\} = \frac{\mathsf{E}\left\{e^{-\int_{\mathbb{X}\times(0,1]} tf(z) - n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{q}}(\mathrm{d}z\,\mathrm{d}t)}\right\}}{\mathsf{E}\left\{e^{\int_{\mathbb{X}\times(0,1]} n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{q}}(\mathrm{d}z\,\mathrm{d}t)}\right\}}. \tag{4.7}$$

The posterior representation resembles available results in the literature; see, for example, (James, 2017, Theorem 3.1), with the fundamental difference that here the distribution of $\mu'$ depends on the previously observed features $\boldsymbol{x}$ via the reduced Palm version of $\Psi$, thus allowing interacting feature labels.

To conclude the general Bayesian analysis of the extended feature models in (4.3), we provide the predictive distribution of the next observation, conditionally on the available sample, which easily follows from Theorem 4.2 by a standard disintegration argument.

**Theorem 4.3.** *Under the same assumptions of Theorem 4.2, the conditional distribution of* $Z_{n+1}$*, given the sample* $\boldsymbol{Z}$*, satisfies the following distributional equality*

$$Z_{n+1} \mid \boldsymbol{Z} \stackrel{d}{=} \sum_{\ell=1}^{k} A_{n+1,\ell}\delta_{X_\ell} + Z'_{n+1}, \tag{4.8}$$

*where*

*(i)* $(A_{n+1,1}, \ldots, A_{n+1,k})$*, conditionally on* $q_1, \ldots, q_k$*, is a vector of independent Bernoulli random variables with parameters* $(q_1, \ldots, q_k)$*, which have been defined in Theorem 4.2;*

*(ii)* $Z'_{n+1} \sim \mathrm{BP}(\mu')$ *and* $\mu'$ *is distributed according to Theorem 4.2.*

Note that, in the point process language, the latter measure $Z'_{n+1}$ in (ii) is obtained by first thinning the process $\Psi' = \sum_{j\geq 1} \delta_{(\tilde{X}'_j, q'_j)}$ with retention probability $p(x, s) = s$, and then discarding the second component. Theorem 4.3 shows that the next individual $Z_{n+1}$ has a positive probability of displaying the features $X_1, \ldots, X_k$ observed out of the initial sample, and it can display hitherto unseen features, which correspond to the atoms of $Z'_{n+1}$.

Interestingly, the distribution of $Z'_{n+1}$ depends on the whole sampling information, including the feature labels $\boldsymbol{x}$, indeed it is a Bernoulli process with parameter $\mu'$, whose Laplace functional equals (4.7). This is a crucial difference with respect to previous works. Indeed, from (James, 2017, Proposition 3.2) it is apparent that if $\tilde{\mu}$ is a CRM, the distribution of the component involving hitherto unseen features depends on $\boldsymbol{Z}$ only through the sample size $n$. As for the scaled processes in (Camerlenghi et al., 2024, Proposition 2), $Z'_{n+1}$ depends on $n$, $k$, and the frequency counts, but again not on the feature labels.

## 4.3 Predictive characterizations

In the same spirit as sufficientness postulates for species sampling models, we now consider the problem of characterizing prior distributions in extended feature allocation models, leading to predictive distributions that satisfy specific properties. In particular, we focus our analysis on how the Bernoulli process $Z'_{n+1}$ in (4.8) depends on the sampling information encoded in $\boldsymbol{Z}$. As we will clarify in Section 4.6, there is no gain in considering $Z_{n+1}$ in place of $Z'_{n+1}$.

From point (ii) of Theorem 4.3, it is clear that the distribution of $Z'_{n+1}$ is uniquely characterized by the random measure $\mu'$ in (6.10) with Laplace functional (4.7). Therefore, in general, the predictive distribution of $Z'_{n+1}$ depends on the sample size $n$, as well as the couples of feature labels $\boldsymbol{x}$ and their frequencies $\boldsymbol{m}$. Of course, from exchangeability, the order in which features are recorded, as well as the indices of observations displaying each feature, is irrelevant. However, special cases of extended feature models, previously studied in the literature, lead to much simpler predictive laws. For instance, the treatment in James (2017) (see also Corollary 4.1 below) shows that if $\tilde{\mu}$ is a CRM, the law of $Z'_{n+1}$ depends only on the sample size $n$. Instead, in the case of the stable beta scaled process in Camerlenghi et al. (2024) and for feature models having a product form EFPF (Battiston et al., 2018), analyzed in Chapter 3, such predictions may also depend on the number of distinct features $k$.

### 4.3.1 Sufficientness postulates

We start by characterizing the class of extended feature allocation models for which the law of $Z'_{n+1}$ depends on the initial sample $\boldsymbol{Z}$ only through the sample size $n$.

**Theorem 4.4** (Sufficientness postulate for the dependence on $n$)**.** *Consider a sample $\boldsymbol{Z}$ from the statistical model* (4.3)*. The distribution of $Z'_{n+1}$ in Theorem 4.3 depends on the observed sample $\boldsymbol{Z}$ solely through the sample size $n$ if and only if $\Psi$ in model* (4.3) *is a Poisson process.*

This theorem in the feature setting is the most natural counterpart of the results by Regazzini (1978); Lo (1991), who characterized the Dirichlet process as the unique species sampling prior in which the probability of observing a new species depends only on the sample size, and the probability of observing a previously recorded species depends on both the sample size and its frequency. The proof of Theorem 4.4 is deferred to the Appendix. Central to the proof is a novel characterization of the Poisson point process in terms of its (higher-order) reduced Palm distribution that might be of independent interest. We report it in the following lemma.

**Lemma 4.1.** *Let $\Phi$ be a point process with locally finite mean measure $M_\Phi$. Then the following assertions are equivalent.*

*(i)* *$\Phi$ is a Poisson process.*

*(ii)* *The law of $\Phi^!_{\boldsymbol{x}}$ does not depend on $\boldsymbol{x}$. That is, for $M_\Phi^{(k)}$-almost all $\boldsymbol{x} = (x_1, \ldots, x_k)$ and $M_\Phi^{(m)}$-almost all $\boldsymbol{y} = (y_1, \ldots, y_m)$, it holds that $\Phi^!_{\boldsymbol{x}} \stackrel{d}{=} \Phi^!_{\boldsymbol{y}}$.*

*Moreover, if (ii) holds, then* $\Phi^!_{\boldsymbol{x}} \stackrel{d}{=} \Phi^!_{\boldsymbol{y}} \stackrel{d}{=} \Phi$.

Other important characterization theorems for species sampling models are contained in Zabell (2005); Bacallado et al. (2017), where both the Pitman-Yor process and Gibbs-type priors are characterized in terms of the prediction rules they induce. According to these characterizations, both the sample size and the number of observed species in the sample play a pivotal role. As for extended feature allocation models, we can prove similar sufficientness postulates. Specifically, we are going to characterize all the models for which the distribution of $Z'_{n+1}$ depends on $\boldsymbol{Z}$ solely on the sample size $n$ and the number of observed features $k$.

**Theorem 4.5** (Sufficientness postulate for the dependence on $n$ and $k$)**.** *Consider a sample* $\boldsymbol{Z}$ *from the statistical model* (4.3)*. The distribution of* $Z'_{n+1}$ *depends on the observed sample* $\boldsymbol{Z}$ *solely through* $n$ *and* $k$ *if and only if* $\Psi$ *in model* (4.3) *is a mixed Poisson or mixed binomial process.*

As for Theorem 4.4, we need a characterization of mixed Poisson and mixed binomial processes to prove Theorem 4.5. We provide it in the next lemma, whose proof is a slight extension of a result in Kallenberg (1973).

**Lemma 4.2.** *Let* $\Phi$ *be a point process with locally finite mean measure* $M_\Phi$*. Then the following statements are equivalent.*

*(i)* $\Phi$ *is a mixed Poisson or a mixed binomial point process.*

*(ii) The law of* $\Phi^!_{\boldsymbol{x}}$ *depends on* $\boldsymbol{x}$ *only through its cardinality. That is, for* $M_\Phi^{(k)}$*-almost all* $\boldsymbol{x} = (x_1, \ldots, x_k)$ *and* $\boldsymbol{y} = (y_1, \ldots, y_k)$*, it holds that* $\Phi^!_{\boldsymbol{x}} \stackrel{d}{=} \Phi^!_{\boldsymbol{y}}$.

Theorem 4.4 fully characterizes the class of CRM priors analyzed in James (2017). Moreover, Theorem 4.5 characterizes a broad class of prior distributions that includes, among others, all feature models having a product form EFPF (Battiston et al., 2018), as well as the stable beta scaled processes analyzed in Camerlenghi et al. (2024). From Battiston et al. (2018) and Chapter 3, it is not difficult to realize that the point process $\Psi$ associated with a feature model having a product form EFPF could be either $(i)$ a mixed binomial point processes, i.e., $\Psi = \sum_{j=1}^{M} \delta_{(\tilde{X}_j, \tilde{q}_j)}$ (cf. Example 4.2), where $M$ is random, the $\tilde{q}_j$'s are i.i.d. beta distributed and the $\tilde{X}_j$'s are i.i.d. from a diffuse measure further independent of the $\tilde{q}_j$'s, or $(ii)$ a mixed Poisson processes $\mathrm{MP}(\nu, f_\gamma)$ with $\nu$ being the Lévy intensity of a three parameter beta process (Teh and Gorur, 2009) and $\gamma$ a positive random variable (cf. Example 4.1). However, these classes of processes are only very special cases of mixed binomial and mixed Poisson processes. Hence, while a general treatment of Bayesian feature models with predictions depending exclusively on the sample size $n$ was already available in James (2017), the corresponding framework for predictions based solely on $n$ and the number of distinct features $k$ remains only partially developed. We address this gap in Section 4.4 below.

### 4.3.2 A fresh look at scaled processes

We now clarify the connection between a stable beta scaled process (Camerlenghi et al., 2024) and a mixed Poisson process, which has been only mentioned at the end of the last section. A scaled process is a random measure $\tilde{\mu} = \sum_{j\geq 1} \tilde{q}_j \delta_{\tilde{X}_j}$, where the $\tilde{X}_j$'s are i.i.d. random variables from a base measure $G_0$ on $\mathbb{X}$, while $(\tilde{q}_j)_{j\geq 1}$ arises as a suitable transformation of a Poisson process on the positive real line with intensity measure $\rho(\mathrm{d}s)$. More precisely, to define the sequence of weights, one considers the jumps $(\Delta_j)_{j\geq 1}$ of a Poisson process in decreasing order and set $\tilde{q}_j^{\Delta_1} := \Delta_{j+1}/\Delta_1$. Then, the distribution of $(\tilde{q}_j)_{j\geq 1}$ is obtained by mixing the conditional distribution of $(\tilde{q}_j^{\Delta_1})_{j\geq 1}$, given $\Delta_1$, with respect to a new distribution for the largest jump $\Delta_1$. See Camerlenghi et al. (2024) for additional details.

Under a scaled process prior for the model (4.3), the distribution of $Z'_{n+1}$ in (4.8) may depend on $n$, $k$ and $\boldsymbol{m}$, and often involves intractable expression (Camerlenghi et al., 2024, Proposition 2). The notable tractable case is represented by scaled processes obtained from stable subordinators, i.e., $\rho(\mathrm{d}s) = cs^{-1-\alpha}\mathrm{d}s$ for some constants $c > 0$ and $\alpha \in (0, 1)$, referred to as stable beta scaled processes. In this case, $Z'_{n+1}$ depends on the sample only through $n$ and $k$. However, Theorem 4.5 claims that such a dependence is retained if and only if $\Psi$ is a mixed Poisson or a mixed binomial process and, at a first glance, the definition of scaled processes does not align with that of mixed Poisson processes. This apparent inconsistency can be solved by resorting to an alternative construction of scaled processes that can be evinced either by (James et al., 2015, Theorem 1.1) or (Camerlenghi et al., 2024, Lemma 1).

**Proposition 4.1.** *Let $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ be the Lévy intensity of a subordinator, $C$ a positive random variable, and $G_0$ a diffuse measure on $\mathbb{X}$. Let $\Psi = \sum_{j\geq 1} \delta_{(\tilde{X}_j, \tilde{q}_j)}$ be such that*

$$\Psi \,|\, C \sim \mathrm{PP}\left(C\rho(Cs)\mathbb{1}_{(0,1]}(s)\mathrm{d}s\, G_0(\mathrm{d}x)\right).$$

*Then, $\tilde{\mu} = \sum_{j\geq 1} \tilde{q}_j \delta_{\tilde{X}_j}$ is a scaled process.*

The construction outlined in Proposition 4.1 is available for any scaled process, as introduced at the beginning of the section. It becomes clear that $\Psi$ is a mixed Poisson process if and only if for any $s$ and any $C$, the following factorization holds $C\rho(Cs) = h(C)\tilde{\rho}(s)$ where $\tilde{\rho}(s)$ is a Lévy density and $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a function. When $\rho(s) = s^{-1-\alpha}$, i.e., the case of stable subordinators, we get $C\rho(Cs) = C^{-\alpha}s^{-1-\alpha}$ so that in this case the scaled process $\tilde{\mu}$ is induced by a mixed Poisson process $\Psi$ directed by the law of $C^{-\alpha}$. Finally, the predictive characterization (Camerlenghi et al., 2024, Theorem 1) of stable beta scaled processes within the class of scaled processes is straightforward from Theorem 4.5. Indeed, the Lévy intensities of stable subordinators are the only $\rho$ such that $C\rho(Cs) = h(C)\tilde{\rho}(s)$. Therefore, scaled processes obtained from stable subordinators are the only scaled processes induced by a mixed Poisson process $\Psi$. Finally, Theorem 4.5 highlights how the dependence of the distribution of $Z'_{n+1}$ solely on $n$ and $k$ is not a distinctive feature of stable beta scaled processes but pertains to the larger class of priors obtained from mixed Poisson and mixed binomial processes.

## 4.4 Detailed analysis of specific extended feature models

The previous section highlights the distinctive role of Poisson, mixed Poisson, and mixed binomial process priors in shaping the dependencies of the induced predictive distributions on the observed sample. In this section, we specialize the general Bayesian analysis developed in Section 4.2.1 to these notable classes of prior distributions, as well as to a prior distribution based on a determinantal point process (Hough et al., 2006) which yields predictions for $Z'_{n+1}$ depending on $n$, $k$ and the observed features $\boldsymbol{x}$. Finally, we introduce a tractable model leading to dependence on the whole frequency spectrum.

### 4.4.1 The Poisson process prior

Consider the model (4.3), where $\tilde{\mu}$ is as in (4.2) and $\Psi$ is a Poisson process on $\mathbb{X} \times (0,1]$ with an (infinite) mean measure $\nu$. This is one of the most popular cases in the literature, and the Indian buffet process arises as a specific example. Under the standard assumption that each observation $Z_i$ exhibits a finite number of features a.s., the $k$-th factorial moment measures are $\sigma$-finite, as shown in Proposition 4.5, point (i). The disintegration in (4.4) writes as $\nu(\mathrm{d}x\,\mathrm{d}s) = \rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$, where $\rho$ is a kernel from $\mathbb{X}$ to $(0,1]$ and $G_0$ is a $\sigma$-finite measure on $\mathbb{X}$. The Bayesian analysis of model (4.3) under a Poisson process prior for $\Psi$ can be addressed by specializing theorems of Section 4.2 as in the next result. Let us introduce some shorthand notations that will be useful throughout this section: $\rho_n(\mathrm{d}s \,|\, x) = (1-s)^n \rho(\mathrm{d}s \,|\, x)$.

**Corollary 4.1** (Bayesian analysis under the Poisson process)**.** *Consider a sample $\boldsymbol{Z}$ from the statistical model* (4.3)*, where $\tilde{\mu}$ is the functional of a Poisson point process $\Psi$ with intensity measure $\nu$ defined via* (4.2)*.*

*(i) The marginal distribution of the sample $\boldsymbol{Z}$, equals*

$$e^{-\varphi_n} \prod_{\ell=1}^{k} \int_{(0,1]} (1-s)^{n-m_\ell} s^{m_\ell} \rho(\mathrm{d}s \,|\, X_\ell) G_0(\mathrm{d}X_\ell),$$

*where we defined*

$$\varphi_n = \sum_{i=0}^{n-1} \int_{\mathbb{X}\times(0,1]} s\rho_i(\mathrm{d}s \,|\, x) G_0(\mathrm{d}x). \tag{4.9}$$

*(ii) The posterior distribution of $\tilde{\mu}$ satisfies the distributional equality in* (6.10)*, where in this case $\mu'$ is a* CRM *with Lévy intensity measure $\rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$. The weights $q_\ell$'s of previously observed features are independent random variables, and independent of $\mu'$, with marginal density $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell}\rho(\mathrm{d}s \,|\, X_\ell)$, as $\ell = 1, \ldots, k$.*

*(iii) The predictive distribution of a future observation $Z_{n+1}$, given the sample $\boldsymbol{Z}$, equals the distribution of the measure defined in* (4.8)*, where $Z'_{n+1}$ is a Poisson process on $\mathbb{X}$ with (finite) intensity measure given by $\int_{(0,1]} s\rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$, and the $A_{n+1,\ell}$'s are independent Bernoulli random variables with parameters $q_\ell$'s, as $\ell = 1, \ldots, k$.*

The previous corollary is in agreement with the theory developed by James (2017) for the Poisson process case. We first point out that the Poisson process $Z'_{n+1}$ in point (iii) depends

on the observable sample $\boldsymbol{Z}$ only through the sample size $n$, and this is not surprising by virtue of Theorem 4.4. We also observe that if $\rho(\mathrm{d}s \,|\, x) = \rho(\mathrm{d}s)$ is independent of the location, then $G_0$ can be taken as a probability measure, as evident from Proposition 4.5, point (ii), and $\tilde{\mu}$ turns out to be a homogeneous CRM. In this case, $Z'_{n+1}$ is a mixed binomial process $\mathrm{MB}(G_0, q_{K'})$, where $q_{K'}$ is the probability mass function of the Poisson distribution with mean $\lambda_n = \int_{(0,1]} s\rho_n(\mathrm{d}s)$, where $\rho_n(\mathrm{d}s) = (1-s)^n\rho(\mathrm{d}s)$. In other words, $Z'_{n+1}$ has $K' \sim \mathrm{Poi}(\lambda_n)$ points, represented as $Z'_{n+1} = \sum_{\ell=1}^{K'} \delta_{X'_\ell}$, where the $X'_\ell$'s are i.i.d. from $G_0$. Again, the distribution of $Z'_{n+1}$ depends on the sample only through the sample size $n$.

### 4.4.2 The mixed Poisson process prior

Let now $\Psi \sim \mathrm{MP}(\nu, f_\gamma)$, where $\nu$ is a (infinite) locally finite measure on $\mathbb{X} \times (0,1]$. That is, $\Psi$ is a Poisson process with random mean measure $\gamma\nu$ and $\gamma \sim f_\gamma$ is a positive random variable, cf. Example 4.1. As usual, we assume that $Z_i(\mathbb{X}) < \infty$ a.s.. Moreover, we suppose that $\gamma$ has finite moments of any order $k \geq 1$. Then, the $k$-th factorial moment measures are $\sigma$-finite for any $k$ (cf. Proposition 4.5) and $\nu(\mathrm{d}x\,\mathrm{d}s) = \rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$, where $\rho$ is a kernel from $\mathbb{X}$ to $(0,1]$ and $G_0$ is a $\sigma$-finite measure on $\mathbb{X}$. The following corollary specifies the Bayesian analysis of this extended feature allocation model.

**Corollary 4.2** (Bayesian analysis under the mixed Poisson process)**.** *Consider a sample $\boldsymbol{Z}$ from the statistical model* (4.3)*, where $\tilde{\mu}$ is the functional of a mixed Poisson point process $\Psi$, i.e., $\Psi \sim \mathrm{MP}(\nu, f_\gamma)$, defined via* (4.2)*.*

*(i) The marginal distribution of $\boldsymbol{Z}$ equals*

$$\mathsf{E}\left(e^{-\gamma\varphi_n}\gamma^k\right)\prod_{\ell=1}^{k}\int_{(0,1]}(1-s)^{n-m_\ell}s^{m_\ell}\rho(\mathrm{d}s \,|\, X_\ell)G_0(\mathrm{d}X_\ell),$$

*where $\varphi_n$ is defined as in* (4.9)*.*

*(ii) The posterior distribution of $\tilde{\mu}$ satisfies the distributional equality in* (6.10)*, where $\mu' = \sum_{j\geq 1} q'_j\delta_{\tilde{X}'_j}$ and $\{(\tilde{X}'_j, q'_j)\}_{j\geq 1}$ are the points of a mixed Poisson point process $\Psi' \sim \mathrm{MP}(\nu', f_{\tilde{\gamma}})$, with*

$$\nu'(\mathrm{d}x\,\mathrm{d}s) = \rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x) \quad \textit{and} \quad f_{\tilde{\gamma}}(\mathrm{d}\gamma) \propto e^{-\gamma\varphi_n}\gamma^k f_\gamma(\mathrm{d}\gamma).$$

*In addition, the $q_\ell$'s in* (6.10) *are independent random variables, and independent of $\mu'$, with marginal density $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell}\rho(\mathrm{d}s \,|\, X_\ell)$, as $\ell = 1, \ldots, k$.*

*(iii) The predictive distribution of $Z_{n+1}$, given the sample $\boldsymbol{Z}$, satisfies the distributional equality in* (4.8)*, where $Z'_{n+1}$ is a mixed Poisson process on $\mathbb{X}$ with distribution*

$$Z'_{n+1} \sim \mathrm{MP}\Big(\int_{(0,1]} s\rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x), f_{\tilde{\gamma}}\Big), \tag{4.10}$$

*where the first parameter is a finite measure on $\mathbb{X}$. Moreover, the $A_{n+1,\ell}$'s in* (4.8) *are independent Bernoulli random variables with parameters $q_\ell$'s, as $\ell = 1, \ldots, k$.*

We emphasize that the process $Z'_{n+1}$ in (4.10) depends on the initial sample $\boldsymbol{Z}$ only through the sample size $n$ and the number of distinct features $k$, which appear in the mixing law $f_{\tilde{\gamma}}$, and on no additional sampling information, as expected from Theorem 4.5. Finally, if the kernel $\rho$ does not depend on the location $x$, i.e., $\rho(\mathrm{d}s \,|\, x) = \rho(\mathrm{d}s)$, then $G_0$ can be taken as a probability measure (see Proposition 4.5, point (ii)) and $\tilde{\mu}$, conditionally on $\gamma$, turns out to be a homogeneous CRM. Thus, thanks to Lemma 4.3, $Z'_{n+1}$ is a mixed binomial process distributed as $Z'_{n+1} \sim \mathrm{MB}(G_0, q_{K'} \,|\, \tilde{\gamma})$, where $q_{K'} \,|\, \tilde{\gamma}$ is a Poisson density with parameter $\tilde{\gamma} \int_{(0,1]} s\rho_n(\mathrm{d}s)$, where $\rho_n(\mathrm{d}s) = (1-s)^n \rho(\mathrm{d}s)$, conditionally on a positive random variable $\tilde{\gamma}$ with probability distribution $f_{\tilde{\gamma}}$. The stable beta scaled process in Camerlenghi et al. (2024) and the mixtures of Indian buffet processes analyzed in Chapter 3 fall under the umbrella of models considered in this section.

### 4.4.3 The mixed binomial process prior

We now consider the other class of priors identified in Theorem 4.5, i.e., mixed binomial processes. Let $\Psi \sim \mathrm{MB}(\nu, q_M)$, where $q_M$ is a probability mass function on the nonnegative integers with finite moments and $\nu$ is a probability measure on $\mathbb{X} \times (0, 1]$. Therefore, the disintegration $\nu(\mathrm{d}x\,\mathrm{d}s) = \rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$ always holds, where $G_0$ is a probability measure on $\mathbb{X}$, and $\rho$ is a probability kernel from $\mathbb{X}$ to $(0, 1]$. This follows from a suitable application of (Kallenberg, 2021, Theorem 3.4). The Bayesian analysis of model (4.3) under the mixed binomial process prior for $\Psi$ is presented in the next result. For ease of notation, we will use $\kappa_n$ to represent the integral $\kappa_n = \int_{\mathbb{X}\times(0,1]} \rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$.

**Corollary 4.3** (Bayesian analysis under the mixed binomial process)**.** *Consider a sample $\boldsymbol{Z}$ from the statistical model* (4.3)*, where $\tilde{\mu}$ is the functional of a mixed binomial point process $\Psi$, i.e., $\Psi \sim \mathrm{MB}(\nu, q_M)$, defined via* (4.2)*.*

*(i) The marginal distribution of $\boldsymbol{Z}$ equals*

$$\mathsf{E}\left\{(\kappa_n)^{\tilde{M}}\right\} \mathsf{E}(M^{(k)}) \prod_{\ell=1}^{k} \int_{(0,1]} s^{m_\ell}(1-s)^{n-m_\ell} \rho(\mathrm{d}s \,|\, X_\ell) G_0(\mathrm{d}X_\ell),$$

*where the first expected value is taken with respect to a nonnegative integer-valued random variable $\tilde{M}$ having probability mass function*

$$q_{\tilde{M}}(m) \propto q_M(m+k) \frac{(m+k)!}{m!}.$$

*(ii) The posterior distribution of $\tilde{\mu}$ satisfies the distributional equality in* (6.10)*, where $\mu' = \sum_{j\geq 1} q'_j \delta_{\tilde{X}'_j}$ and $\{(\tilde{X}'_j, q'_j)\}_{j\geq 1}$ are the points of a mixed binomial process $\Psi' \sim \mathrm{MB}(\nu', q_{M'})$ with*

$$\nu'(\mathrm{d}s\,\mathrm{d}x) = \rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x) \quad \text{and} \quad q_{M'}(m) \propto (\kappa_n)^m q_{\tilde{M}}(m).$$

*Moreover, the $q_\ell$'s are independent random variables, and independent of $\mu'$, with marginal density $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell}\rho(\mathrm{d}s \,|\, X_\ell)$, as $\ell = 1, \ldots, k$.*

*(iii) The predictive distribution of a future observation $Z_{n+1}$, given the sample $\boldsymbol{Z}$, satisfies the distributional equality in* (4.8)*, where $Z'_{n+1}$ is a mixed binomial point process on*

$\mathbb{X}$ *with parameters* $(\tilde{G}, q_{K'})$, *with* $\tilde{G}(\mathrm{d}x) = \int_{(0,1]} s\rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$ *and*

$$q_{K'}(m) \propto \sum_{z \geq m} \binom{z}{m} \kappa_n^z q_M(z+k) c_p^m (1-c_p)^{z-m} \frac{(z+k)!}{z!},$$

*having set*

$$c_p = \frac{\int_{\mathbb{X}\times(0,1]} s\rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)}{\int_{\mathbb{X}\times(0,1]} \rho_n(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)}.$$

*Moreover, the* $A_{n+1,\ell}$*'s in* (4.8) *are independent Bernoulli random variables with parameters* $q_\ell$*'s, as* $\ell = 1, \ldots, k$.

As expected, the distribution of $Z'_{n+1}$ in point (iii) of the previous corollary depends on the sample only through the sample size $n$ and the number of observed features $k$. A special case of interest is obtained by assuming that $\rho(\mathrm{d}s \,|\, x) = \rho(\mathrm{d}s)$, so that $\rho_n(\mathrm{d}s \,|\, x) = \rho_n(\mathrm{d}s) = (1-s)^n\rho(\mathrm{d}s)$. Under this assumption, $Z'_{n+1}$ is a mixed binomial process $\mathrm{MB}(G_0, q_{K'})$ and $q_{K'}$ does not depend on $G_0$. We further specialize this case to two choices of $M \sim q_M$.

If $M \sim \mathrm{Poi}(\lambda)$, then $\Psi$ is a Poisson process with finite intensity measure. In this case, $\Psi'$ is a Poisson process with finite intensity $\lambda\rho_n(\mathrm{d}s)G_0(\mathrm{d}x)$, having $M' \sim \mathrm{Poi}(\rho_n(0,1])$ points. In addition, $Z'_{n+1}$ is a Poisson process with finite intensity, having $K'$ points, with $K'$ being a Poisson random variable with mean $\lambda \int_{(0,1]} s\rho_n(\mathrm{d}s)$.

Let now $M$ be a negative binomial random variables with parameters $(r, p)$, where $p \in [0,1]$ denotes the success probability and $r > 0$ is the number of successes. In this case, $\Psi'$ is a mixed binomial process with $M'$ points, where $M'$ is a negative binomial random variable with updated parameters $r' = r + k$ and $p' = 1 - \kappa_n(1-p)$. Moreover, the process involving new features $Z'_{n+1}$ is a mixed binomial process with $K'$ points, where $K'$ is a negative binomial with parameters $r^{th} = r + k$ and

$$p^{th} = \frac{p'}{p' + (1-p)\int_{(0,1]} s\rho_n(\mathrm{d}s)}.$$

We finally point out that the Gibbs-type feature models with finitely many features introduced in Chapter 3 correspond to specific choices of the two models specified above.

### 4.4.4 The independently marked (repulsive) determinantal process prior

In this section, we provide a prior for $\Psi$ such that the distribution of $Z'_{n+1}$ in the prediction rule (4.8) depends on the sample also through the observed feature labels $\boldsymbol{x}$. To this end, consider a determinantal point process (DPP) $\xi$ on a compact region $R \subset \mathbb{R}^d$ (Hough et al., 2006; Lavancier et al., 2015). The DPP $\xi$ is specified by a covariance kernel $C : R \times R \to \mathbb{C}$, such that $M_\xi^{(k)}$ has density with respect to the $k$-fold product of the Lebesgue measure given by

$$\eta^{(k)}(x_1, \ldots, x_k) = \det\{C(x_h, x_w)_{h,w=1,\ldots,k}\}, \qquad x_1, \ldots, x_k \in R,$$

where $C(x_h, x_w)_{h,w=1,\ldots,k}$ is the $k \times k$ matrix with entries $C(x_h, x_w)$. Now $\Psi = \sum_{j=1}^M \delta_{(\tilde{X}_j, \tilde{q}_j)}$ in (4.2) is obtained by marking each point in $\xi = \sum_{j=1}^M \delta_{\tilde{X}_j}$ by independent random variables $\tilde{q}_j \,|\, \tilde{X}_j = x_j \sim H(\cdot \,|\, x_j)$ with values in $(0,1]$, that is $\Psi$ is an *independently*

*marked point process with ground point process $\xi$ and mark kernel $H$*, according to Baccelli et al. (2020). We write for convenience $\Psi \sim \text{imDPP}(C, H)$, where $C$ is the covariance kernel of the determinantal ground process and $H$ is the mark probability kernel. As for any independently marked process, the $k$-th factorial moment measure equals $M_\Psi^{(k)}(\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{s}) = M_\xi^{(k)}(\mathrm{d}\boldsymbol{x}) \prod_{\ell=1}^k H(\mathrm{d}s_\ell \,|\, x_\ell)$ and it is $\sigma$-finite, since $M_\xi^{(k)}$ is $\sigma$-finite.

We also recall that the reduced Palm version of a DPP is still a DPP. Namely, for any $\boldsymbol{x} = (x_1, \ldots, x_k)$ such that the $x_j$'s are distinct, $\xi^!_{\boldsymbol{x}}$ is a DPP with kernel $K_{\boldsymbol{x}}(y_1, y_2) = C(y_1, y_2) - \tilde{c}_{\boldsymbol{x}}(y_1)^T \tilde{\boldsymbol{C}}^{-1} \tilde{c}_{\boldsymbol{x}}(y_2)$, for any $y_1, y_2 \in R$, where $\tilde{c}_{\boldsymbol{x}}(y) = (C(y, x_1), \ldots, C(y, x_k))^T$ and $\tilde{\boldsymbol{C}} = C(x_h, x_w)_{h,w=1,\ldots,k}$. See, e.g., Lavancier and Rubak (2023) for further details.

**Corollary 4.4** (Bayesian analysis under the independently marked DPP)**.** *Consider a sample $\boldsymbol{Z}$ from the statistical model* (4.3)*, where $\tilde{\mu}$ is the functional of an independently marked point process $\Psi$, i.e., $\Psi \sim \text{imDPP}(C, H)$, defined via* (4.2)*. Let $\boldsymbol{x} = (X_1, \ldots, X_k)$.*

*(i) The marginal distribution of the sample $\boldsymbol{Z}$ equals*

$$\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left[-\log\left\{\int_{(0,1]}(1-s)^n H(\mathrm{d}s \,|\, x)\right\}\right] \prod_{\ell=1}^k \int_{(0,1]} s^{m_\ell}(1-s)^{n-m_\ell} H(\mathrm{d}s \,|\, X_\ell) \cdot M_\xi^{(k)}(\mathrm{d}\boldsymbol{x}),$$

*where we remind that $\mathcal{L}_{\xi^!_{\boldsymbol{x}}}$ denotes the Laplace functional of the DPP $\xi^!_{\boldsymbol{x}}$.*

*(ii) The posterior distribution of $\tilde{\mu}$ satisfies the distributional equality in* (6.10)*, where the weights $q_\ell$'s are independent random variables, further independent of $\mu'$, with marginal density $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell} H(\mathrm{d}s \,|\, X_\ell)$, as $\ell = 1, \ldots, k$. Moreover, $\mu'$ in* (6.10) *can be represented as $\mu' = \sum_{j=1}^{M'} q'_j \delta_{\tilde{X}'_j}$, where the $\tilde{X}'_j$'s are the atoms of a point process $\xi'$ on $\mathbb{X}$ specified by the Laplace functional*

$$\mathcal{L}_{\xi'}(f) = \frac{\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left\{f - \log\int_{(0,1]}(1-s)^n H(\mathrm{d}s \,|\, x)\right\}}{\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left\{-\log\int_{(0,1]}(1-s)^n H(\mathrm{d}s \,|\, x)\right\}}, \tag{4.11}$$

*and the $q'_j$'s are independent marks with conditional density $q'_j \,|\, \tilde{X}'_j = x'_j \sim H'(\cdot \,|\, x'_j)$, where $H'(\mathrm{d}s \,|\, x'_j) \propto (1-s)^n H(\mathrm{d}s \,|\, x'_j)$.*

*(iii) The predictive distribution of $Z_{n+1}$, given the sample $\boldsymbol{Z}$, satisfies the distributional equality in* (4.8)*, where $Z'_{n+1}$ is a Bernoulli process with parameter $\mu'$, and the $A_{n+1,\ell}$'s are independent Bernoulli random variables with parameters $q_\ell$'s, as $\ell = 1, \ldots, k$.*

Some remarks are in order. First, observe that in the Poisson, mixed Poisson, and mixed binomial examples, the atoms $\tilde{X}_j$'s of the underlying point process $\Psi$ are marginally i.i.d. from some distribution on $\mathbb{X}$. On the other hand, if $\Psi$ is built starting from a DPP or any other repulsive process, the $\tilde{X}_j$'s have a joint law that encourages a priori "well separated" configurations, i.e., for $h \neq w$, the probability that $\tilde{X}_h$ "is close to" $\tilde{X}_w$ is small. In addition, under this prior choice, the process $Z'_{n+1}$ appearing in the predictive distribution of Corollary 4.4, point (iii), depends on the sample $\boldsymbol{Z}$ through the feature labels $\boldsymbol{x}$, since the parameter $\mu'$ depends on $\boldsymbol{x}$, as well as the sample size $n$ and the cardinality $k$ of $\boldsymbol{x}$. However, such a distribution is not affected by other sample statistics, i.e., the frequency counts. We conclude with an example, where we focus on a specific choice of the distribution of the $\tilde{q}_j$'s.

**Example 4.3.** We consider $\Psi \sim \text{imDPP}(C, H)$, where $H(\cdot \,|\, x)$ does not depend on $x$ and equals the beta distribution with parameters $(a, b)$, namely $\tilde{q}_j \overset{\text{iid}}{\sim} \text{Beta}(a, b)$, where $\text{Beta}(a, b)$ denotes the beta distribution. In this particular case, the distributional results in Corollary 4.4 simplify. It is worth noticing that the posterior distribution of $\tilde{\mu}$ satisfies the distributional equality in (6.10), where each weight $q_\ell$ has a beta distribution with parameters $(m_\ell + a, n - m_\ell + b)$, as $\ell = 1, \ldots, k$. Moreover, the distribution of the point process $\xi'$, and, as a consequence, of $\mu'$ in (6.10), becomes much more manageable. Indeed, from (4.11), the distribution of $\xi'$ has a density with respect to the distribution of $\xi^!_{\boldsymbol{x}}$ given by

$$f_{\xi'}(\nu) \propto \left\{ \frac{B(a, b+n)}{B(a, b)} \right\}^{\nu(\mathbb{X})} =: g(n; a, b)^{\nu(\mathbb{X})},$$

having denoted by $B(a, b)$ the Euler's beta function. The associated marks $q'_j$'s are i.i.d. from a beta distribution with parameters $(a, b+n)$. In more detail, the distribution of the number of points in $\xi'$, denoted with $M'$, has a density with respect to the distribution of $\xi^!_{\boldsymbol{x}}(\mathbb{X})$ given by

$$f_{M'}(m) \propto \left\{ \frac{B(a, b+n)}{B(a, b)} \right\}^{m}. \tag{4.12}$$

Finally, the mean measure $M_{\xi'}$ has density with respect to $M_{\xi^!_{\boldsymbol{x}}}$ defined as

$$m_{\xi'}(y) := g(n; a, b) \mathsf{E}\left[ f_{\xi'}\left\{ \xi^!_{(\boldsymbol{x}, y)} \right\} \right].$$

### 4.4.5 Predictions depending on the whole frequency spectrum

In the examples analyzed so far, we have found predictive distributions for $Z'_{n+1}$ depending either solely on $n$, or $n$ and $k$, or $n$ and $\boldsymbol{x}$ (including $k$). As shown in James et al. (2015); Camerlenghi et al. (2024) and as discussed at the end of Section 4.3, in general, scaled processes yield predictive distributions for the newly discovered features which depend on the sampling information through $n, k$ and the frequency spectrum $m_1, \ldots, m_k$. However, outside the case of stable beta scaled processes, the resulting expressions for the posterior and marginal distributions (and the associated computations) are somewhat involved. For instance, if one considers scaled processes built from gamma and generalized gamma subordinators, the distribution of $\mu'$ involves intractable integrals that must be evaluated numerically.

Here, we propose a different and straightforward strategy to induce predictive distribution depending on the whole frequency spectrum while maintaining computational convenience. The idea is simple and generally applicable: starting from the expression of the Lévy intensity $\rho$ of a subordinator, assign a prior distribution to a parameter that does not enter only multiplicatively in the expression. For simplicity and specificity, we will consider here the case of the three-parameter beta process (Teh and Gorur, 2009; Broderick et al., 2015), i.e., model (4.3) with $\tilde{\mu}$ a homogeneous CRM having Lévy intensity measure

$$\rho(\mathrm{d}s)\, G_0(\mathrm{d}x) = s^{-1-\alpha}(1-s)^{\beta+\alpha-1} \mathbb{1}_{(0,1]}(s) \mathrm{d}s\, G_0(\mathrm{d}x), \tag{4.13}$$

where $\alpha \in (0, 1)$ and $\beta > -\alpha$. The next proposition shows how considering $\alpha$ random yields the desired predictive distribution.

**Proposition 4.2.** *Consider a sample $\boldsymbol{Z}$ from the statistical model* (4.3)*, where $\tilde{\mu}$ is the functional of a point process $\Psi$ defined via* (4.2) *and $\Psi$ is such that*

$$\begin{aligned}\Psi \,|\, \alpha &\sim \mathrm{PP}(\gamma s^{-1-\alpha}(1-s)^{\beta+\alpha-1}\mathbb{1}_{(0,1]}(s)\mathrm{d}s\, G_0(\mathrm{d}x)),\\ \alpha &\sim \pi_\alpha,\end{aligned}$$

*where $\gamma > 0, \beta > 0$, $G_0$ is a diffuse probability measure on $\mathbb{X}$ and $\pi_\alpha$ is supported on $(0,1)$. Then, if $\pi_\alpha$ is not a degenerate (Dirac) measure, the predictive distribution of $Z'_{n+1}$, given the sample $\boldsymbol{Z}$, depends on $n$, $k$ and $m_1, \ldots, m_k$.*

To the best of our knowledge, there is no conjugate prior for $\alpha$. However, inference under this model is straightforward by means of Markov chain Monte Carlo algorithms. Indeed, conditionally to $\alpha$, marginal, posterior, and predictive distributions are readily available, and the posterior distribution of $\alpha$ has a simple density function that is amenable to posterior simulation algorithms.

## 4.5 An application of extended feature allocation models to spatial statistics

We consider here an extension to the problem proposed by Ord (1978), see also Diggle (2013). Informally, assume that trees are randomly distributed in a forest. Each observation $Z_i$ is generated by a surveyor who, walking in the forest, annotates the location of some trees. Thus, $Z_i$ contains the locations recorded by the $i$-th surveyor. Ord (1978) assumes the scenario in which a single surveyor is present, selecting trees at random, and considers the problem of estimating the total number of trees in the forest using only a single set of recorded locations. We adopt a slightly weakened hypothesis by assuming that the probability of observing each tree is independent of the location but could be influenced by some unavailable tree-specific characteristic (e.g., height), eventually modeled as a random variable. Our goals extend beyond those of Ord (1978). In addition to estimating the number of trees in the forest by leveraging multiple observations, we also aim to identify the locations of unobserved trees.

### 4.5.1 Fitting details and numerical implementation

Formally, we assume that $(Z_i)_{i\geq 1}$ follows (4.3), where $\tilde{\mu} = \sum_{j\geq 1} \tilde{q}_j \delta_{\tilde{X}_j}$, and the $\tilde{X}_j$'s correspond to the locations of the trees in the forest, whereas the $\tilde{q}_j$'s are the tree-specific probabilities of observing tree $j$ in any survey. We also indicate the point process containing all the locations by $\xi = \sum_{j\geq 1} \delta_{\tilde{X}_j}$. We further assume that the $\tilde{q}_j$'s are i.i.d. beta random variables with common parameters $(a, b)$, and $\xi$ is a DPP on a rectangular region $R \subset \mathbb{R}^2$, as in Section 4.4.4. Indeed, it is well understood that trees often exhibit repulsive behavior. For the purpose of illustration, we will assume that $\xi$ follows a Gaussian DPP (Lavancier et al., 2015) with parameters $(\rho, \alpha)$, subject to the condition $\rho < (\pi\alpha^2)^{-1}$ to ensure the process is well-defined. The covariance kernel of this Gaussian DPP is given by $C(x, y) = \rho \exp\{-\|(x-y)/\alpha\|^2\}$. We refer to Møller and Waagepetersen (2003); Lavancier et al. (2015) for other examples of repulsive point processes.

Supposing we observe $n$ surveys $Z_1, \ldots, Z_n$, which report the locations of $k$ distinct trees, we now address the problem of predicting the number and locations of the missing

trees. The posterior distribution of the total number of trees is equal to $M' + k$, where the law of $M'$ is given in Example 4.3. In addition to estimating the number of trees in the forest, a natural and more challenging question is to locate the unobserved trees. With the notation of Example 4.3, the infinitesimal probability that an unobserved tree would occupy position $\mathrm{d}x$ equals $\mathsf{E}\{\Psi'(\mathrm{d}x \times (0,1])\} = M_{\xi'}(\mathrm{d}x)$.

Both tasks require to handle the distribution of $\xi^!_{\boldsymbol{x}}(\mathbb{X})$, for some specific points $\boldsymbol{x}$. From Corollary 4.4, $\xi^!_{\boldsymbol{x}}$ is a DPP on $R$ with kernel $K_{\boldsymbol{x}}$, and associated Mercer decomposition $K_{\boldsymbol{x}}(y_1, y_2) = \sum_{k\geq 1} \lambda^*_k \varphi^*_k(y_1)\overline{\varphi^*_k(y_2)}$. Then, from Hough et al. (2006) (see also Lavancier et al., 2015), we have that $\xi^!_{\boldsymbol{x}}(\mathbb{X})$ follows a Poisson-binomial distribution with parameters $(\lambda^*_k)_{k\geq 1}$. Unfortunately, as discussed in Lavancier and Rubak (2023), the eigendecomposition of $K_{\boldsymbol{x}}$ is generally not analytically available. We consider two approaches: the first consists of approximating the eigendecomposition numerically, and the second exploits an approximation of the Poisson-binomial distribution that does not require such an eigendecomposition.

To numerically approximate the $\lambda^*_k$'s, we proceed as follows. Let $R_g$ be a grid of $(N_g)^2$ equispaced points in $R$, define the $N_g^2 \times N_g^2$ matrix $\tilde{\boldsymbol{K}}$ by $\tilde{\boldsymbol{K}}_{i,j} = K_{\boldsymbol{x}}(y_{1,i}, y_{2,j})$ for $(y_{1,i}, y_{2,j}) \in R_g \times R_g$. Let $\Delta = \prod_{i=1}^2 L_i/N_g$ where $L_i$ is the length of the $i$-th side of $R$. Of course, $\tilde{\boldsymbol{K}}$ is positive definite since $K_{\boldsymbol{x}}$ is a covariance kernel, with eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{N_g^2}$. Then, we set $\lambda^*_k \approx \tilde{\lambda}_k \Delta$ for $k = 1, \ldots, N_g^2$ and $\lambda^*_k = 0$ for $k > N_g^2$, and approximate the law of $\xi^!_{\boldsymbol{x}}(\mathbb{X})$ with a Poisson-binomial distribution of parameters $(\tilde{\lambda}_1\Delta, \ldots, \tilde{\lambda}_{N_g^2}\Delta)$. In our experiments presented in the subsequent sections, the decreasingly ordered sequence of $\lambda^*_k$ decreases extremely fast with $k$ so that the truncation error is negligible. We checked the accuracy of the numerical approximation of $\lambda^*_k$ against the analytical value in the case of a Gaussian covariance, reporting errors of the order of 0.01 for the largest 50 eigenvalues. To compute the probability mass function of a Poisson-binomial distribution, we use the Python package `fast-poibin`. Evaluating the distribution on the whole support takes less than one second for $N_g^2 = 2500$. Further speed-ups can be achieved by truncating the series of eigenvalues earlier, for example, by keeping only those eigenvalues exceeding a pre-specified threshold.

An alternative strategy would focus on approximating the Poisson-binomial distribution via Le Cam's theorem (Steele, 1994), i.e., approximating the law of $\xi^!_{\boldsymbol{x}}(\mathbb{X})$ with a Poisson distribution with parameter $\sum_{k\geq 1} \lambda^*_k$. Since the sum of eigenvalues is equal to the trace of $K_{\boldsymbol{x}}$, i.e., $\int_R K_{\boldsymbol{x}}(y, y)\mathrm{d}y$, using Le Cam's approximation does not require performing any numerical eigendecomposition. However, Le Cam's approximation introduces a non-negligible error. In particular, the total variation between the true distribution of $\xi^!_{\boldsymbol{x}}(\mathbb{X})$ and its approximation is bounded by above by $\sum_{k\geq 1}(\lambda^*_k)^2$.

Finally, as illustrated in this context, the distribution of the independently marked DPP $\Psi$ is controlled by a set of hyperparameters that must be specified. Specifically, we need to elicit the parameters $(a, b)$ of the beta distribution for the marks $\tilde{q}_j$'s, as well as the parameters $(\rho, \alpha)$ of the Gaussian DPP $\xi$. To this end, we resort to an empirical Bayes approach, choosing such hyperparameters by maximizing the likelihood function (4.3) of the model with respect to these four parameters.

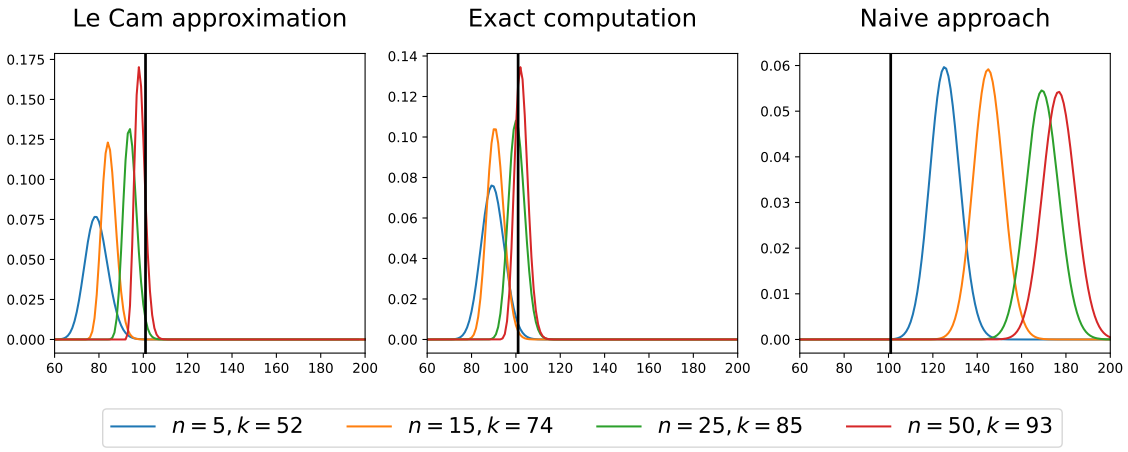



Figure 4.5.1: Posterior distribution of the total number of trees in the synthetic scenario of Section 4.5.2. From left to right: calculations performed using Le Cam's approximation of the Poisson-binomial, exact computations, and posterior of $\xi^!_{\boldsymbol{x}}(\mathbb{X})$. Different line colors correspond to different sample sizes; the black vertical line indicates the true number of trees.

### 4.5.2 Synthetic scenarios

We generate data by simulating the true point process of the "trees", denoted with $\xi_0$, from a Gaussian DPP on $[0,1]^2$ with intensity parameter $\rho = 100$ and scale $\alpha = 0.0535$. To the points in $\xi_0$ we attach i.i.d. marks $\tilde{q}_j$'s from the beta distribution of parameters $(1,5)$, therefore obtaining a realization for $\Psi$ and $\tilde{\mu}$. Observations $Z_i$'s are obtained by simulating i.i.d. Bernoulli processes conditionally to $\tilde{\mu}$, mimicking the collection of data from the surveyors.

For different sample sizes $n \in \{5, 15, 25, 50\}$, we compute the distribution of the total number of trees as $M' + k$, where $M'$ is defined in Example 4.3, and compare it with the total number of trees in $\xi_0$. We compare inference obtained using the exact computation of the Poisson-binomial distribution and Le Cam's approximation. Additionally, we consider a naive approach where we disregard the $Z_i$'s, retaining only the distinct locations $\boldsymbol{x}$ and computing the law of $\xi^!_{\boldsymbol{x}}$. Beyond predicting the total number of trees, we address the problem of locating the unobserved trees through $M_{\xi'}$, as discussed in Section 4.5.1. When adopting the naive approach, this task is tackled by considering the mean measure of $\xi^!_{\boldsymbol{x}}$. All results presented here are obtained by estimating the hyperparameters of the model via the empirical Bayes approach described in Section 4.5.1. For completeness, Section 4.G in the Appendix provides the corresponding analyses under an oracle scenario, where all hyperparameters are fixed at their true values.

Figures 4.5.1 and 4.5.2 highlight the key features of our predictions: in estimating the total number of trees, Figure 4.5.1 clearly shows that the naive approach performs poorly across all sample sizes. Le Cam's approximation, albeit faster, introduces some errors that lead to slightly underestimating the number of trees. In contrast, the exact posterior looks centered around the true number of trees, with its variance shrinking as the sample size increases. Figure 4.5.2 addresses the problem of locating unseen trees for a sample size of $n = 15$, comparing our proposed model with the naive approach. Refer to

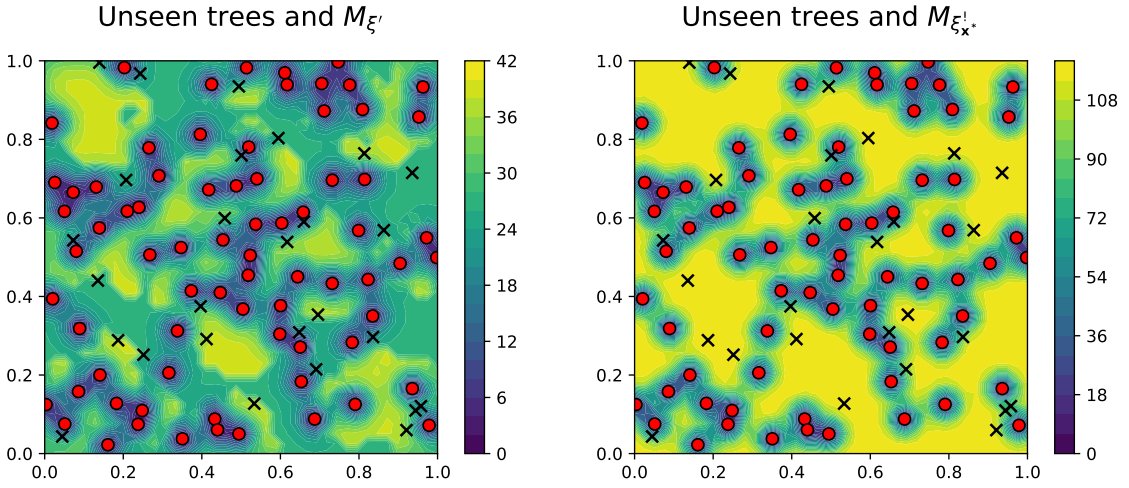



Figure 4.5.2: Locating the unobserved trees for $n = 15$ in the synthetic scenario of Section 4.5.2: infinitesimal probability of observing an unseen tree in a given location. Left plot: the mean measure of $\xi'$. Right plot: the mean measure of $\xi^!_{\boldsymbol{x}}$. The red dots represent the observed trees in the sample. The black crosses indicate the unseen trees. Note that the color scales of the two plots are different.

Section 4.G in the Appendix for the corresponding plots when $n \in \{5, 25, 50\}$. Beyond the substantial discrepancy between the two scales, caused by the naive approach significantly overestimating the number of trees, further remarks are needed. Specifically, consider the infinitesimal probability of observing an unseen tree at location $\mathrm{d}x$. Since $\xi^!_{\boldsymbol{x}}$ is a DPP, the naive approach assumes that the farther $\mathrm{d}x$ is from the observed trees $\boldsymbol{x}$, the more likely it is to contain an unseen tree. On the other hand, the repulsive pattern of $\xi'$ is peculiar, leading to a different behavior: our model predicts that it is unlikely to find a tree at $\mathrm{d}x$ if it is too close to an observed tree, as expected. However, it also predicts that locations "too far" from observed trees may have small probabilities of containing a tree, as they might not align with the accumulation patterns estimated from the data.

Finally, we briefly discuss the utility of imposing repulsiveness between the points $\tilde{X}_j$'s of $\Psi$. Consider any prior process for $\Psi$ where the points $\tilde{X}_j$'s are marginally i.i.d. from some diffuse distribution $G_0$, such as the Poisson, mixed Poisson, or mixed binomial processes. From the predictive characterizations in Theorem 4.4 and Theorem 4.5, we know that the locations of the unseen trees $\tilde{X}'_j$'s in $\Psi'$ are independent of the observed locations $\boldsymbol{x}$. Furthermore, these locations remain marginally i.i.d. from $G_0$, as demonstrated in Corollary 4.1, Corollary 4.2, and Corollary 4.3. Therefore, for these models, the task of locating the unobserved trees is answered by $G_0$, which provides no informative structure about the spatial arrangement of the unseen trees. This highlights the advantage of incorporating repulsiveness in this spatial illustration, as our model provides a more informative and structured prediction of tree locations.

### 4.5.3 Analysis of Norwegian spruces

We analyze the `spruces` dataset from the R package `spatstat`, which contains the spatial locations of 134 Norwegian spruce trees in a natural forest stand in Saxony, Germany. These tree locations are represented as the point configuration $\xi_0$. We assign an i.i.d.

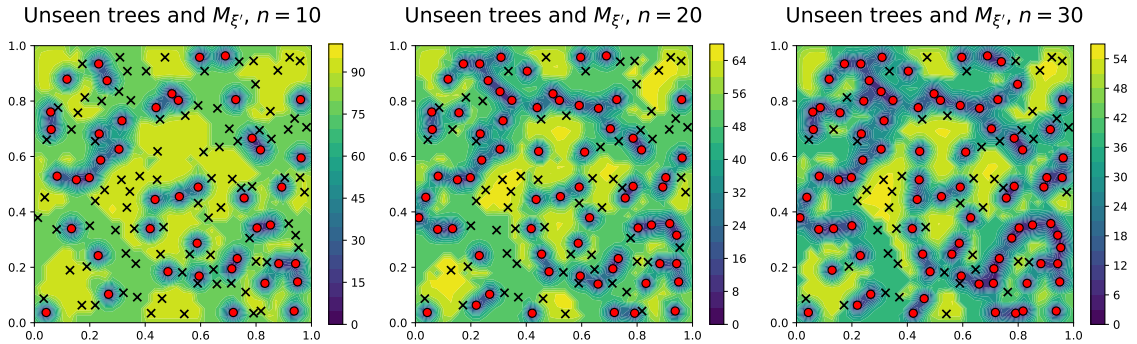



Figure 4.5.3: Locating the unobserved trees for $n \in \{10, 20, 30\}$ in the analysis of the spruces dataset of Section 4.5.3: infinitesimal probability of observing an unseen tree in a given location. The three plots report $M_{\xi'}$ for the three sample sizes. The red dots represent the observed trees in the sample. The black crosses indicate the unseen trees. Note that the plots have different color scales.

mark $\tilde{q}_j$ sampled from a beta distribution with parameters $(1, 20)$ to each tree. Using these marks, we generate surveys $Z_i$'s by retaining each tree with probability $\tilde{q}_j$. We consider samples of increasing sizes $n \in \{10, 20, 30\}$ and estimate model hyperparameters using the empirical Bayes approach.

We infer the total number of trees as $M' + k$, where $M'$ is defined in Example 4.3. Employing Le Cam's approximation, our model provides reliable predictions, although with a slight underestimation. Specifically, the expected values of $M' + k$ are respectively $115, 111, 114$, for the increasing sample sizes. This systematic underestimation aligns with the comments of Figure 4.5.1, where exact computations yielded more accurate results than Le Cam's approximation. The choice of Le Cam's approximation is justified by its computational efficiency.

The main problem we address is the prediction of the locations of unobserved trees. In our framework, this is achieved via $M_{\xi'}$, as detailed in Section 4.5.1, which is reported in Figure 4.5.3. Across all sample sizes, our predictions exhibit the peculiar repulsive structure highlighted in the simulated example. The regions predicted to most likely contain unseen trees align well with the locations of the actual unobserved trees, as indicated by the black crosses in the figure. The difference in plot scales arises from the reduction in the number of unseen trees as the sample size increases.

## 4.6 Discussion

Feature allocations are routinely employed in several applied fields ranging from text mining to online A/B testing. In this work, we have investigated the class of extended feature allocation models, which allows for generic dependence structures across the feature labels and weights. As noted by Lee et al. (2023), assuming a repulsive dependence among features reduces overfitting and leads to more interpretable posterior summaries of hierarchical models where features are present at a latent level, which are common in tasks such as image segmentation (Griffiths and Ghahramani, 2011; Broderick et al., 2015), matrix factorization (Zhou et al., 2012, 2017), and cellular biology (Lee et al., 2023). Our analysis lays the foundation for such models and their applications since the posterior and

predictive distributions can be leveraged to design posterior inference algorithms.

Within the class of extended feature models, we have characterized those priors leading to *simple* predictive distributions, whereby the probability distribution of "new" features depends exclusively on the sample size, or on the sample size and the number of distinct features in the sample. These postulates are entirely absent from the existing literature and serve as essential tools to guide the prior elicitation. Moreover, predictive characterizations offer a bridge between the classical Bayesian paradigm, based on likelihood and prior, and the *predictive* approach to inference, predicated by de Finetti, that recently gained popularity (Fong et al., 2023; Berti et al., 2025), by allowing the statistician to choose a system of prior and likelihood based on their subjective idea of the predictive distribution.

Sufficientness postulates for species sampling models characterize specific prior distributions such as the finite Dirichlet distribution, the Dirichlet process, and the Pitman-Yor process (Bacallado et al., 2017), by imposing conditions on the probability of observing a new species and the probability of re-observing a species recorded in the sample. By contrast, our postulates focus only on the distribution of new features and characterize broad classes of priors. It is then natural to wonder if, by adding further conditions on previously observed features, it is possible to restrict the characterization to specific prior distributions. This question has a negative answer, as clarified by a simple counterexample. In fact, for any Poisson, mixed Poisson, or mixed binomial prior, the probability of re-observing a feature depends exclusively on the sample size and the frequency of that feature. This excludes the possibility of distinguishing within the class of CRMs by adding structural constraints on the predictive law for the previously observed features.

The present work opens several opportunities for future research. First, a promising research direction is the further development of the model based on the determinantal point process, as described in Section 4.4.4, along with its extension to high dimensional settings. The resulting prior will be particularly useful as a latent structure in latent factor models. In such cases, we expect computational challenges that may require the development of suitable algorithms for approximate posterior inference. Second, one can explore trait allocations (Campbell et al., 2018), a generalization of feature models in which an expression level is also recorded for each feature. We expect that our theorems extend to the trait allocation setting by virtue of our Palm-calculus-based framework and by adopting the spike-and-slab formulation of trait allocations in James (2017). Finally, we plan to explore more complex models for partially exchangeable data, both generalizing the hierarchical beta process of Thibaux and Jordan (2007) (further developed by James et al., 2024) and proposing alternative prior distributions inspired by the rich literature on partially exchangeable priors for species sampling models. For instance, one could develop analogous of nested (Rodriguez et al., 2008), additive (Lijoi et al., 2014) and compound (Griffin and Leisen, 2017) processes, as well as draw inspiration from the more general constructions in Ascolani et al. (2024); Franzolini et al. (2023); Beraha and Griffin (2023) to design new models in the feature setting. Work on these problems and related ones is ongoing.

# Appendix

## Organization of the Appendix

The Appendix is structured as follows. Section 4.A presents useful results on extended feature allocation models, which are referenced in Remark 4.1 and throughout the chapter. Section 4.B provides additional results on the class of mixed binomial processes. Section 4.C recalls a key formula for working with Palm distributions, namely the Campbell-Little-Mecke (CLM) formula, as well as a fundamental characterization result for mixed Poisson and mixed binomial processes, as stated in (Kallenberg, 1973, Theorem 5.3). Section 4.D contains the proofs of Theorems 4.1 and 4.2, which establish the full Bayesian analysis presented in Section 4.2.1. Section 4.E includes proofs for all results related to the sufficientness postulates discussed in Section 4.3. Section 4.F provides the proofs for Section 4.4, where specific examples are analyzed. Finally, Section 4.G offers additional details on the synthetic experiment examined in Section 4.5.

To facilitate the reading of the Appendix, we recall the *extended* feature allocation model for the exchangeable sequence of observations $(Z_i)_{i\geq 1}$, presented in model (4.3) of the main text. In particular, we consider $Z_i = \sum_{j\geq 1} \tilde{A}_{ij}\delta_{\tilde{X}_j}$, for $i \geq 1$, and the statistical model is given by

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \text{BP}(\tilde{\mu}), \\ \tilde{\mu} &\sim \mathscr{Q}, \end{aligned} \tag{4.14}$$

where $\mathscr{Q}$ denotes the law of the random measure $\tilde{\mu}$, which is a functional of the point process $\Psi = \sum_{j\geq 1} \delta_{(\tilde{X}_j, \tilde{q}_j)}$ on $\mathbb{X} \times (0,1]$. Specifically, $\tilde{\mu}$ is defined as

$$\tilde{\mu}(B) = \int_{\mathbb{X}\times(0,1]} s\mathbb{1}_B(x)\Psi(\mathrm{d}x\,\mathrm{d}s), \quad B \in \mathscr{X}. \tag{4.15}$$

## 4.A Some useful results on extended feature allocation models

The following proposition formally states the necessary condition, discussed in Remark 4.1, on the mean measure of $\Psi$ to ensure $Z_i(\mathbb{X}) < \infty$ a.s. for any generic element $Z_i$ in the sequence $(Z_i)_{i\geq 1}$ defined in (4.14).

**Proposition 4.3.** *Let $Z_i$ be the generic element of the sequence $(Z_i)_{i\geq 1}$ defined in* (4.14)*, where $\tilde{\mu}$ is the functional of a point process $\Psi$ defined via* (4.15)*. Assume that $\Psi$ has infinite points a.s. and let $M_\Psi$ denote the mean measure of $\Psi$. If $Z_i(\mathbb{X}) < \infty$ a.s., then*

$$\int_{\mathbb{X}\times(0,1]} (1-s) M_\Psi(\mathrm{d}x\,\mathrm{d}s) = \infty.$$

*Proof.* We proceed by proving that $\int_{\mathbb{X}\times(0,1]}(1-s)M_\Psi(\mathrm{d}x\,\mathrm{d}s) < \infty$ implies $\mathsf{P}(Z_i(\mathbb{X}) < \infty) = 0$. Indeed, define the sequence of events $V_j = \{\tilde{A}_{ij} = 0\} \in \mathscr{A}$, $j \geq 1$ and observe that the event $Z_i(\mathbb{X}) < \infty$ coincides with the event $\liminf V_j$. We now prove that $\mathsf{P}(\liminf V_j) \leq \mathsf{P}(\limsup V_j) = 0$. Consider

$$\sum_{j\geq 1}\mathsf{P}(V_j) = \sum_{j\geq 1}\mathsf{P}(\tilde{A}_{ij} = 0) = \sum_{j\geq 1}\mathsf{E}(1-\tilde{q}_j) = \mathsf{E}\sum_{j\geq 1}(1-\tilde{q}_j),$$

and define the process $\mathcal{T} = \sum_{j\geq 1}\delta_{(\tilde{X}_j,\pi_j)}$ on $\mathbb{X}\times[0,1)$, where $\pi_j = 1-\tilde{q}_j$. We have that

$$\sum_{j\geq 1}\mathsf{P}(V_j) = \mathsf{E}\sum_{j\geq 1}\pi_j = \mathsf{E}\left\{\int_{\mathbb{X}\times[0,1)} t\mathcal{T}(\mathrm{d}x\,\mathrm{d}t)\right\} = \int_{\mathbb{X}\times[0,1)} tM_{\mathcal{T}}(\mathrm{d}x\,\mathrm{d}t), \tag{4.16}$$

where the last equality follows from Campbell averaging formula. Now, defining the function $g : \mathbb{X}\times(0,1] \to \mathbb{X}\times[0,1)$ as $g(x,s) = (x, 1-s)$, we have that $\mathcal{T} = \Psi \circ g^{-1}$ is the image of $\Psi$ by $g$. Consequently, $M_{\mathcal{T}}(\mathrm{d}x\,\mathrm{d}t) = M_\Psi(g^{-1}(\mathrm{d}x, \mathrm{d}t)) = M_\Psi(\mathrm{d}x\,\mathrm{d}s)$, where $s = 1-t$. Thus, it follows that

$$\int_{\mathbb{X}\times[0,1)} tM_{\mathcal{T}}(\mathrm{d}x\,\mathrm{d}t) = \int_{\mathbb{X}\times(0,1]}(1-s)M_\Psi(\mathrm{d}x\,\mathrm{d}s) < \infty, \tag{4.17}$$

where the inequality holds by hypothesis. Therefore, from (4.16) and (4.17), we have that $\sum_{j\geq 1}\mathsf{P}(V_j) < \infty$. By applying the Borel-Cantelli lemma, we obtain $\mathsf{P}(\limsup V_j) = 0$ and the proof is complete. □

The following proposition establishes the necessary and sufficient condition, also discussed in Remark 4.1, on the mean measure of $\Psi$ to ensure that $Z_i(\mathbb{X}) < \infty$ a.s., under the assumption that $\Psi$ is a Poisson or a mixed Poisson process. In particular, when $\Psi$ follows either of these processes, the necessary condition given in Proposition 4.3 is also sufficient.

**Proposition 4.4.** *Let $Z_i$ be the generic element of the sequence $(Z_i)_{i\geq 1}$ defined in (4.14), where $\tilde{\mu}$ is the functional of a point process $\Psi$ defined via (4.15). Assume $\Psi$ is a Poisson process with (infinite) mean measure $\nu$ or a mixed Poisson process $\mathrm{MP}(\nu, f_\gamma)$. Then, $Z_i(\mathbb{X}) < \infty$ a.s. if and only if $\int_{\mathbb{X}\times(0,1]} s\nu(\mathrm{d}x\,\mathrm{d}s) < \infty$.*

*Proof.* We start by proving the statement for $\Psi$ distributed as a Poisson process with mean measure $\nu$. As discussed in Remark 4.1, the condition $\int_{\mathbb{X}\times(0,1]} sM_\Psi(\mathrm{d}x\,\mathrm{d}s) < \infty$ is sufficient for $Z_i(\mathbb{X}) < \infty$ a.s.. Under the Poisson assumption for $\Psi$, this sufficient condition writes as $\int_{\mathbb{X}\times(0,1]} s\nu(\mathrm{d}x\,\mathrm{d}s) < \infty$. We are left to show the inverse implication. Consider the Laplace transform of the random variable $Z_i(\mathbb{X})$ evaluated in 1,

$$\begin{aligned}
\mathsf{E}\left\{e^{-Z_i(\mathbb{X})}\right\} &= \mathsf{E}\left[\mathsf{E}\left\{e^{-Z_i(\mathbb{X})} \mid \tilde{\mu}\right\}\right] = \mathsf{E}\left[\prod_{j\geq 1}\left\{1-\tilde{q}_j(1-e^{-1})\right\}\right] \\
&= \mathsf{E}\left(\exp\left[\sum_{j\geq 1}\log\left\{1-\tilde{q}_j(1-e^{-1})\right\}\right]\right) \\
&= \exp\left\{-(1-e^{-1})\int_{\mathbb{X}\times(0,1]} s\nu(\mathrm{d}x\,\mathrm{d}s)\right\}.
\end{aligned}$$

Therefore, if $\int_{\mathbb{X}\times(0,1]} s\nu(\mathrm{d}x\,\mathrm{d}s) = \infty$, then $\mathsf{E}\left\{e^{-Z_i(\mathbb{X})}\right\} = 0$ and $Z_i(\mathbb{X}) = \infty$ a.s. and the proof is complete.

To prove the result when $\Psi$ is a mixed Poisson process $\mathrm{MP}(\nu, f_\gamma)$, we leverage the result just established for the Poisson process. In particular, $Z_i(\mathbb{X}) < \infty$ a.s. is equivalent to

$$1 = \mathsf{P}\left(Z_i(\mathbb{X}) < \infty\right) = \mathsf{E}\left\{\mathsf{P}\left(Z_i(\mathbb{X}) < \infty \,|\, \gamma\right)\right\},$$

which holds true if and only if $\mathsf{P}\left(Z_i(\mathbb{X}) < \infty \,|\, \gamma\right) = 1$ a.s.. Conditionally to $\gamma$, the process $\Psi$ is a Poisson process with mean measure $\gamma\nu$. Thus, leveraging on the result just shown for the Poisson case, it holds that $\mathsf{P}\left(Z_i(\mathbb{X}) < \infty \,|\, \gamma\right) = 1$ if and only if $\nu$ satisfies the condition stated for a Poisson process, namely $\int_{\mathbb{X}\times(0,1]} s\nu(\mathrm{d}x\,\mathrm{d}s) < \infty$. □

The next proposition examines notable consequences of satisfying the condition $Z_i(\mathbb{X}) < \infty$ a.s. under the assumption that $\Psi$ is a Poisson process.

**Proposition 4.5.** *Let $Z_i$ be the generic element of the sequence $(Z_i)_{i\geq 1}$ defined in* (4.14)*, where $\tilde{\mu}$ is the functional of a point process $\Psi$ defined via* (4.15) *and $\Psi$ is a Poisson process with (infinite) mean measure $\nu$. If $Z_i(\mathbb{X}) < \infty$ a.s., then:*

*(i) the $k$-th factorial moment measure $M_\Psi^{(k)} = \nu^k$ is $\sigma$-finite, for any $k$;*

*(ii) the following disintegration holds: $\nu(\mathrm{d}x\,\mathrm{d}s) = \kappa(\mathrm{d}x \,|\, s)A_0(\mathrm{d}s)$, where $\kappa$ is a probability kernel from $(0,1]$ to $\mathbb{X}$ and $A_0(\mathrm{d}s) := \nu(\mathbb{X}\times\mathrm{d}s)$ is $\sigma$-finite. Consequently, $\Psi$ can be seen as an independently marked point process with Poisson ground process on $(0,1]$ with mean measure $A_0$ and marks on $\mathbb{X}$ with mark probability kernel $\kappa$.*

*If the mean measure $\nu$ is finite, points (i) and (ii) always hold with additional trivial simplifications.*

*Proof.* To prove point (i), we observe that $\nu(\mathbb{X}\times(\epsilon,1]) < \infty$, for any $\epsilon > 0$. Indeed, when $\Psi$ is a Poisson process with mean measure $\nu$, the condition $Z_i(\mathbb{X}) < \infty$ a.s. is equivalent to $\int_{\mathbb{X}\times(0,1]} s\nu(\mathrm{d}x\,\mathrm{d}s) < \infty$ (Proposition 4.4). Therefore, for any $\epsilon > 0$,

$$\nu(\mathbb{X}\times(\epsilon,1]) = \int_{\mathbb{X}\times(\epsilon,1]} \nu(\mathrm{d}x\,\mathrm{d}s) \leq \epsilon^{-1}\int_{\mathbb{X}\times(\epsilon,1]} s\nu(\mathrm{d}x\,\mathrm{d}s) < \infty.$$

Thus, it follows that $\nu$ is $\sigma$-finite, as well as any $\nu^k$.

We now focus on point (ii). Consider the projected measure $\nu(\mathbb{X}\times\mathrm{d}s) =: A_0(\mathrm{d}s)$ on $(0,1]$. From point (i), $A_0$ is a $\sigma$-finite measure. Then, thanks to (Kallenberg, 2021, Theorem 3.4), the following disintegration holds: $\nu(\mathrm{d}x\,\mathrm{d}s) = \kappa(\mathrm{d}x \,|\, s)A_0(\mathrm{d}s)$, where $\kappa$ is a probability kernel from $(0,1]$ to $\mathbb{X}$. □

## 4.B Auxiliary results on mixed binomial processes

This section presents three distributional results related to mixed binomial processes. The first proposition establishes that the $k$-th reduced Palm version of any mixed binomial process remains a mixed binomial process. This result is crucial for proving Lemma 4.2 (see Section 4.E) and Corollary 4.3 (see Section 4.F).

**Proposition 4.6** (Palm distribution of mixed binomial processes)**.** *Let $\Phi$ be a mixed binomial process* $\mathrm{MB}(\nu, q_M)$ *defined on* $\mathbb{X}$*. Then, the reduced Palm version of $\Phi$ at $x$, denoted with $\Phi^!_x$, is a mixed binomial process* $\mathrm{MB}(\nu, q_{\tilde{M}})$ *where*

$$q_{\tilde{M}}(m) = \frac{(m+1)}{\mathsf{E}(M)} q_M(m+1).$$

*Similarly, the $k$-th reduced Palm version of $\Phi$ at $\boldsymbol{x} = (x_1, \ldots, x_k)$ (distinct points), denoted with $\Phi^!_{\boldsymbol{x}}$, is a mixed binomial process* $\mathrm{MB}(\nu, q^{(k)}_{\tilde{M}})$ *where*

$$q^{(k)}_{\tilde{M}}(m) = \frac{(m+k)!}{\mathsf{E}\{M^{(k)}\} m!} q_M(m+k).$$

*Proof.* Let us focus on the reduced Palm distribution of order $k = 1$. Let $\mathcal{L}_\Phi$ be the Laplace functional of $\Phi$. By (Baccelli et al., 2020, Proposition 3.2.1), for any measurable functions $f, g : \mathbb{X} \to \mathbb{R}_+$, the Palm distribution of a point process satisfies

$$\frac{\partial}{\partial t} \mathcal{L}_\Phi(f + tg)\big|_{t=0} = -\int_{\mathbb{X}} g(x) \mathcal{L}_{\Phi_x}(f) M_\Phi(\mathrm{d}x).$$

Since $\Phi$ is a mixed binomial process $\mathrm{MB}(\nu, q_M)$, its Laplace functional writes as

$$\mathcal{L}_\Phi(f) = \mathsf{E}\left[\mathsf{E}\left\{e^{-f(X)}\right\}^M\right], \tag{4.18}$$

where $M$ has probability mass function $q_M$ and $X$ has law $\nu$, so that

$$\frac{\partial}{\partial t} \mathcal{L}_\Phi(f + tg)\big|_{t=0} = -\int_{\mathbb{X}} g(x) e^{-f(x)} \left[\sum_{m \geq 1} m q_M(m) \mathsf{E}\left\{e^{-f(X)}\right\}^{m-1}\right] \nu(\mathrm{d}x).$$

Multiplying and dividing by $\mathsf{E}(M)$, and thanks to a change of index $k = m - 1$ in the inner summation, we recognize the following expression

$$\frac{\partial}{\partial t} \mathcal{L}_\Phi(f + tg)\big|_{t=0} = -\int_{\mathbb{X}} g(x) e^{-f(x)} \left[\sum_{k \geq 0} \frac{(k+1) q_M(k+1)}{\mathsf{E}(M)} \mathsf{E}\left\{e^{-f(X)}\right\}^k\right] M_\Phi(\mathrm{d}x).$$

Hence, $\Phi_x = \delta_x + \Phi^!_x$ where $\Phi^!_x$ is distributed as in the statement.

The proof for the Palm distribution of order $k$ follows by induction, by using the Palm algebra property (Baccelli et al., 2020, Proposition 3.3.9), i.e., $\Phi^!_{(x_1,x_2)} \stackrel{\mathrm{d}}{=} \left(\Phi^!_{x_1}\right)^!_{x_2}$, for $M^{(2)}_\Phi$-a.a. $(x_1, x_2)$. □

The second proposition examines the effect of thinning a mixed binomial process and establishes that its probability law remains that of a mixed binomial process. This result is applied in the proof of Corollary 4.3; see the mixed binomial case of Section 4.F.

**Proposition 4.7** (Thinning of mixed binomial processes)**.** *Let $\Phi$ be a mixed binomial process* $\mathrm{MB}(\nu, q_M)$ *defined on* $\mathbb{X}$*. Let the retention probability $p : \mathbb{X} \to [0, 1]$ be a measurable function. Then the thinning of $\Phi$ by $p$, denoted by $\Phi_p$, is a mixed binomial process* $\mathrm{MB}(\nu_p, q_{M_p})$*, where $\nu_p(\mathrm{d}x) = p(x)\nu(\mathrm{d}x)/c_p$ is a probability distribution and*

$$q_{M_p}(m) = \sum_{z \geq m} \binom{z}{m} q_M(z) c_p^m (1 - c_p)^{z-m}.$$

*Proof.* Given a point process $\Phi$, the thinned process $\Phi_p$ is characterized by the Laplace functional described in (Baccelli et al., 2020, Proposition 2.2.6), that is

$$\mathcal{L}_{\Phi_p}(f) = \mathcal{L}_{\Phi}\left[-\log\left\{p(x)e^{-f(x)} + 1 - p(x)\right\}\right].$$

Under the assumption that $\Phi$ is a mixed binomial process, its Laplace functional is expressed in (4.18). Consequently,

$$\mathcal{L}_{\Phi_p}(f) = \mathsf{E}\left\{\left(1 + \mathsf{E}\left[p(X)\left\{e^{-f(X)} - 1\right\}\right]\right)^M\right\},$$

where $M$ has probability mass function $q_M$ and $X$ has law $\nu$. Let $Z$ be distributed according to $\nu_p$, where $\nu_p$ is as in the statement. It follows that

$$\mathsf{E}\left[p(X)\left\{e^{-f(X)} - 1\right\}\right] = c_p\mathsf{E}\left\{e^{-f(Z)}\right\} - c_p.$$

An application of the binomial theorem leads to

$$\mathcal{L}_{\Phi_p}(f) = \mathsf{E}\left[\sum_{j=0}^{M}\binom{M}{j}{c_p}^j\mathsf{E}\left\{e^{-f(Z)}\right\}^j(1-c_p)^{M-j}\right]$$

and the result follows from Fubini's theorem. □

Finally, we highlight a simple yet useful result stating that mixed Poisson processes $\mathrm{MP}(\nu, f_\gamma)$, where $\nu$ is finite, can be viewed as mixed binomial processes.

**Lemma 4.3.** *Let $\Phi \sim \mathrm{MP}(\nu, f_\gamma)$, where $\nu$ is a finite measure on $\mathbb{X}$ and some $f_\gamma$. Then $\Phi \sim \mathrm{MB}(\nu, q_M)$ for some $q_M$. In particular, letting $\varrho(t) := \mathsf{E}(e^{-t\gamma})$ and $\varphi(\nu(\mathbb{X})(1-z)) := \mathsf{E}(z^M)$, it holds that $\varrho = \varphi$ on $[0, \nu(\mathbb{X})]$.*

*Proof.* Assume that $\Phi \sim \mathrm{MP}(\nu, f_\gamma)$. Define a mixed binomial process

$$\tilde{\Phi} \,|\, M = \sum_{j=1}^{M}\delta_{\tilde{X}_j},$$

with $M \,|\, \gamma \sim \mathrm{Poi}(\gamma\nu(\mathbb{X}))$ and $\gamma \sim f_\gamma$. Moreover, consider $\tilde{X}_1, \ldots, \tilde{X}_M \,|\, M \overset{\text{iid}}{\sim} \nu(\cdot)/\nu(\mathbb{X})$. The probability-generating function of $M$ can be written as

$$\mathcal{G}_M(z) = \mathsf{E}(z^M) = \mathsf{E}[\exp\{-\gamma\nu(\mathbb{X})(1-z)\}]$$

and the Laplace transform of $f(\tilde{X}_1)$ is given by

$$\mathcal{L}_{f(\tilde{X}_1)}(t) = \int_{\mathbb{X}}\exp\{-tf(x)\}\nu(\mathrm{d}x)/\nu(\mathbb{X}).$$

Consequently, the Laplace functional of $\tilde{\Phi}$ equals

$$\mathcal{L}_{\tilde{\Phi}}(f) = \mathcal{G}_M\left(\mathcal{L}_{f(\tilde{X}_1)}(1)\right) = \mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}}(1-e^{-f(x)})\gamma\nu(\mathrm{d}x)\right\}\right],$$

therefore $\tilde{\Phi} \sim \mathrm{MP}(\nu, f_\gamma)$ and in particular $\Phi \overset{\mathrm{d}}{=} \tilde{\Phi}$. By definition, $\Phi \overset{\mathrm{d}}{=} \tilde{\Phi} \sim \mathrm{MB}(\nu, q_M)$, with $\varphi(\nu(\mathbb{X})(1-z)) := \mathsf{E}(z^M) = \mathsf{E}[\exp\{-\gamma\nu(\mathbb{X})(1-z)\}]$, thus $\varphi(t) = \mathsf{E}(e^{-t\gamma}) =: \varrho(t)$ for $t \in [0, \nu(\mathbb{X})]$. □

## 4.C Key results from point process theory and Palm distributions

First, we recall the primary technical tool used in the proof of Theorems 4.1 and 4.2, namely the Campbell-Little-Mecke formula. This formula can be viewed as an extension of Fubini's theorem to the case when both expectation and the integration are taken with respect to a point process.

**Lemma 4.4** (Campbell-Little-Mecke (CLM) formula)**.** *Let $\Phi$ be a point process over $(\mathbb{X}, \mathscr{X})$ such that $M_\Phi^{(k)}$ is $\sigma$-finite. For all measurable $f : \mathbb{X}^k \times \mathbb{M}_\mathbb{X} \to \mathbb{R}_+$, it holds*

$$\mathsf{E}\left\{\int_{\mathbb{X}^k} f\left(\boldsymbol{x}, \Phi - \sum_{j=1}^{k} \delta_{x_j}\right) \Phi^{(k)}(\mathrm{d}\boldsymbol{x})\right\} = \int_{\mathbb{X}^k} \mathsf{E}\left\{f(\boldsymbol{x}, \Phi^!_{\boldsymbol{x}})\right\} M_\Phi^{(k)}(\mathrm{d}\boldsymbol{x}).$$

Next, we present a key characterization result (Kallenberg, 1973, Theorem 5.3) for mixed Poisson and mixed binomial processes, which relies on properties of the reduced Palm distributions. This result will be crucial for the proof of Lemma 4.2. However, it is important to note that in Kallenberg's theorem, a point process is defined as a random boundedly finite counting measure, whereas we consider a point process as a random locally finite counting measure. This distinction, while subtle, is crucial. Working with boundedly finite measures excludes Poisson processes with infinite activity. Specifically, in infinite-activity Poisson processes considered in the literature, $\Psi(B \times (0, r]) = \infty$ for any bounded Borel set $B \subset \mathbb{X}$, meaning that $\Psi$ is not a boundedly finite measure. Therefore, we verify that the next result remains valid under our definition of a point process.

**Lemma 4.5** (Theorem 5.3 in Kallenberg (1973))**.** *Let $\Phi$ be a point process with locally finite mean measure $M_\Phi$. Then the following assertions are equivalent.*

*(i) The distribution of $\Phi^!_x$ is independent of $x$, for $M_\Phi$-a.a. $x$.*

*(ii) $\Phi$ is either a mixed Poisson or mixed binomial process.*

*Proof.* We proceed by demonstrating the validity of the result in two steps, by separately proving the two implications.
*Proof of (ii) $\implies$ (i).* We start by remarking that if $\Phi$ is either a mixed Poisson or a mixed binomial process, then $\Phi^!_x$ is independent of $x \in \mathbb{X}$, for $M_\Phi$-a.a. $x$. Assume $\Phi$ to be a mixed Poisson, i.e., $\Phi \,|\, \gamma \sim \mathrm{PP}(\gamma\nu)$ and $\gamma \sim f_\gamma$ is an almost surely positive random variable. Then, by (Baccelli et al., 2020, Lemma 2.2.27), its Laplace functional equals

$$\mathcal{L}_\Phi(f) = \mathsf{E}\left[\mathsf{E}\left\{\mathcal{L}_{\Phi_\gamma}(f) \,|\, \gamma\right\}\right] = \mathsf{E}\left(\exp\left[-\int_\mathbb{X} \left\{1 - e^{-f(x)}\right\} \gamma\nu(\mathrm{d}x)\right]\right).$$

By (Baccelli et al., 2020, Proposition 3.2.1), the Palm distribution of $\Phi$ satisfies

$$\frac{\partial}{\partial t}\mathcal{L}_\Phi(f + tg)\big|_{t=0} = -\int_\mathbb{X} g(x)\mathcal{L}_{\Phi_x}(f) M_\Phi(\mathrm{d}x). \tag{4.19}$$

We proceed by computing the left-hand side of (4.19), under the assumption that $\Phi$ is a mixed Poisson process. First, note that $M_\Phi(B) = \mathsf{E}\left[\mathsf{E}\left\{\Phi(B) \,|\, \gamma\right\}\right] = \nu(B)\mathsf{E}(\gamma)$, for $B \in \mathscr{X}$. Hence, one has

$$\frac{\partial}{\partial t}\mathcal{L}_\Phi(f + tg)\big|_{t=0} = \mathsf{E}\left(\mathsf{E}\left[h'(0)\exp\{h(0)\} \,|\, \gamma\right]\right),$$

where $h(t) := -\int_{\mathbb{X}}(1-e^{-f(x)-tg(x)})\gamma\nu(\mathrm{d}x)$, and thus $h'(t) := -\int_{\mathbb{X}} e^{-f(x)-tg(x)}g(x)\gamma\nu(\mathrm{d}x)$. It follows that

$$\begin{aligned}
&\frac{\partial}{\partial t}\mathcal{L}_{\Phi}(f+tg)\big|_{t=0} \\
&\quad= \mathsf{E}\left[-\int_{\mathbb{X}} e^{-f(x)}g(x)\gamma\nu(\mathrm{d}x)\cdot\exp\left\{-\int_{\mathbb{X}}(1-e^{-f(x)})\gamma\nu(\mathrm{d}x)\right\}\right] \\
&\quad= -\int_{\mathbb{R}_+}\int_{\mathbb{X}} e^{-f(x)}g(x)\exp\left\{-\int_{\mathbb{X}}(1-e^{-f(y)})\gamma\nu(\mathrm{d}y)\right\}\gamma\nu(\mathrm{d}x)f_{\gamma}(\mathrm{d}\gamma) \\
&\quad= -\int_{\mathbb{X}} g(x)\left[e^{-f(x)}\int_{\mathbb{R}_+}\exp\left\{-\int_{\mathbb{X}}(1-e^{-f(y)})\gamma\nu(\mathrm{d}y)\right\}\frac{\gamma}{\mathsf{E}(\gamma)}f_{\gamma}(\mathrm{d}\gamma)\right]M_{\Phi}(\mathrm{d}x).
\end{aligned}$$

Therefore, by comparing the previous expression with (4.19), we deduce that, for $M_{\Phi}$-a.a. $x\in\mathbb{X}$,

$$\mathcal{L}_{\Phi_x}(f) = e^{-f(x)}\int_{\mathbb{R}_+}\exp\left\{-\int_{\mathbb{X}}(1-e^{-f(y)})\gamma\nu(\mathrm{d}y)\right\}\frac{\gamma}{\mathsf{E}(\gamma)}f_{\gamma}(\mathrm{d}\gamma)$$

and the Laplace functional of the Palm version is

$$\mathcal{L}_{\Phi^!_x}(f) = \int_{\mathbb{R}_+}\exp\left\{-\int_{\mathbb{X}}(1-e^{-f(y)})\gamma\nu(\mathrm{d}y)\right\}\frac{\gamma}{\mathsf{E}(\gamma)}f_{\gamma}(\mathrm{d}\gamma),$$

which does not depend on $x$. In particular, it turns out that $\Phi^!_x$ is a mixed Poisson process such that $\Phi^!_x\,|\,\tilde{\gamma}\sim\mathrm{PP}(\tilde{\gamma}\nu)$ and $\tilde{\gamma}\sim f_{\tilde{\gamma}}$ with $f_{\tilde{\gamma}}(\mathrm{d}\gamma)\propto\gamma f_{\gamma}(\mathrm{d}\gamma)$.

If instead $\Phi$ is a mixed binomial process, $\Phi^!_x$ is described by Proposition 4.6: it is still a mixed binomial process with law independent of $x$, for $M_{\Phi}$-a.a. $x\in\mathbb{X}$.

*Proof of (i) $\implies$ (ii).* Conversely, to prove the opposite implication, we need some preliminary lemmas, which are of independent interest. We start by recalling (Kallenberg, 1973, Lemma 5.1). Given the point process $\Phi$ and the set $C\in\mathscr{X}$, we indicate with $\Phi^C$ the restriction of $\Phi$ on $C$, i.e., $\Phi^C(B)=\Phi(B\cap C)$, for any $B\in\mathscr{X}$. Moreover, let $\mathbb{N}$ denote the set of natural numbers $\{1,2,\ldots\}$.

**Lemma 4.6.** *If $\Phi$ is a point process with locally finite mean measure $M_{\Phi}$, then*

$$\mathsf{P}(\Phi\in L\,|\,\Phi(C)=n,\eta\in\mathrm{d}x) = \mathsf{P}(\Phi_x\in L\,|\,\Phi_x(C)=n),$$

*for $L\in\mathcal{M}_{\mathbb{X}}, C\in\mathscr{X}, n\in\mathbb{N}$, and $\lambda_{C,n}$-a.a. $x\in C$, where $\eta$ is the position of a randomly chosen atom of $\Phi^C$ and*

$$\lambda_{C,n}(\mathrm{d}x) = \mathsf{E}\{\Phi(\mathrm{d}x)\,|\,\Phi(C)=n\}\mathsf{P}(\Phi(C)=n) = \mathsf{P}(\Phi_x(C)=n)M_{\Phi}(\mathrm{d}x).$$

*Proof.* Preliminarily, we need to remark an alternative and limiting characterization of Palm versions in terms of the so-called *local Palm probabilities*, denoted by $\mathrm{P}^x$ and defined on $(\Omega,\mathscr{A})$ (see Baccelli et al., 2020). Specifically, for $M_{\Phi}$-a.a. $x\in\mathbb{X}$, the following definition is given, for $A\in\mathscr{A}$,

$$\mathrm{P}^x(A) = \frac{\mathsf{E}(\Phi(\mathrm{d}x)\mathbb{1}_A)}{M_{\Phi}(\mathrm{d}x)}.$$

Define the family of processes $\{\Phi^{loc}_x\}_{x\in\mathbb{X}}$ whose laws are given as follows: $\mathsf{P}(\Phi^{loc}_x\in L) = \mathrm{P}^x(\Phi^{-1}(L))$, for any $L\in\mathcal{M}_{\mathbb{X}}$. Following Baccelli et al. (2020), we observe that $\Phi^{loc}_x\overset{\mathrm{d}}{=}\Phi_x$, thus the Palm versions of $\Phi$ can be characterized as follows

$$\mathsf{P}(\Phi_x\in L) = \frac{\mathsf{E}(\Phi(\mathrm{d}x)\,|\,\Phi\in L)\mathsf{P}(\Phi\in L)}{M_{\Phi}(\mathrm{d}x)}. \tag{4.20}$$

To prove the lemma, consider $x \in C$, then

$$\begin{aligned}\mathsf{P}(\Phi \in L,&\Phi(C) = n, \eta \in \mathrm{d}x) = \mathsf{P}(\eta \in \mathrm{d}x \,|\, \Phi \in L, \Phi(C) = n)\mathsf{P}(\Phi \in L, \Phi(C) = n) \\ &= \mathsf{E}(\delta_\eta(\mathrm{d}x) \,|\, \Phi \in L, \Phi(C) = n)\mathsf{P}(\Phi \in L, \Phi(C) = n) \\ &= \frac{1}{n}\mathsf{E}(\Phi(\mathrm{d}x) \,|\, \Phi \in L, \Phi(C) = n)\mathsf{P}(\Phi \in L, \Phi(C) = n),\end{aligned} \tag{4.21}$$

where the last equality holds since $\eta$ is a randomly chosen atom of $\Phi$. Moreover, applying (4.20) to the event $L' := \{\tilde{\mu} \in \mathbb{M}_{\mathbb{X}} : \tilde{\mu} \in L, \tilde{\mu}(C) = n\}$, we get

$$\mathsf{E}(\Phi(\mathrm{d}x) \,|\, \Phi \in L, \Phi(C) = n)\mathsf{P}(\Phi \in L, \Phi(C) = n) = \mathsf{P}(\Phi_x \in L, \Phi_x(C) = n)M_\Phi(\mathrm{d}x).$$

Consequently, plugging the last expression into (4.21), it follows that

$$\mathsf{P}(\Phi \in L, \Phi(C) = n, \eta \in \mathrm{d}x) = \frac{1}{n}\mathsf{P}(\Phi_x \in L, \Phi_x(C) = n)M_\Phi(\mathrm{d}x) \tag{4.22}$$

and, in particular, choosing $L = \mathbb{M}_{\mathbb{X}}$,

$$\mathsf{P}(\Phi(C) = n, \eta \in \mathrm{d}x) = \frac{1}{n}\mathsf{P}(\Phi_x(C) = n)M_\Phi(\mathrm{d}x). \tag{4.23}$$

Finally,

$$\begin{aligned}\mathsf{P}(\Phi \in L \,|\, \Phi(C) = n, \eta \in \mathrm{d}x) &= \frac{\mathsf{P}(\Phi \in L, \Phi(C) = n, \eta \in \mathrm{d}x)}{\mathsf{P}(\Phi(C) = n, \eta \in \mathrm{d}x)} \\ &= \frac{\mathsf{P}(\Phi_x \in L, \Phi_x(C) = n)}{\mathsf{P}(\Phi_x(C) = n)} \\ &= \mathsf{P}(\Phi_x \in L \,|\, \Phi_x(C) = n),\end{aligned}$$

where the second equality follows from (4.22) and (4.23). Clearly, the previous computation holds when $\mathsf{P}(\Phi(C) = n, \eta \in \mathrm{d}x) > 0$, that is $\lambda_{C,n}(\mathrm{d}x) = \mathsf{P}(\Phi_x(C) = n)M_\Phi(\mathrm{d}x) > 0$. □

Secondly, we state another key result which we need for the proof, corresponding to (Kallenberg, 2017, Theorem 3.7).

**Lemma 4.7.** *Let $\Phi$ be a point process on $\mathbb{X}$ and let $C_j \uparrow \mathbb{X}$, $C_j \in \mathscr{X}$ and relatively compact, $j \geq 1$. Then $\Phi$ is either a mixed Poisson or mixed binomial process if and only if $\Phi^{C_j}$ is a mixed binomial process for every $j \geq 1$.*

We can now prove that (i) implies (ii). Assume that $\Phi_x^!$ is independent of $x$, that is $\Phi_x^! = \Phi_x - \delta_x \overset{\mathrm{d}}{=} \xi$, for some point process $\xi$, for $M_\Phi$-a.a. $x \in \mathbb{X}$. Consider $C \in \mathscr{X}$ relatively compact set and $n \in \mathbb{N}$ such that $\mathsf{P}(\xi(C) = n) > 0$. Let $k \in \mathbb{N}$ and $\boldsymbol{B} = (B_1, \ldots, B_k) \in \mathscr{X}^{\otimes k}$ whose components are a partition of $C$. Let $\boldsymbol{z} \in (\mathbb{N} \cup \{0\})^k$ and $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$ with $k-1$ zeros. Let $\eta$ as in Lemma 4.6. Assume, without loss of generality, that $x \in B_1$; for $\lambda_{C,n}$-a.a. $x$, it holds that

$$\begin{aligned}\mathsf{P}((\Phi - \delta_\eta)(\boldsymbol{B}) = \boldsymbol{z} \,|\, \Phi(C) = n, \eta \in \mathrm{d}x) &= \mathsf{P}(\Phi(\boldsymbol{B}) = \boldsymbol{z} + \boldsymbol{e}_1 \,|\, \Phi(C) = n, \eta \in \mathrm{d}x) \\ &= \mathsf{P}(\Phi_x(\boldsymbol{B}) = \boldsymbol{z} + \boldsymbol{e}_1 \,|\, \Phi_x(C) = n) \\ &= \mathsf{P}((\Phi_x - \delta_x)(\boldsymbol{B}) = \boldsymbol{z} \,|\, (\Phi_x - \delta_x)(C) = n - 1) \\ &= \mathsf{P}(\xi(\boldsymbol{B}) = \boldsymbol{z} \,|\, \xi(C) = n - 1),\end{aligned}$$

where the second equality follows from Lemma 4.6 and the last one from the definition of $\xi$. Therefore, the following holds:

$$\begin{aligned}\mathsf{P}((\Phi-\delta_\eta)(\boldsymbol{B})=\boldsymbol{z},\eta\in\mathrm{d}x\,|\,\Phi(C)=n)&\\ =\mathsf{P}((\Phi-\delta_\eta)(\boldsymbol{B})=\boldsymbol{z}\,|\,\Phi(C)=n,\eta\in\mathrm{d}x)\mathsf{P}(\eta\in\mathrm{d}x\,|\,\Phi(C)=n)&\\ =\mathsf{P}(\xi(\boldsymbol{B})=\boldsymbol{z}\,|\,\xi(C)=n-1)\mathsf{P}(\eta\in\mathrm{d}x\,|\,\Phi(C)=n).&\end{aligned}\tag{4.24}$$

Then, define $\boldsymbol{B}=(B_1,\dots,B_k)$ as follows: let $k=\bar{k}+1$, $\bar{k}\leq n$, consider $\bar{k}$ points in $C$, denoted with $\bar{x}_1,\dots,\bar{x}_{\bar{k}}$, and let $B_j=\mathrm{d}\bar{x}_j$, $j\leq\bar{k}$, be a neighborhood of $\bar{x}_j$, such that $B_i\cap B_j=\emptyset$, for any $i,j\leq\bar{k}, i\neq j$. Finally, define $B_k=C\setminus\cup_{j=1}^{\bar{k}}\mathrm{d}\bar{x}_j$ and $\boldsymbol{z}=(z_1,\dots,z_{\bar{k}},0)$ such that $\sum_{j=1}^{\bar{k}}z_j=n-1$. Then, the factorization in (4.24) gives

$$\begin{aligned}&\mathsf{P}\Big(\bigcap_{j=1}^{\bar{k}}(\Phi-\delta_\eta)(\mathrm{d}\bar{x}_j)=z_j,(\Phi-\delta_\eta)(C\setminus\cup_{j=1}^{\bar{k}}\mathrm{d}\bar{x}_j)=0,\eta\in\mathrm{d}x\,|\,\Phi(C)=n\Big)\\ &=\mathsf{P}\Big(\bigcap_{j=1}^{\bar{k}}\xi(\mathrm{d}\bar{x}_j)=z_j,\xi(C\setminus\cup_{j=1}^{\bar{k}}\mathrm{d}\bar{x}_j)=0\,|\,\xi(C)=n-1\Big)\mathsf{P}(\eta\in\mathrm{d}x\,|\,\Phi(C)=n),\end{aligned}$$

from which we conclude that the distribution of $\eta$, conditionally to $\Phi(C)=n$, is independent of the position of the remaining $n-1$ points. Then, we conclude that a randomly chosen point of $\Phi-\delta_\eta$ is conditionally independent of the others, given that $(\Phi-\delta_\eta)(C)=n-1$, by repeating the same argument above for the process $\Phi-\delta_\eta$. Note that such an argument applies to $\Phi-\delta_\eta$ since $(\Phi-\delta_\eta)_x-\delta_x=\Phi_x-\delta_x-\delta_\eta=\xi-\delta_\eta$ which does not depend on $x$. Wrapping up, we have just shown that, conditionally on $\Phi(C)=n$, the $n$ points of $\Phi$ are independent between them. Equivalently, $\Phi^C$ is a binomial process given $\Phi(C)=n$, that is $\Phi^C$ is marginally a mixed binomial process. Then, taking a sequence of sets $C_j\uparrow\mathbb{X}, C_j\in\mathscr{X}$ and relatively compact, $j\geq1$, it holds that $\Phi^{C_j}$ is a mixed binomial process for every $j\geq1$. Finally, by an application of Lemma 4.7, we conclude that $\Phi$ is either a mixed Poisson or mixed binomial process. □

## 4.D Proof of Theorems 4.1 and 4.2

The proof is based on the study of the Laplace functional of $\tilde{\mu}$. We remind that, for any measurable function $f:\mathbb{X}\to\mathbb{R}_+$, the Laplace functional of $\tilde{\mu}$ equals

$$\mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}}f(x)\tilde{\mu}(\mathrm{d}x)\right\}\right]=\mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}\times(0,1]}sf(x)\Psi(\mathrm{d}x\,\mathrm{d}s)\right\}\right].$$

In the following, we will use the shorthand notations $\tilde{\mu}(f)=\int_{\mathbb{X}}f(x)\tilde{\mu}(\mathrm{d}x)$ and $\Psi(sf)=\int_{\mathbb{X}\times(0,1]}sf(x)\Psi(\mathrm{d}x\,\mathrm{d}s)$. Let $\boldsymbol{Z}:=(Z_1,\dots,Z_n)$ and denote by $\boldsymbol{z}$ its realization, with associated likelihood function $L(\boldsymbol{z}\,|\,\tilde{\mu})$. An application of Bayes' theorem entails

$$\mathsf{E}\left\{e^{-\tilde{\mu}(f)}\,|\,\boldsymbol{Z}=\boldsymbol{z}\right\}=\frac{\mathsf{E}\left\{e^{-\Psi(sf)}L(\boldsymbol{z}\,|\,\tilde{\mu})\right\}}{\mathsf{E}\left\{L(\boldsymbol{z}\,|\,\tilde{\mu})\right\}},\tag{4.25}$$

where, at the denominator, we recognize the marginal likelihood.

Let $X_1,\dots,X_k$ be the $k$ distinct features displayed in $\boldsymbol{z}$ and denote by $m_1,\dots,m_k$ the corresponding frequency counts. Let $\Theta=\mathbb{X}\times(0,1]$; arguing as in (James, 2017, Appendix

A), the likelihood function can be shown to be

$$L(\boldsymbol{z} \,|\, \tilde{\mu}) = \int_{\Theta^k} e^{n \int \log(1-t)\Psi(\mathrm{d}z\,\mathrm{d}t)} \prod_{\ell=1}^{k} \frac{s_\ell^{m_\ell}}{(1-s_\ell)^{m_\ell}} \delta_{X_\ell}(x_\ell) \Psi^k(\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{s})$$

where $\Psi^k$ is the $k$-th power of $\Psi$. Note that the term $\prod_{\ell=1}^{k} \delta_{X_\ell}(x_\ell)$ entails that the integrand is zero on sets of the type

$$\{(\boldsymbol{x}, \boldsymbol{s}) \in \Theta^k : \; x_i = x_j \text{ for } i \neq j\}.$$

Therefore we can replace $\Psi^k$ with the $k$-th factorial power $\Psi^{(k)}$. Then, an application of the CLM formula (cf. Lemma 4.4) yields the following expression for the numerator of (4.25),

$$\begin{aligned}
\mathsf{E}\left\{e^{-\Psi(sf)} L(\boldsymbol{z} \,|\, \tilde{\mu})\right\} &= \mathsf{E}\left\{ \int_{\Theta^k} e^{-\Psi(tf - n\log(1-t))} \prod_{\ell=1}^{k} \frac{s_\ell^{m_\ell}}{(1-s_\ell)^{m_\ell}} \delta_{X_\ell}(x_\ell) \Psi^{(k)}(\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{s}) \right\} \\
&= \int_{\Theta^k} \mathsf{E}\left\{ e^{-\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(tf - n\log(1-t))} \right\} \\
&\qquad \times \prod_{\ell=1}^{k} e^{-s_\ell f(x_\ell) + n\log(1-s_\ell)} \frac{s_\ell^{m_\ell}}{(1-s_\ell)^{m_\ell}} \delta_{X_\ell}(x_\ell) M^{(k)}_{\Psi}(\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{s}).
\end{aligned}$$

By assumptions, the disintegration in (4.4) holds true. Moreover, integrating with respect to the $x_\ell$'s, we obtain

$$\begin{aligned}
\mathsf{E}\left\{e^{-\Psi(sf)} L(\boldsymbol{z} \,|\, \tilde{\mu})\right\} &= \int_{(0,1]^k} \mathsf{E}\left\{ e^{-\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(tf - n\log(1-t))} \right\} \\
&\qquad \times \prod_{\ell=1}^{k} e^{-s_\ell f(X_\ell)} s_\ell^{m_\ell} (1-s_\ell)^{n-m_\ell} \rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x}) \tilde{m}^{(k)}_{\xi}(\mathrm{d}\boldsymbol{x}),
\end{aligned} \tag{4.26}$$

where $\boldsymbol{x} = (X_1, \ldots, X_k)$. The expression in Theorem 4.1 follows by setting $f = 0$ in the equation above, which coincides with the marginal likelihood of the data.

As for the proof of Theorem 4.2, we can simply evaluate (4.25), where the numerator equals (4.26), while the denominator coincides with the marginal likelihood, i.e., Equation (4.26) when $f = 0$. Thus, the posterior Laplace functional in (4.25) boils down to

$$\begin{aligned}
&\mathsf{E}\left\{e^{-\tilde{\mu}(f)} \,|\, \boldsymbol{Z} = \boldsymbol{z}\right\} \\
&\quad = \frac{\int_{(0,1]^k} \mathsf{E}\left\{ e^{-\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(tf - n\log(1-t))} \right\} \prod_{\ell=1}^{k} e^{-s_\ell f(X_\ell)} s_\ell^{m_\ell} (1-s_\ell)^{n-m_\ell} \rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x})}{\int_{(0,1]^k} \mathsf{E}\left\{ e^{\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(n\log(1-t))} \right\} \prod_{\ell=1}^{k} s_\ell^{m_\ell} (1-s_\ell)^{n-m_\ell} \rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x})}.
\end{aligned}$$

To conclude the proof of Theorem 4.2, we need to show that the right-hand side above coincides with the Laplace transform of the measure

$$\eta := \sum_{\ell=1}^{k} q_\ell \delta_{X_\ell} + \mu'$$

where the laws of $q_1, \dots, q_k$ and $\mu'$ are as in the statement of Theorem 4.2. This is indeed true, since the Laplace functional of $\eta$ is

$$\begin{aligned}
\mathsf{E}\left\{e^{-\eta(f)}\right\} &= \mathsf{E}\left\{e^{-\sum_{\ell=1}^{k} q_\ell f(X_\ell) - \mu'(f)}\right\} \\
&= \int_{(0,1]^k} \mathsf{E}\left\{e^{-\mu'(f)} \mid \boldsymbol{q} = \boldsymbol{s}\right\} \prod_{\ell=1}^{k} e^{-s_\ell f(X_\ell)} f_{\boldsymbol{q}}(\mathrm{d}\boldsymbol{s}) \\
&= \int_{(0,1]^k} \frac{\mathsf{E}\left\{e^{-\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(tf - n\log(1-t))}\right\}}{\mathsf{E}\left\{e^{\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(n\log(1-t))}\right\}} \\
&\quad \times \frac{\mathsf{E}\left\{e^{\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(n\log(1-t))}\right\} \prod_{\ell=1}^{k} e^{-s_\ell f(X_\ell)} s_\ell^{m_\ell} (1-s_\ell)^{n-m_\ell} \rho^{(k)}(\mathrm{d}\boldsymbol{s} \mid \boldsymbol{x})}{\int_{(0,1]^k} \mathsf{E}\left\{e^{\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(n\log(1-t))}\right\} \prod_{\ell=1}^{k} s_\ell^{m_\ell} (1-s_\ell)^{n-m_\ell} \rho^{(k)}(\mathrm{d}\boldsymbol{s} \mid \boldsymbol{x})}
\end{aligned}$$

which is exactly the Laplace functional of $\tilde{\mu}$ that we have found before, and this concludes the proof.

## 4.E Results and proofs of Section 4.3

### 4.E.1 Proof of Theorem 4.4

It is clear that the predictive distribution of $Z'_{n+1}$ depends on the sampling information as the law of $\mu'$, or equivalently $\Psi'$, in Theorem 4.2 does. Therefore, the statement of Theorem 4.4 is equivalent to saying that $\Psi'$ depends only on $n$ if and only if $\Psi$ is a Poisson process. From (4.6), it is clear that $\Psi'$ depends on $n$ and $\Psi^!_{\boldsymbol{x},\boldsymbol{q}}$, where $\boldsymbol{x}$ are the observed distinct feature labels and $\boldsymbol{q}$ is distributed as in Theorem 4.2. Then, the proof follows from Lemma 4.1 which characterizes the Poisson process as the unique point process for which the reduced Palm kernel $\Psi^!_{\boldsymbol{x},\boldsymbol{q}}$ does not depend on $(\boldsymbol{x}, \boldsymbol{q})$.

### 4.E.2 Proof of Lemma 4.1

If $\Phi$ is a Poisson process, then $\Phi^!_{\boldsymbol{x}} \overset{\mathrm{d}}{=} \Phi^!_{\boldsymbol{y}} \overset{\mathrm{d}}{=} \Phi$ by the multivariate Mecke equation (Last and Penrose, 2017, Theorem 4.4). Thus, this shows that (i) implies (ii).

On the other hand, assume that (ii) holds true, we need to show that $\Phi$ is a Poisson process. To this end, consider $\boldsymbol{x} = (x_1, \dots, x_k)$ and $\boldsymbol{y} = (x_1, \dots, x_k, y)$, i.e., $\boldsymbol{y}$ is defined by adding a single point $y$ to $\boldsymbol{x}$. Then, by the Palm algebra (Baccelli et al., 2020, Proposition 3.3.9) we have

$$\Phi^!_{\boldsymbol{y}} \overset{\mathrm{d}}{=} \left(\Phi^!_{\boldsymbol{x}}\right)^!_y \overset{\mathrm{d}}{=} \Phi^!_{\boldsymbol{x}}$$

where the last equality holds by hypothesis. That is, we have proven that $\Phi^!_{\boldsymbol{x}}$ is a Poisson process using Slivnyak-Mecke (Baccelli et al., 2020, Theorem 3.2.4). Setting $k = 1$, we have that $\Phi^!_x$ is Poisson and by hypothesis, the law does not depend on $x$. Therefore, we conclude that $\Phi$ is a Poisson process since the family of Palm distributions of a point process characterizes its law. See, e.g., (Baccelli et al., 2020, Proposition 3.1.17).

### 4.E.3 Proof of Theorem 4.5

The proof follows by arguing as in the proof of Theorem 4.4 above but invoking Lemma 4.2 instead of Lemma 4.1.

### 4.E.4 Proof of Lemma 4.2

The proof of our lemma requires the use of Lemma 4.5. Using this characterization, proceed as follows to prove Lemma 4.2 . First, (ii) implies (i) by using (Kallenberg, 1973, Theorem 5.3). Indeed, by setting $k = 1$, condition (ii) states that $\Phi^!_x$ does not depend on $x$, which implies $\Phi$ to be either a mixed Poisson or a mixed binomial process (Lemma 4.5). On the other side, (i) implies (ii). Indeed, if $\Phi$ is a mixed Poisson process, then $\Phi^!_{x_1}$ does not depend on $x_1$ (Lemma 4.5) and it is still a mixed Poisson (see the proof of the inverse implication of Lemma 4.5). Conversely, if $\Phi$ is a mixed binomial process, then $\Phi^!_{x_1}$ is still a mixed binomial process which does not depend on $x_1$ (see Proposition 4.6). Therefore, in both cases, $\Phi^!_{(x_1,x_2)} = \left(\Phi^!_{x_1}\right)^!_{x_2}$ does not depend on $(x_1, x_2)$ and it is still either a mixed Poisson or a mixed binomial process. Continuing with this argument, the proof follows for any $k$.

## 4.F Proofs of Section 4.4

### 4.F.1 The Poisson process prior: proof of Corollary 4.1

*Proof of point (i) of Corollary 4.1.* The marginal distribution of $\boldsymbol{Z}$ follows from specializing Theorem 4.1. In particular, letting $\boldsymbol{x} = (X_1, \ldots, X_k)$, we first observe that $\tilde{m}^{(k)}_\xi(\mathrm{d}\boldsymbol{x}) = \prod_{\ell=1}^k G_0(\mathrm{d}X_\ell)$ and $\rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x}) = \prod_{\ell=1}^k \rho(\mathrm{d}s_\ell \,|\, X_\ell)$. Moreover, since $\Psi$ is a Poisson process, by Lemma 4.1, it holds $\Psi^!_{\boldsymbol{x},\boldsymbol{s}} \stackrel{\mathrm{d}}{=} \Psi$. Consequently, the expected value in the marginal expression of Theorem 4.1 equals

$$\mathsf{E}\left\{e^{\int_{\mathbb{X}\times(0,1]} n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(\mathrm{d}z\,\mathrm{d}t)}\right\} = \exp\left[-\int_{\mathbb{X}\times(0,1]}\{1-(1-s)^n\}\rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)\right]. \quad (4.27)$$

It is easy to verify that $\int_{\mathbb{X}\times(0,1]}\{1-(1-s)^n\}\rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$ equals $\varphi_n$, where $\varphi_n$ is defined in (4.9). The resulting marginal distribution recovers the marginal expression found in (James, 2017, Proposition 3.1).

*Proof of point (ii) of Corollary 4.1.* For the posterior distribution of $\tilde{\mu}$, expressed in Theorem 4.2, we need to determine the law of the vector $\boldsymbol{q} = (q_1, \ldots, q_k)$ and the law of $\mu'$, conditionally to $\boldsymbol{q}$. From point (i) of Theorem 4.2 and (4.27), the $q_\ell$'s are independent with marginal laws $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell}\rho(\mathrm{d}s \,|\, X_\ell)$. Moreover, since $\Psi^!_{\boldsymbol{x},\boldsymbol{q}}$ is a Poisson point process with intensity $\rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$, simple algebra applied to (4.7) leads to recognizing

$$\mathsf{E}\left\{e^{-\int f(z)\mu'(\mathrm{d}z)} \,|\, \boldsymbol{q}\right\} = \exp\left\{-\int_{\mathbb{X}\times(0,1]}(1-e^{-sf(x)})(1-s)^n\rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)\right\},$$

that is the Laplace functional of a CRM with Lévy measure $(1-s)^n\rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$, thus $\Psi'$ is a Poisson process with the same mean measure. Further note that $\mu'$ is independent of $\boldsymbol{q}$.

*Proof of point (iii) of Corollary 4.1.* To obtain the predictive distribution of $Z'_{n+1}$, we observed in Theorem 4.3 how $Z'_{n+1}$ is obtained by first thinning the process $\Psi'$ with retention probability $p(x, s) = s$, and then discarding the second component. Since $\Psi'$ is a Poisson process with intensity $(1-s)^n\rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$, the thinned process is still Poisson with intensity $s(1-s)^n\rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$ and the resulting $Z'_{n+1}$ is a Poisson process on $\mathbb{X}$ with intensity $\int_{(0,1]} s(1-s)^n\rho(\mathrm{d}s \,|\, x)G_0(\mathrm{d}x)$.

### 4.F.2 The mixed Poisson process prior: proof of Corollary 4.2

*Proof of point (i) of Corollary 4.2.* The marginal distribution of $\boldsymbol{Z}$ follows from exploiting that $\Psi \,|\, \gamma \sim \mathrm{PP}(\gamma\nu)$, for which the marginal law is given in point (i) of Corollary 4.1, and then integrating out the variable $\gamma$.

*Proof of point (ii) of Corollary 4.2.* For the posterior distribution of $\tilde{\mu}$, expressed in Theorem 4.2, we need to determine the law of the vector $\boldsymbol{q} = (q_1, \ldots, q_k)$ and the law of $\mu'$, conditionally to $\boldsymbol{q}$. To this end, it is convenient to exploit the disintegration of the law of $\Psi$ into $\Psi \,|\, \gamma$ and $\gamma$. First, as for the posterior distribution of $\Psi \,|\, \gamma$, since $\Psi \,|\, \gamma \sim \mathrm{PP}(\gamma\nu)$, we resort to point (ii) of Corollary 4.1. Specifically, such a posterior is equal in distribution to $\Psi' + \sum_{\ell=1}^{k} \delta_{(X_\ell, q_\ell)}$, where $\Psi' \,|\, \gamma \sim \mathrm{PP}(\gamma(1-s)^n \nu(\mathrm{d}x\,\mathrm{d}s))$ and $q_\ell \,|\, \gamma$ are independent random variables, further independent of $\Psi' \,|\, \gamma$, with laws $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell}\rho(\mathrm{d}s \,|\, X_\ell)$. Since the laws of $q_\ell \,|\, \gamma$ do not depend on $\gamma$, then the $q_\ell$'s and $\Psi' \,|\, \gamma$ are independent. Second, the posterior distribution of $\gamma$ is obtained from the likelihood in point (i) of Corollary 4.1 and the prior $f_\gamma$, thus $\gamma \,|\, \boldsymbol{Z} \sim f_{\tilde{\gamma}}$, with $f_{\tilde{\gamma}}(\mathrm{d}\gamma) \propto e^{-\gamma\varphi_n}\gamma^k f_\gamma(\mathrm{d}\gamma)$. The thesis in point (ii) follows.

*Proof of point (iii) of Corollary 4.2.* To describe the predictive distribution for $Z'_{n+1}$ we proceed as follows: first consider the thinning of the process $\Psi'$ with retention probability $p(x,s) = s$, and then discard the second component. Since $\Psi'$ is a mixed Poisson process $\mathrm{MP}((1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s), f_{\tilde{\gamma}})$, the thinned process is still a mixed Poisson $\mathrm{MP}(s(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s), f_{\tilde{\gamma}})$ and the resulting $Z'_{n+1}$ is a mixed Poisson process (on $\mathbb{X}$) distributed as $\mathrm{MP}(\int_{(0,1]} s(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s), f_{\tilde{\gamma}})$.

### 4.F.3 The mixed binomial process prior: proof of Corollary 4.3

*Proof of point (i) of Corollary 4.3.* The marginal distribution of $\boldsymbol{Z}$ is recovered from Theorem 4.1 as follows. First, letting $\boldsymbol{x} = (X_1, \ldots, X_k)$, it is straightforward to see that $\tilde{m}_\xi^{(k)}(\mathrm{d}\boldsymbol{x}) = \mathsf{E}(M^{(k)})G_0^k(\mathrm{d}\boldsymbol{x})$ and $\rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x}) = \prod_{\ell=1}^{k}\rho(\mathrm{d}s_\ell \,|\, X_\ell)$ satisfy (4.4). Second, from Proposition 4.6, we have that if $\Psi$ is a mixed binomial process $\mathrm{MB}(\nu, q_M)$, then $\Psi^!_{\boldsymbol{x},\boldsymbol{s}}$ is a mixed binomial process $\mathrm{MB}(\nu, q_{\tilde{M}})$ with $q_{\tilde{M}}(m) = q_M(m+k)(m+k)!/(\mathsf{E}(M^{(k)})m!)$. Consequently, the expected value in the marginal expression of Theorem 4.1 equals

$$\begin{aligned} &\mathsf{E}\left\{e^{\int_{\mathbb{X}\times(0,1]} n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(\mathrm{d}z\,\mathrm{d}t)}\right\} = \mathsf{E}\left[\exp\left\{n\textstyle\sum_{j=1}^{\tilde{M}}\log(1-\tilde{S}_j)\right\}\right] \\ &= \mathsf{E}\left[\left\{\int_{\mathbb{X}\times(0,1]}(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s)\right\}^{\tilde{M}}\right] = \mathcal{G}_{\tilde{M}}(\kappa_n), \end{aligned} \tag{4.28}$$

where $\tilde{M}$ has probability mass function $q_{\tilde{M}}$ and $(\tilde{X}_j, \tilde{S}_j) \overset{\text{iid}}{\sim} \nu$, for $j = 1, \ldots, \tilde{M}$; moreover, $\mathcal{G}_{\tilde{M}}(z) = \mathsf{E}(z^{\tilde{M}})$ is the probability-generating function of $\tilde{M}$ and $\kappa_n = \int_{\mathbb{X}\times(0,1]}(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s)$. It follows that the marginal distribution in Theorem 4.1 boils down to the expression in point (i) of the statement.

*Proof of point (ii) of Corollary 4.3.* For the posterior distribution of $\tilde{\mu}$, expressed in Theorem 4.2, we need to determine the law of the vector $\boldsymbol{q} = (q_1, \ldots, q_k)$ and the law of $\mu'$, conditionally to $\boldsymbol{q}$. From point (i) of Theorem 4.2 and (4.28), the $q_\ell$'s are independent with marginal laws $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell}\rho(\mathrm{d}s \,|\, X_\ell)$, as for the Poisson case. Moreover, from

point (ii) of Theorem 4.2 we have that for any measurable function $g : \mathbb{X} \times (0,1] \to \mathbb{R}_+$,

$$\begin{aligned}
\mathcal{L}_{\Psi'}(g) &= \mathcal{L}_{\Psi^!_{\boldsymbol{x},\boldsymbol{q}}}(g(x,s) - n\log(1-s))/\mathcal{L}_{\Psi^!_{\boldsymbol{x},\boldsymbol{q}}}(-n\log(1-s)) \\
&= \mathcal{G}_{\tilde{M}}\left(\int_{\mathbb{X}\times(0,1]} \exp\left\{-g(x,s) + n\log(1-s)\right\}\nu(\mathrm{d}x\,\mathrm{d}s)\right)/\mathcal{G}_{\tilde{M}}(\kappa_n) \\
&= \mathcal{G}_{\tilde{M}}\left(\int_{\mathbb{X}\times(0,1]} e^{-g(x,s)}(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s)\right)/\mathcal{G}_{\tilde{M}}(\kappa_n) \\
&= \mathcal{G}_{M'}\left(\int_{\mathbb{X}\times(0,1]} e^{-g(x,s)}(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s)/\kappa_n\right),
\end{aligned}$$

where the second equality follows from the fact that $\Psi^!_{\boldsymbol{x},\boldsymbol{q}}$ is a mixed binomial process $\mathrm{MB}(\nu, q_{\tilde{M}})$; moreover, $M'$ is a nonnegative integer-valued random variable with probability mass function $q_{M'}$ with $q_{M'}(m) \propto q_{\tilde{M}}(m)\kappa_n^m \propto \kappa_n^m q_M(m+k)(m+k)!/m!$. Therefore, $\Psi'$ is a mixed binomial process $\mathrm{MB}((1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s), q_{M'})$ and it is independent of $\boldsymbol{q}$.

*Proof of point (iii) of Corollary 4.3.* To describe the predictive distribution of $Z_{n+1}$ remind that $Z'_{n+1}$ is obtained by first thinning the process $\Psi'$ with retention probability $p(x,s) = s$, and then discarding the second component, as described in Theorem 4.3. By Proposition 4.7, since $\Psi'$ is a mixed binomial process, the thinned process is still mixed binomial, specifically $\mathrm{MB}(s(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s), q_{K'})$, with

$$\begin{aligned}
q_{K'}(m) &= \sum_{z\geq m}\binom{z}{m} q_{M'}(z) c_p^m (1-c_p)^{z-m} \\
&\propto \sum_{z\geq m}\binom{z}{m}\kappa_n^z q_M(z+k) c_p^m (1-c_p)^{z-m}(z+k)!/z!,
\end{aligned}$$

where $c_p = \int_{\mathbb{X}\times(0,1]} s(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s)/\int_{\mathbb{X}\times(0,1]}(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s)$. Removing the second component of the retained points, we obtain $Z'_{n+1}$, which is then a mixed binomial $\mathrm{MB}(\int_{(0,1]} s(1-s)^n\nu(\mathrm{d}x\,\mathrm{d}s), q_{K'})$.

### 4.F.4 The independently marked (repulsive) determinantal process prior: proof of Corollary 4.4 and details of Example 4.3

*Proof of point (i) of Corollary 4.4.* The marginal distribution of $\boldsymbol{Z}$ is recovered from Theorem 4.1 as follows. Since $\Psi$ is an independently marked process with ground process $\xi$ and mark kernel $H$, then from (Baccelli et al., 2020, Proposition 3.2.14), the reduced Palm version $\Psi^!_{x,s}$ is still an independently marked process, with ground process $\xi^!_x$ and mark kernel $H$, thus it does not depend on $s$. By Palm algebra, we can extend this property by claiming that $\Psi^!_{\boldsymbol{x},\boldsymbol{s}}$ is an independently marked process, with ground process $\xi^!_{\boldsymbol{x}}$ and mark kernel $H$. Then,

$$\begin{aligned}
\mathsf{E}\left\{e^{\int_{\mathbb{X}\times(0,1]} n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(\mathrm{d}z\,\mathrm{d}t)}\right\} &= \mathcal{L}_{\Psi^!_{\boldsymbol{x},\boldsymbol{s}}}(-n\log(1-t)) \\
&= \mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left[-\log\left\{\int_{(0,1]}(1-t)^n H(\mathrm{d}t\,|\,x)\right\}\right],
\end{aligned} \tag{4.29}$$

where the second equality follows from (Baccelli et al., 2020, Proposition 2.2.20). The thesis in point (i) of the statement follows.

*Proof of point (ii) of Corollary 4.4.* For the posterior distribution of $\tilde{\mu}$, expressed in Theorem 4.2, we need to determine the law of the vector $\boldsymbol{q} = (q_1, \ldots, q_k)$ and the law of $\mu'$, conditionally to $\boldsymbol{q}$. From point (i) of Theorem 4.2 and (4.29), which does not depend on $\boldsymbol{s}$, the $q_\ell$'s are independent with marginal laws $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell}H(\mathrm{d}s \,|\, X_\ell)$. Moreover, by point (ii) in Theorem 4.2, the law of $\Psi'$, conditionally to $\boldsymbol{q}$, has Laplace functional given by

$$\begin{aligned}\mathcal{L}_{\Psi' \,|\, \boldsymbol{q}}(g) &= \frac{\mathcal{L}_{\Psi^!_{\boldsymbol{x},\boldsymbol{q}}}(g(x,s) - n\log(1-s))}{\mathcal{L}_{\Psi^!_{\boldsymbol{x},\boldsymbol{q}}}(-n\log(1-s))} \\ &= \frac{\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left[-\log\left\{\int_{(0,1]} e^{-g(x,s)}(1-s)^n H(\mathrm{d}s \,|\, x)\right\}\right]}{\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left[-\log\left\{\int_{(0,1]}(1-s)^n H(\mathrm{d}s \,|\, x)\right\}\right]}.\end{aligned}$$

By (Baccelli et al., 2020, Proposition 2.2.20), this equals the Laplace transform of the independently marked point process $\Psi' = \sum_{j\geq 1}\delta_{(\tilde{X}'_j, q'_j)}$ where $q'_j \,|\, \tilde{X}'_j = x'_j \sim H'(\cdot \,|\, x'_j) \propto (1-s)^n H(\mathrm{d}s \,|\, x'_j)$ and $\xi' = \sum_{j\geq 1}\delta_{\tilde{X}'_j}$ has Laplace transform as in the statement of Corollary 4.4.

*Details of Example 4.3.* The results in Example 4.3 are obtained by specializing the treatment to the case where the mark kernel $H(\cdot \,|\, x)$ corresponds to the law of a beta with parameters $(a, b)$. Under this assumption, we get that the law of $M'$ can be determined as follows

$$\begin{aligned}\mathcal{L}_{M'}(u) &= \mathsf{E}\left\{e^{-u\xi'(\mathbb{X})}\right\} = \mathcal{L}_{\xi'}(u\mathbb{1}_{\mathbb{X}}) \\ &= \frac{\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left[u - \log\left\{\int_{(0,1]}(1-s)^n \mathrm{Beta}(\mathrm{d}s\,; a, b)\right\}\right]}{\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left[-\log\left\{\int_{(0,1]}(1-s)^n \mathrm{Beta}(\mathrm{d}s\,; a, b)\right\}\right]} \\ &= \frac{\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left[u - \log\left\{B(a, b+n)/B(a,b)\right\}\right]}{\mathcal{L}_{\xi^!_{\boldsymbol{x}}}\left[-\log\left\{B(a, b+n)/B(a,b)\right\}\right]}.\end{aligned}$$

Moreover, we focus on the mean measure of $\xi'$, denoted with $M_{\xi'}$. Indicating with $f_{\xi'}$ the density of the law of $\xi'$ with respect to the law of $\xi^!_{\boldsymbol{x}}$, we have

$$M_{\xi'}(A) = \mathsf{E}\left\{f_{\xi'}(\xi^!_{\boldsymbol{x}})\xi^!_{\boldsymbol{x}}(A)\right\} = \mathsf{E}\left\{\int_{\mathbb{X}} \mathbb{1}_A(y) f_{\xi'}(\xi^!_{\boldsymbol{x}})\xi^!_{\boldsymbol{x}}(\mathrm{d}y)\right\}.$$

By applying the alternative statement of the CLM formula in Lemma 4.4 in terms of the Palm distributions, we obtain that the mean measure of $\xi'$ equals

$$\begin{aligned}M_{\xi'}(A) &= \int_{\mathbb{X}} \mathsf{E}\left[\mathbb{1}_A(y) f_{\xi'}\left\{\left(\xi^!_{\boldsymbol{x}}\right)_y\right\}\right] M_{\xi^!_{\boldsymbol{x}}}(\mathrm{d}y) \\ &= \int_{\mathbb{X}} \mathsf{E}\left[\mathbb{1}_A(y) f_{\xi'}\left\{\left(\xi^!_{\boldsymbol{x}}\right)^!_y + \delta_y\right\}\right] M_{\xi^!_{\boldsymbol{x}}}(\mathrm{d}y) \\ &= \int_{\mathbb{X}} \mathsf{E}\left[\mathbb{1}_A(y) f_{\xi'}\left\{\xi^!_{(\boldsymbol{x},y)} + \delta_y\right\}\right] M_{\xi^!_{\boldsymbol{x}}}(\mathrm{d}y) \\ &= g(n; a, b)\int_A \mathsf{E}\left[f_{\xi'}\left\{\xi^!_{(\boldsymbol{x},y)}\right\}\right] M_{\xi^!_{\boldsymbol{x}}}(\mathrm{d}y).\end{aligned}$$

### 4.F.5 Proof of Proposition 4.2

Conditionally to $\alpha$, the posterior distribution of $\tilde{\mu}$ and the marginal law of the sample $\boldsymbol{Z}$ are as in Corollary 4.1. Moreover, the posterior law of $\alpha$ is given by

$$\pi_{\alpha \mid \boldsymbol{Z}}(\mathrm{d}a) \propto \exp\{-\varphi_n(a)\} \prod_{\ell=1}^{k} B(m_\ell - a, n - m_\ell + \beta + a) \times \pi_\alpha(\mathrm{d}a) \tag{4.30}$$

where $\varphi_n(a) = \gamma \int_{(0,1]} \{1-(1-s)^n\} s^{-1-a}(1-s)^{\beta+a-1}\mathrm{d}s = \sum_{h=0}^{n-1} B(-a+1, h+\beta+a)$.

From the posterior distribution of $\tilde{\mu}$ given $\alpha$ and the posterior density of $\alpha$ in (4.30), it is evident that the predictive distribution for the newly discovered features depends on the observed sample through $n$, $k$ and $m_1, \ldots, m_k$, while it does not depend on the labels $X_1, \ldots, X_k$.

## 4.G Additional details about the synthetic scenarios

We present here the analysis of the synthetic scenarios from Section 4.5.2 under the oracle strategy, which assumes knowledge of the true values of all hyperparameters used to generate the data. First, we examine inference on the total number of trees, $M' + k$, across different sample sizes, $n \in \{5, 15, 25, 50\}$. Figure 4.G.1 compares results obtained via the exact computation of the Poisson-binomial distribution and Le Cam's approximation. Additionally, we assess the naive approach described in Section 4.5.2. The same observations made for Figure 4.5.1 apply here. Specifically, the naive approach performs poorly across all sample sizes when estimating the total number of trees. While Le Cam's approximation is computationally faster, it introduces some errors, leading to a slight underestimation of the tree count. In contrast, the exact posterior distribution is centered around the true number of trees, with its variance decreasing as the sample size increases.

Next, we address the problem of locating unobserved trees through $M_{\xi'}$, as discussed in Section 4.5.1. Under the naive approach, this task is performed using the mean measure of $\xi^!_{\boldsymbol{x}}$. Figure 4.G.2 compares inference from our model and the naive alternative for $n = 15$, while Figure 4.G.3 presents results from our model for $n \in \{5, 25, 50\}$. Similar insights to those discussed for Figure 4.5.2 apply. In particular, Figure 4.G.2 shows that the naive approach assumes a higher probability of finding an unobserved tree at locations farther from the observed trees $\boldsymbol{x}$, as inference is based on $\xi^!_{\boldsymbol{x}}$, which is a DPP. In contrast, Figures 4.G.2 and 4.G.3 reveal the distinctive repulsive structure of $\xi'$ across different sample sizes. Specifically, our model predicts a low probability of finding a tree at $\mathrm{d}x$ if it is too close to an observed tree, as expected. However, it also suggests that locations "too far" from observed trees may have a small probability of containing a tree, as they might not align with the accumulation patterns inferred from the data.

For completeness, Figure 4.G.4 presents inference on the locations of unobserved trees through $M_{\xi'}$ for $n \in \{5, 25, 50\}$ under the scenario where hyperparameters are estimated via the empirical Bayes approach from Section 4.5.2. As aimed for, the results are consistent with those obtained under the oracle strategy, as evidenced by the clear similarity between Figures 4.G.4 and 4.G.3.

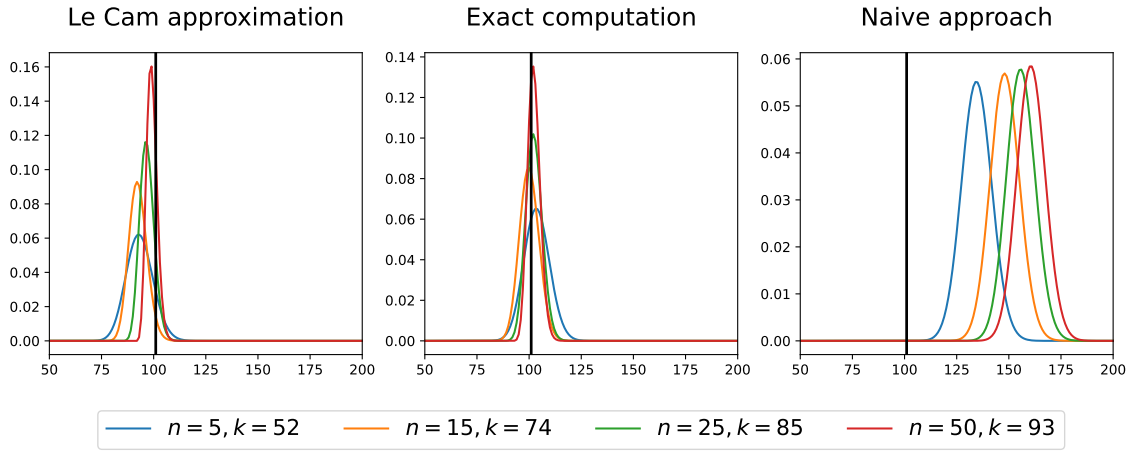



Figure 4.G.1: Posterior distribution of the total number of trees. From left to right: calculations performed using Le Cam's approximation of the Poisson-binomial, exact computations, and posterior of $\xi^!_{\boldsymbol{x}}(\mathbb{X})$. Different line colors correspond to different sample sizes; the black vertical line indicates the true number of trees.

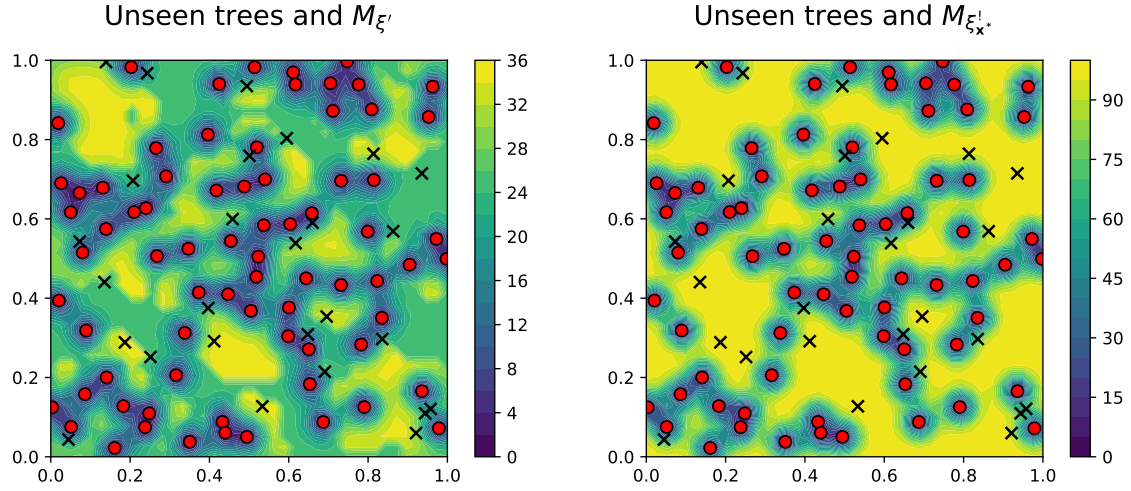



Figure 4.G.2: Locating the unobserved trees for $n = 15$: infinitesimal probability of observing an unseen tree in a given location. Left plot: the mean measure of $\xi'$. Right plot: the mean measure of $\xi^!_{\boldsymbol{x}}$. The red dots represent the observed trees in the sample. The black crosses indicate the unseen trees. Note that the color scales of the two plots are different.

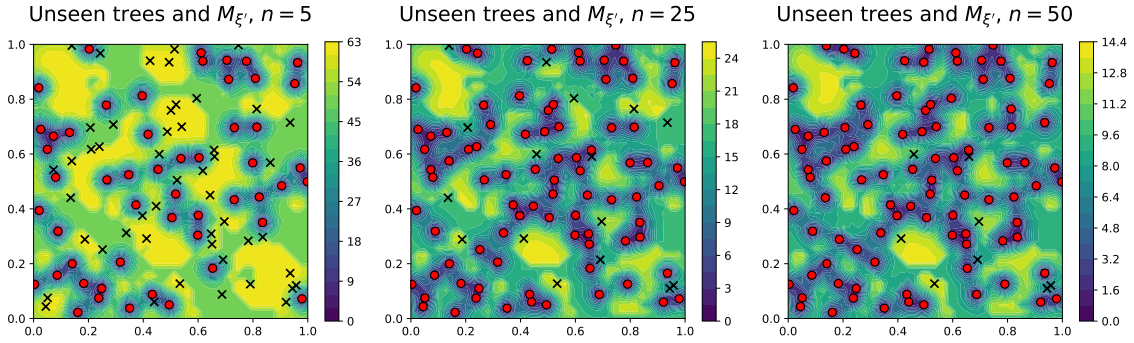



Figure 4.G.3: Locating the unobserved trees for $n \in \{5, 25, 50\}$: infinitesimal probability of observing an unseen tree in a given location. The three plots report $M_{\xi'}$ for the three sample sizes. The red dots represent the observed trees in the sample. The black crosses indicate the unseen trees. Note that the color scales of the plots are different.

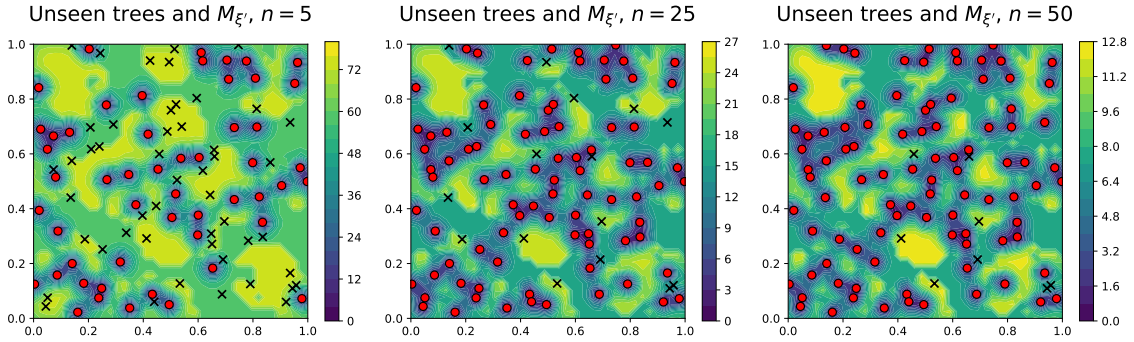



Figure 4.G.4: Locating the unobserved trees for $n \in \{5, 25, 50\}$: infinitesimal probability of observing an unseen tree in a given location. The three plots report $M_{\xi'}$ for the three sample sizes. The red dots represent the observed trees in the sample. The black crosses indicate the unseen trees. Note that the color scales of the plots are different.

# 5 Bayesian nonparametric modeling of multivariate count data with an unknown number of traits

This chapter introduces a novel general and tractable class of Bayesian nonparametric priors suitable for modeling partially exchangeable trait allocations. Our proposal relies on completely random vectors (CRVs), defined in Catalano et al. (2021). Also, see Kallenberg (2017). The proposed theoretical framework is very broad and therefore we focus on a notable subclass, i.e., finite completely random vectors (FCRVs), which prescribe that the number of traits in the population is finite but random. We provide a comprehensive theoretical analysis of partially exchangeable trait allocations under CRVs and FCRVs, including: (i) the marginal distribution of a sample and (ii) posterior representations. To illustrate the applicability of our framework, we discuss two examples involving binary traits (i.e. features) and Poisson counts. In these key special cases, the analytical derivations are extremely tractable.

To illustrate the practical relevance of our methodological framework, we analyze the criminal network dataset of the 'Ndrangheta, previously examined in Legramanti et al. (2022); Lu et al. (2025). The data were collected during *Operazione Infinito* (Calderoni et al., 2017), a large-scale law enforcement initiative aimed at dismantling the core branch of the 'Ndrangheta Mafia in the Milan area. The dataset records multivariate binary outcomes describing the attendance of known affiliates (subjects) at a series of meetings (binary traits, i.e., features). In addition, affiliates can be grouped according to their *locali* affiliation, as documented in juridical records. This partition naturally suggests a partially exchangeable (or *known-groups*) framework, where inference can be carried out in closed form by leveraging our theoretical results. A distinctive feature of the proposed trait allocation model, in contrast with classical approaches to multivariate count data, is that the number of traits, that is, the columns of the data matrix, is itself a random variable. In our motivating application, this means that some meetings may remain unobserved because they were not detected by law enforcement. Our methodology explicitly accounts for this possibility while also enabling the estimation of the number of unseen traits.

Unfortunately, the *locali* partition, while highly informative, is insufficient to fully capture the complexity of the relationships among subjects (Legramanti et al., 2022; Lu et al., 2025). To address this limitation, we extend the partially exchangeable model by allowing the partition structure itself to be inferred from the data, thereby leading to a clustering problem. In this setting, the groups are not predetermined but are learned from the data; for this reason, we refer to it as the *unknown-groups* framework. The second main contribution of this work is thus a novel mixture model designed to cluster trait allocations. Inference is enabled by the distributional results established for the simpler known-groups setting, leading to an efficient Gibbs sampling strategy. Importantly, by treating the parti-

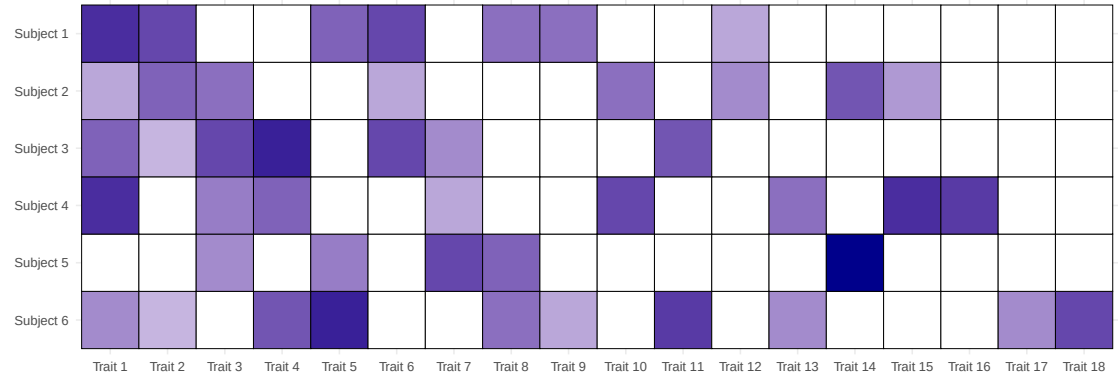



Figure 5.1.1: Observed data from an exchangeable trait model: matrix of counts $\boldsymbol{A}$, with $n = 6$ subjects and $K_n = 18$ observed traits. Rows and columns are arranged in no particular order. White cells, such as $A_{13} = 0$, indicate the absence of a trait for a given subject, while darker shades of blue represent higher values of the corresponding counts $A_{i\ell} \in \{1, 2, \dots\}$.

tion as a random quantity, the proposed model generalizes Bayesian nonparametric latent class models, such as Dunson and Xing (2009). A key distinction, however, is that the total number of traits is unknown and must be estimated from the data. We formally demonstrate that ignoring this aspect inevitably results in overclustering.

The chapter is organized as follows. Section 5.1 reviews the classical framework of exchangeable trait allocation models, which serves as the foundation for our approach to modeling multivariate count data. Section 5.2 introduces the partially exchangeable setting for trait allocation models and provides a complete Bayesian analysis, including closed-form expressions for the marginal and posterior distributions. Section 5.3 develops our more elaborate methodology for clustering count data with potentially unobserved traits. We also examine a naïve specification that disregards the unseen traits, and we demonstrate, both theoretically and empirically, that such an approach can bias the resulting analyses. Section 5.4 presents simulation studies assessing the performance of our methodology, while Section 5.5 applies it to the criminal network dataset from the *Operazione Infinito* investigation.

## 5.1 Background on exchangeable trait allocation models

We begin by reviewing exchangeable trait allocation models (Campbell et al., 2018), in which data for each of the $n$ subjects are conditionally i.i.d. draws from a common and simple distribution. These highly tractable models serve as building blocks for the more flexible approaches introduced later in Sections 5.2-5.3.

Exchangeable trait allocation models describe how a set of traits is distributed across a sample of $n$ subjects, where the presence of each trait in a subject is associated with a quantitative measurement, typically an integer, that reflects the expression or abundance of that trait in the subject. More formally, suppose we observe $n$ subjects and $K_n = k$ distinct traits. The data can be represented by an $n \times k$ matrix $\boldsymbol{A}$, where each entry $A_{i\ell} \in \{0, 1, 2, \dots\}$ denotes the count of the $\ell$th trait (column) for the $i$th subject (row),

as illustrated in Figure 5.1.1. We say that trait $\ell$ is *absent* (or not observed) in subject $i$ if and only if $A_{i\ell} = 0$. Note that each column of $\boldsymbol{A}$ contains at least one non-zero entry. Each trait is associated with a distinct *label*, denoted $X_\ell \in \mathbb{X}$, which serves as a placeholder and is not explicitly modeled; it simply identifies the column. Here $\mathbb{X}$ denotes the space of the labels, say $\mathbb{X} = (0, 1)$ for simplicity. Importantly, a given trait $X_\ell$ may be shared by multiple subjects. A defining feature of trait allocation models, unlike classical multivariate count data models, is that the number of traits (columns) is itself random. In other words, some traits may remain *unseen*. To model this explicitly, let $(\tilde{X}_j)_{j\geq 1}$ denote the sequence of all possible trait labels and let $\tilde{A}_{ij} \in \{0, 1, 2, \dots\}$ represent the abundance of trait $\tilde{X}_j$ in subject $i$. In a sample of size $n$, a trait $\tilde{X}_j$ is observed only if $\tilde{A}_{ij} > 0$ for at least one subject. In other words, the observed traits $X_1, \dots, X_{K_n}$ form a subsample of the latent traits $(\tilde{X}_j)_{j\geq 1}$, and the corresponding observed counts satisfy $\sum_{i=1}^n A_{i\ell} > 0$; otherwise, the trait would not be observed. For mathematical convenience, we can organize the pairs $((\tilde{X}_j, \tilde{A}_{ij}))_{j\geq 1}$ by means of subject-specific counting measures $(Z_i)_{i\geq 1}$ on $\mathbb{X}$, namely

$$Z_i(\cdot) = \sum_{j\geq 1} \tilde{A}_{ij}\delta_{\tilde{X}_j}(\cdot), \tag{5.1}$$

where $\delta_x$ denotes the Dirac delta mass at the point $x \in \mathbb{X}$. Common assumption in trait allocation models requires that each subject may exhibit only a finite number of traits, ensuring that the total number of distinct traits $K_n$ is almost surely finite in any given sample.

In the exchangeable case, the random variables $\tilde{A}_{ij}$, given a sequence of parameters $(\theta_j)_{j\geq 1}$, are conditionally i.i.d. across subjects (rows) for any fixed $j$, that is

$$\tilde{A}_{ij} \,|\, \theta_j \overset{\text{iid}}{\sim} P(\cdot\,; \theta_j), \qquad i \geq 1, \tag{5.2}$$

and they are also conditionally independent across traits (columns) for $j \geq 1$. Here, $P(\cdot\,; \theta)$ denotes any parametric distribution supported on the non-negative integers, such as a Poisson distribution, depending on a positive parameter $\theta > 0$. The parameters $(\theta_j)_{j\geq 1}$ can be organized in a discrete measure $\tilde{\mu}$ on $\mathbb{X}$, defined as

$$\tilde{\mu}(\cdot) = \sum_{j\geq 1} \theta_j \delta_{\tilde{X}_j}(\cdot). \tag{5.3}$$

Note that the atoms of the discrete measure $\tilde{\mu}$, i.e., the trait labels $\tilde{X}_j$, are common across all subjects, so that the same traits are allowed to be observed in multiple subjects. Summarizing, the full Bayesian specification is

$$\begin{aligned} Z_i \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \text{CP}(\tilde{\mu}), \qquad i \geq 1, \\ \tilde{\mu} &\sim \mathcal{Q}, \end{aligned}$$

which means that $Z_i$ in (5.1) are i.i.d. from a *process of counts* (CP) with parameter $\tilde{\mu}$ defined by (5.2)-(5.3). Here $\mathcal{Q}$ denotes the de Finetti measure, i.e., the prior distribution of the random measure $\tilde{\mu}$.

The exchangeable setting has been investigated extensively, e.g., in James (2017); Campbell et al. (2018), and, in the binary case, more recently by us in Chapter 3 and Chapter 4. In the following, we present two relevant examples for the distribution $P(\cdot; \theta)$.

**Example 5.1** (Exchangeable binary traits)**.** In our motivating application involving meetings of criminals, we track the attendance of 'Ndrangheta affiliates (subjects) at various meetings (traits). Thus, the $K_n = k$ observed traits correspond to distinct meetings where at least one of the $n$ affiliates has been identified by investigators. Clearly, it is likely that a few meetings have not been spotted therefore it is reasonable to model the total number of meetings as a random variable. In this case study, the count measurement $\tilde{A}_{ij}$ are binary, reducing the trait allocation framework to the special case of feature allocation models (Broderick et al., 2013). Thus, we let $\tilde{A}_{ij} \,|\, \theta_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta_j)$ for $i \geq 1$ and any fixed $j$ with success probabilities $\theta_j \in (0, 1)$, so that $Z_i \,|\, \tilde{\mu} \overset{\text{iid}}{\sim} \text{CP}(\tilde{\mu})$ are i.i.d. draws from a Bernoulli process (Griffiths and Ghahramani, 2011). A broad class of tractable prior distributions for $\tilde{\mu}$ are described in Chapter 3 to which we refer for further details and applications to ecology.

**Example 5.2** (Exchangeable Poisson counts)**.** When the data $\tilde{A}_{ij}$ take values in $\{0, 1, 2, \dots\}$, as those depicted in Figure 5.1.1, a natural choice is $P(a; \theta) = (a!)^{-1}\theta^a e^{-\theta}$ for $a \in \{0, 1, \dots\}$, that is a Poisson distribution with mean $\theta > 0$. In other words, we assume $\tilde{A}_{ij} \,|\, \theta_j \overset{\text{iid}}{\sim} \text{Poisson}(\theta_j)$ for $i \geq 1$ and any fixed $j$. This specification has been less explored compared to the binary case.

**Remark 5.1.** *Throughout the chapter, we assume that the parametric distribution $P(\cdot\,; \theta)$ is governed by a single positive parameter $\theta > 0$. In the binary case this assumption is not restrictive. However, it may be limiting when modeling count data or categorical data. For instance, more flexible alternatives such as zero-inflated Poisson or multinomial distributions may be preferable. In principle, a theoretical analysis allowing for a general parameter space is possible, but it would require moving beyond completely random vectors—the main technical tool discussed in Section 5.2—and instead rely on general Poisson processes. To streamline the presentation, we therefore focus on this subclass of models, particularly in light of our motivating application involving binary data. Nonetheless, in Section 5.B.1 of the Appendix, we discuss that all theoretical results remain valid with only minor modifications, and we provide an example.*

## 5.2 Partially exchangeable finite trait allocation models

### 5.2.1 Model specification and completely random vectors

In this section, we introduce a novel class of trait allocation models that relaxes the assumption of exchangeability. We consider a framework in which subjects are partitioned into subpopulations: subjects from different groups are conditionally independent but not identically distributed, while exchangeability holds within each group. This structure, known as partial exchangeability, is well-suited to the 'Ndrangheta network data, where affiliates can be naturally grouped according to their membership in specific *locali*. Notably, this extension preserves full analytical tractability, as the posterior distribution of remains available in closed form.

Let $d$ be the number of subpopulations, and suppose we observe a sample of size $n$, with $n_q$ subjects from group $q$, for $q = 1, \dots, d$, so that $\sum_{q=1}^{d} n_q = n$. Let $K_n = k$ denote the total number of traits observed across all subjects and groups. The data can be represented by a collection of matrices $\boldsymbol{A}_q$, each of dimension $n_q \times k$, where the entry

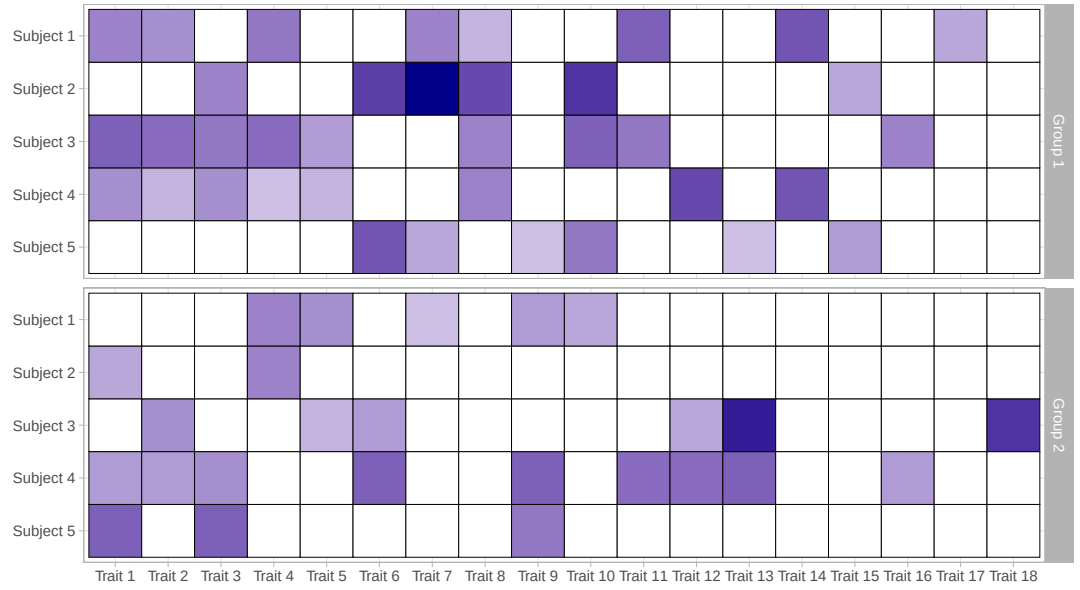



Figure 5.2.1: Observed data from a partially exchangeable trait model ($d = 2$): two matrices of counts $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, each with $n_1 = n_2 = 5$ subjects and $K_n = 18$ observed traits. As in Figure 5.1.1, white cells, such as $A_{221} = 0$, indicate the absence of a trait for a given subject, while darker shades of blue represent higher values of the corresponding counts $A_{i\ell q} \in \{1, 2, \dots\}$.

$A_{i\ell q} \in \{0, 1, 2, \dots\}$ denotes the count of the $\ell$th trait for the $i$th subject in group $q$, as illustrated in Figure 5.2.1. Note that a trait may be unobserved within the sample of a specific group. As before, let $(\tilde{X}_j)_{j\geq 1}$ denote the sequence of all possible trait labels and let $\tilde{A}_{ijq} \in \{0, 1, 2, \dots\}$ be the abundance of trait $\tilde{X}_j$ for subject $i$ in group $q$. In a sample of size $n$, a trait $\tilde{X}_j$ is observed only if $\tilde{A}_{ijq} > 0$ for at least one subject belonging to any group.

We organize these quantities into counting measures $Z_{iq}(\cdot) = \sum_{j\geq 1} \tilde{A}_{ijq}\, \delta_{\tilde{X}_j}(\cdot)$ for each subject $i$ in subpopulation $q$, with $i \geq 1$ and $q = 1, \dots, d$. In the partially exchangeable case, the random variables $\tilde{A}_{ijq}$, given the sequences of parameters $(\theta_{j1})_{j\geq 1}, \dots, (\theta_{jd})_{j\geq 1}$, are conditionally i.i.d. across subjects belonging to the same group and for a given trait $j$ and group $q$, that is

$$\tilde{A}_{ijq} \,|\, \theta_{jq} \overset{\text{iid}}{\sim} P(\cdot\,; \theta_{jq}), \qquad i \geq 1,$$

and they are also conditionally independent across traits for $j \geq 1$ and subpopulations $q = 1, \dots, d$. Thus, the main difference compared to the exchangeable case is that the random variables $\tilde{A}_{ijq}$ have different parameters when they refer to subjects belonging to different subpopulations. Moreover, the parameters $(\theta_{jq})_{j\geq 1}$ can be organized in a group-specific discrete measure $\tilde{\mu}_q(\cdot) = \sum_{j\geq 1} \theta_{jq} \delta_{\tilde{X}_j}(\cdot)$ for $q = 1, \dots, d$. Summarising, the full Bayesian specification for partially exchangeable data is

$$\begin{aligned} Z_{iq} \,|\, \tilde{\mu}_q &\overset{\text{ind}}{\sim} \text{CP}(\tilde{\mu}_q), \qquad i \geq 1, \quad q = 1, \dots, d, \\ (\tilde{\mu}_1, \dots, \tilde{\mu}_d) &\sim \mathcal{Q}_d, \end{aligned} \tag{5.4}$$

where $\mathcal{Q}_d$ denotes the de Finetti measure of the vector of random measures $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_d)$, that is, the *a priori* distribution. As a clarification, note that for any fixed group $q$ the

measures $Z_{iq}$ are conditionally i.i.d. draws for $i \geq 1$. Moreover, the exchangeable case discussed in Section 5.1 is recovered when $d = 1$.

Introducing some form of dependence among $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ is desirable, as it enables borrowing of information across subpopulations. The challenging task is to elicit a prior $\mathcal{Q}_d$ that induces dependence among the measures while remaining analytically tractable. In pursuing this, we note that inferential goals might relate to: (i) the estimation of trait- and group-specific parameters $\theta_{jq}$; (ii) the estimation of the number of unseen traits in the observed sample. To enable the latter goal, we assume that the total number of traits in the population, denoted with $M$, is finite and random. This implies there will be a finite collection of random variables $\tilde{A}_{i1q}, \ldots, \tilde{A}_{iMq}$ for each subject and group, with associated parameters $\theta_{1q}, \ldots, \theta_{Mq}$ for $q = 1, \ldots, d$. Moreover, we assume $M$ is a Poisson random variable with parameter $\lambda > 0$. In other terms, the group-specific measures $\tilde{\mu}_q$ in (5.4) take the form

$$\tilde{\mu}_q(\cdot) = \sum_{j=1}^{M} \theta_{jq} \delta_{\tilde{X}_j}(\cdot), \qquad M \sim \text{Poisson}(\lambda), \tag{5.5}$$

as $q = 1, \ldots, d$. Clearly, if $K_n = k$ traits are observed in the finite sample, the number of unseen traits equals $M - k$. Moreover, we assume the parameter vectors $(\theta_{j1}, \ldots, \theta_{jd})$ are i.i.d. draws from a probability law $H^{(d)}$ defined on $(0, \infty)^d$, namely

$$(\theta_{j1}, \ldots, \theta_{jd}) \overset{\text{iid}}{\sim} H^{(d)}, \qquad j = 1, \ldots, M. \tag{5.6}$$

In this model specification, information is borrowed across groups in two ways: (i) by assuming a common total number $M$ of traits, and (ii) by inducing dependence between the parameters $\theta_{jq_1}$ and $\theta_{jq_2}$ for $q_1 \neq q_2$ through a joint probability distribution $H^{(d)}$. In practice, subpopulations are often exchangeable, making it natural to consider a factorized measure of the form $H^{(d)}(\cdot\,; \psi) = H(\cdot\,; \psi) \times \cdots \times H(\cdot\,; \psi)$, where $\psi$ is a common parameter endowed with a hyperprior. Under this formulation, the parameters $\theta_{jq}$ are conditionally i.i.d. for $j = 1, \ldots, M$ and $q = 1, \ldots, d$, given $\psi$, according to a probability measure $H(\cdot\,; \psi)$. Throughout the chapter, we focus on this special case, as it is the most relevant in practice. General results are provided in Section 5.B of the Appendix. Finally, we need to specify a prior on the atoms $\tilde{X}_j$. As previously mentioned, since the common atoms $\tilde{X}_j$ serve only to label different traits in this formulation, it is sufficient for them to be almost surely distinct. For example, one may assume $\tilde{X}_j \overset{\text{iid}}{\sim} G_0$ for $j = 1, \ldots, M$, where $G_0$ is any non-atomic distribution.

In equations (5.5)–(5.6), we specified a prior distribution $\mathcal{Q}_d$ for the vector of random measures $\tilde{\boldsymbol{\mu}}$ in (5.4) as a finite collection of random variables. This specification is both transparent and constructive, but it does not make explicit the connection with infinite-dimensional trait models (e.g. James, 2017; Shen et al., 2024). In Section 5.A of the Appendix, we prove that $\tilde{\boldsymbol{\mu}}$ is a finite completely random vector, which is a special case of the broader class of completely random vectors (Catalano et al., 2021); see also Kallenberg (2017) for a comprehensive treatment and Section 5.A of the Appendix for a concise overview. This result crucially relies on the Poisson specification for $M$. More precisely, $\tilde{\boldsymbol{\mu}}$ can be interpreted as an FCRV with parameters $H^{(d)}$, $\lambda$, and $G_0$, that is, a CRV whose Lévy intensity takes the form $H^{(d)}(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) \cdot \lambda G_0(\mathrm{d}x)$, and we write $\tilde{\boldsymbol{\mu}} \sim \text{FCRV}(H^{(d)}, \lambda, G_0)$. This connection to the theory of completely random vectors is helpful, since the results

presented in Section 5.2.2 follow as special cases of this general theory. In Section 5.B of the Appendix we present results for the whole class of CRVs characterized by Lévy intensities of the form $\rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) \cdot \lambda G_0(\mathrm{d}x)$, where $\rho_d$ is a possibly infinite measure. This general framework encompasses both finite- and infinite-dimensional trait allocation models. The main practical challenge is identifying suitable, non-degenerate choices for $\rho_d$. Our proposal focuses on the finite case $\rho_d = H^{(d)}$ with $H^{(d)}$ being a probability distribution, which allows for the estimation of the total number of traits, while Shen et al. (2024) provide an alternative construction in the infinite-dimensional setting.

### 5.2.2 Distribution theory and posterior inference

Here we provide a full Bayesian analysis of the proposed model in the known-groups case. More specifically, we obtain tractable closed-form expressions of the marginal distribution of a sample, the posterior distribution of $\tilde{\boldsymbol{\mu}}$, and the predictive distribution of a future observation, under model (5.4) with the prior $\tilde{\boldsymbol{\mu}} \sim \mathrm{FCRV}(H^{(d)}, \lambda, G_0)$ as specified in equations (5.5)–(5.6). As mentioned earlier, we focus on the case where $H^{(d)}(\cdot\,;\psi) = H(\cdot\,;\psi) \times \cdots \times H(\cdot\,;\psi)$ factorizes, for the sake of simplicity. Readers interested in the more general setting are referred to the Appendix, in particular Section 5.B.

We start by describing the marginal distribution of a sample from model (5.4), and here we offer a simple and constructive proof. With *marginal distribution of the sample* $\boldsymbol{Z} = (Z_{iq} : i = 1, \ldots, n_q; q = 1, \ldots, d)$, we specifically mean determining the probabilities of the event $(\boldsymbol{A} = \boldsymbol{a}, K_n = k)$, having denoted by $\boldsymbol{A} = (A_{i\ell q} : i = 1, \ldots, n_q; \ell = 1, \ldots, k; q = 1, \ldots, d)$ the observed counts, where the $K_n = k$ observed traits in the sample are randomly ordered. To this end, suppose for now that the total number of traits $M$ is fixed and let $\tilde{\boldsymbol{A}} = (\tilde{A}_{ijq} : i = 1, \ldots, n_q; j = 1, \ldots, M; q = 1, \ldots, d)$ denote the latent counts. Focus on any event $(\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{a}}, K_n = k)$ that contains the observed event $(\boldsymbol{A} = \boldsymbol{a}, K_n = k)$. Let $\mathcal{A} = \{j = 1, \ldots, M : \sum_{q=1}^{d} \sum_{i=1}^{n_q} \tilde{a}_{ijq} > 0\}$ be the indexes of observed traits, therefore $j \in \mathcal{A}$ if the $j$th latent trait is observed and $|\mathcal{A}| = k$. Conditionally on the parameters $\boldsymbol{\theta} = (\theta_{jq} : j = 1, \ldots, M; q = 1, \ldots, d)$ and $M$, the likelihood function $\mathscr{L}(\boldsymbol{\theta}, M; \tilde{\boldsymbol{a}})$ for the event $(\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{a}}, K_n = k)$ is

$$\mathscr{L}(\boldsymbol{\theta}, M; \tilde{\boldsymbol{a}}) = \prod_{q=1}^{d} \prod_{j=1}^{M} \prod_{i=1}^{n_q} P(\tilde{a}_{ijq}; \theta_{jq}) = \left[ \prod_{q=1}^{d} \prod_{j \notin \mathcal{A}} P(0; \theta_{jq})^{n_q} \right] \left[ \prod_{q=1}^{d} \prod_{j \in \mathcal{A}} \prod_{i=1}^{n_q} P(\tilde{a}_{ijq}; \theta_{jq}) \right]. \tag{5.7}$$

In the last term, the first product accounts for the unobserved traits, whereas the second relates to the observed ones. Note that the quantity $\prod_{q=1}^{d} P(0; \theta_{jq})^{n_q}$ may be interpreted as the probability of not observing trait $\tilde{X}_j$ among all subjects and groups. Then, we can combine the likelihood (5.7) with all the (i.i.d.) prior distributions $H(\mathrm{d}\theta; \psi)$, take the integral over $\boldsymbol{\theta}$ and sum over all possible sets $\mathcal{A}$, leading to the following marginal probability for the event $(\boldsymbol{A} = \boldsymbol{a}, K_n = k)$:

$$\pi_n(\boldsymbol{a}; M, \psi) = \binom{M}{k} \left[ \prod_{q=1}^{d} \int P(0; \theta)^{n_q} H(\mathrm{d}\theta; \psi) \right]^{M-k} \prod_{\ell=1}^{k} \prod_{q=1}^{d} \int \prod_{i=1}^{n_q} P(a_{i\ell q}; \theta) H(\mathrm{d}\theta; \psi), \tag{5.8}$$

where the binomial coefficient is introduced to account for all the possible ways we can arrange the $k$ observed traits among all the $M$ traits. However, since the total number

of traits $M$ is random and follows a Poisson distribution with mean $\lambda$, the final formula requires marginalizing the conditional law in (5.8) with respect to $M$. By exploiting well-known properties of the Poisson distribution, we obtain a simple closed-form expression, summarized in the theorem below.

**Theorem 5.1** (Marginal distribution, pETPF)**.** *Let $\boldsymbol{Z}$ be a sample from the statistical model (5.4), assuming the prior $\tilde{\boldsymbol{\mu}} \sim \text{FCRV}(H^{(d)}, \lambda, G_0)$ with $H^{(d)}(\cdot\,;\psi) = H(\cdot\,;\psi) \times \cdots \times H(\cdot\,;\psi)$. The probability that $\boldsymbol{Z}$ displays $K_n = k$ distinct traits with counts $\boldsymbol{A} = \boldsymbol{a}$ is given by*

$$\pi_n(\boldsymbol{a}; \lambda, \psi) = \frac{\lambda^k}{k!} \exp\left\{-\lambda\left(1 - \prod_{q=1}^{d} \int P(0;\theta)^{n_q} H(\mathrm{d}\theta;\psi)\right)\right\} \prod_{\ell=1}^{k}\prod_{q=1}^{d} \int \prod_{i=1}^{n_q} P(a_{i\ell q};\theta) H(\mathrm{d}\theta;\psi), \tag{5.9}$$

*where $n = \sum_{q=1}^{d} n_q$ and $\boldsymbol{n} = (n_1, \ldots, n_d)$ are the sample sizes.*

A rigorous and general proof of Theorem 5.1, not based on the above argument, is provided in the Appendix, Section 5.B, under general CRV priors, covering the case with infinitely many traits. It is interesting to note that the probabilistic quantity in (5.9) appears related to the partially Exchangeable Partition Probability Function (pEPPF) discussed in Franzolini et al. (2025) for partially exchangeable species sampling models; however, here we consider the distribution of a random trait allocation rather than a random partition. By analogy, we refer to equations (5.8) and (5.9) as a *partially Exchangeable Trait Probability Function*s (pETPF), with the former being conditional on $M$ and the latter being a Poisson mixture of (5.8) over $M$.

**Example 5.3** (Binary traits)**.** If the traits are binary and $P(a;\theta) = \theta^a(1-\theta)^{1-a}$, for $a \in \{0,1\}$, the marginal law of $\boldsymbol{Z}$ depends on a sufficient statistic of $\boldsymbol{a}$, corresponding to the feature frequencies in different groups. Define the random frequency of feature $X_\ell$ in group $q$ as $M_{\ell q} := \sum_{i=1}^{n_q} A_{i\ell q}$ with observed values $m_{\ell q}$, stored in matrices $\boldsymbol{M}$ and $\boldsymbol{m}$, respectively. Suppose that the prior law $H(\cdot\,;\psi)$ corresponds to a beta distribution with parameters $\psi = (-\alpha, \alpha+\beta)$, with $\alpha < 0$ and $\beta > -\alpha$. Then, the probability that $\boldsymbol{Z}$ displays $K_n = k$ distinct traits with feature frequencies $\boldsymbol{M} = \boldsymbol{m}$ is

$$\pi_n(\boldsymbol{m}; \lambda, \psi) = \frac{\lambda^k}{k!} \exp\left\{-\lambda\left[1 - \prod_{q=1}^{d} \frac{(\alpha+\beta)_{n_q}}{(\beta)_{n_q}}\right]\right\} \prod_{\ell=1}^{k}\prod_{q=1}^{d} \frac{-\alpha}{(\beta)_{n_q}} (1-\alpha)_{m_{\ell q}-1} (\alpha+\beta)_{n_q - m_{\ell q}}, \tag{5.10}$$

where $(a)_n = a(a+1)\cdots(a+n-1)$ is the ascending factorial with $a > 0$. When $d = 1$, this corresponds to the EFPF of the Poisson mixture of beta Bernoulli (Chapter 3).

**Example 5.4** (Poisson counts)**.** Another remarkable simplification is obtained when the trait counts $\tilde{A}_{ijq}$ take values in $\{0, 1, 2, \ldots\}$, and $P(a;\theta) = (a!)^{-1}\theta^a e^{-\theta}$ for $a \in \{0, 1, \ldots\}$, i.e., a Poisson distribution with mean $\theta > 0$, and $H(\mathrm{d}\theta;\psi) = \beta^\alpha/\Gamma(\alpha)\theta^{\alpha-1}e^{-\beta\theta}\mathrm{d}\theta$, i.e., a gamma prior with parameters $\psi = (\alpha, \beta)$. In this case, the probability that $\boldsymbol{Z}$ displays $K_n = k$ distinct traits with counts $\boldsymbol{A} = \boldsymbol{a}$ equals

$$\pi_n(\boldsymbol{a}; \lambda, \psi) = \frac{\lambda^k}{k!} \exp\left\{-\lambda\left[1 - \prod_{q=1}^{d} \left(\frac{\beta}{\beta+n_q}\right)^{\alpha}\right]\right\} \prod_{\ell=1}^{k}\prod_{q=1}^{d} \frac{\beta^\alpha (\alpha)_{m_{\ell q}}}{(\beta+n_q)^{\alpha+m_{\ell q}}} \prod_{i=1}^{n_q} \frac{1}{a_{i\ell q}!},$$

where $m_{\ell q} = \sum_{i=1}^{n_q} a_{i\ell q}$.

We now characterize the posterior distribution of $\tilde{\boldsymbol{\mu}}$, conditionally to a sample $\boldsymbol{Z}$, which is essential to provide Bayesian estimators of the parameters $\theta_{jq}$ and the total number of traits $M$.

**Theorem 5.2** (Posterior distribution)**.** *Let $\boldsymbol{Z}$ be a sample from the statistical model* (5.4)*, assuming the prior $\tilde{\boldsymbol{\mu}} \sim \text{FCRV}(H^{(d)}, \lambda, G_0)$ with $H^{(d)}(\cdot\,;\psi) = H(\cdot\,;\psi) \times \cdots \times H(\cdot\,;\psi)$. If $\boldsymbol{Z}$ displays $K_n = k$ distinct traits labeled $X_1, \ldots, X_k$, with associated counts $\boldsymbol{a}$, then the posterior distribution of $\tilde{\boldsymbol{\mu}}$ satisfies the distributional equality*

$$(\tilde{\mu}_1, \ldots, \tilde{\mu}_d) \,|\, \boldsymbol{Z} \stackrel{d}{=} (\mu_1^*, \ldots, \mu_d^*) + (\mu_1', \ldots, \mu_d'), \tag{5.11}$$

*where $\boldsymbol{\mu}^* := (\mu_1^*, \ldots, \mu_d^*)$ and $\boldsymbol{\mu}' := (\mu_1', \ldots, \mu_d')$ are independent random vectors such that*

*(i) the components of the vector $\boldsymbol{\mu}^*$ are defined as $\mu_q^*(\cdot) = \sum_{\ell=1}^{k} \theta_{\ell q}^* \delta_{X_\ell}(\cdot)$, for $q = 1, \ldots, d$, and the random variables $\theta_{\ell q}^*$ are independent with distribution $H_{\ell q}(\mathrm{d}\theta; \psi) \propto \prod_{i=1}^{n_q} P(a_{i\ell q}; \theta) H(\mathrm{d}\theta; \psi)$;*

*(ii) the vector $(\mu_1', \ldots, \mu_d')$ is a $\text{FCRV}(H'^{(d)}, \lambda', G_0)$, where $H'^{(d)}(\cdot\,;\psi) = H_1'(\cdot\,;\psi) \times \cdots \times H_d'(\cdot\,;\psi)$ and*

$$\lambda' = \lambda \prod_{q=1}^{d} \int P(0;\theta)^{n_q} H(\mathrm{d}\theta; \psi), \quad H_q'(\mathrm{d}\theta; \psi) \propto P(0;\theta)^{n_q} H(\mathrm{d}\theta; \psi).$$

*This is equivalent to say that for each $q = 1, \ldots, d$*

$$\mu_q'(\cdot) = \sum_{j=1}^{M'} \theta_{jq}' \delta_{\tilde{X}_j'}(\cdot), \qquad \theta_{jq}' \overset{iid}{\sim} H_q', \qquad \tilde{X}_j' \overset{iid}{\sim} G_0, \qquad j = 1, \ldots, M', \tag{5.12}$$

*where $M' \sim \text{Poisson}(\lambda')$.*

See Section 5.B of the Appendix for a proof. Despite the complex notation, Theorem 5.2 has actually a very intuitive and clear interpretation. For each subpopulation $q$, the posterior distribution of $\tilde{\mu}_q$ describes both the observed traits out of the initial sample and the unseen traits, and indeed it consists of the sum of two terms in (5.11). The first term $\mu_q^*$, referring to the observed traits $X_1, \ldots, X_k$, simply describes the posterior distribution of the associated parameters, indicated with variables $\theta_{\ell q}^*$. Indeed, from point (i) of Theorem 5.2, we see that the law of each $\theta_{\ell q}^*$ is $H_{\ell q}(\mathrm{d}\theta; \psi) \propto \prod_{i=1}^{n_q} P(a_{i\ell q}; \theta) H(\mathrm{d}\theta; \psi)$, which corresponds to the plain application of Bayes theorem. The second term represents the novel and interesting component, since $\mu_q'$ takes into account potentially unobserved traits in the sample. From point (ii) of the theorem, $M'$ represents the number of unseen traits *a posteriori*, that is $M \stackrel{d}{=} M' + k$, which is distributed as a Poisson random variable with updated parameter $\lambda'$. Inference on the number of unseen traits can be carried out by inspecting the value of $\lambda'$. The expressions in theorem 5.2 substantially simplify for the cases considered in Examples 5.3 and 5.4.

**Example 5.5** (Binary traits, cont'd)**.** If the traits are binary and assuming a beta prior law $H(\cdot;\psi)$ with parameters $(-\alpha, \alpha + \beta)$, as in Example 5.3. The posterior distribution

of $\tilde{\boldsymbol{\mu}}$ given $\boldsymbol{Z}$ decomposes as the sum of $\boldsymbol{\mu}^*$ and $\boldsymbol{\mu}'$, as in the general case. However, the probabilities $\theta^*_{\ell q}$ of re-observing an old feature $X_\ell$ are distributed as $\theta^*_{\ell q} \overset{\text{ind}}{\sim} \text{Beta}(m_{\ell q} - \alpha, \alpha + \beta + n_q - m_{\ell q})$, for $\ell = 1, \dots, k$, $q = 1, \dots, d$. Furthermore, the random parameters of each $\mu'_q$, which governs unobserved traits in subpopulation $q$, also have a tractable form. In particular, the distribution of the number of unseen features $M'$ is

$$M' \sim \text{Poisson}(\lambda'), \qquad \lambda' = \lambda \prod_{q=1}^{d} \frac{(\alpha + \beta)_{n_q}}{(\beta)_{n_q}},$$

whereas the probabilities $\theta'_{jq}$ of observing these traits are distributed as $\theta'_{jq} \overset{\text{iid}}{\sim} \text{Beta}(-\alpha, \alpha + \beta + n_q)$.

**Example 5.6** (Poisson counts, cont'd)**.** Suppose the traits are Poisson distributed and assume a gamma prior law $H(\cdot; \psi)$ with parameters $(\alpha, \beta)$, as in Example 5.4. The parameters $\theta^*_{\ell q}$ of the Poisson random variables associated to an observed $X_\ell$ are distributed as $\theta^*_{\ell q} \overset{\text{ind}}{\sim} \text{Gamma}(\alpha + m_{\ell q}, \beta + n_q)$, for $\ell = 1, \dots, k$, $q = 1, \dots, d$, where $m_{\ell q} = \sum_{i=1}^{n_q} a_{i\ell q}$. Furthermore, the distribution of the number of unseen traits $M'$ is

$$M' \sim \text{Poisson}(\lambda'), \qquad \lambda' = \lambda \prod_{q=1}^{d} \left( \frac{\beta}{\beta + n_q} \right)^{\alpha},$$

whereas the parameters $\theta'_{jq}$ of the unseen traits are distributed as $\theta'_{jq} \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta + n_q)$.

### 5.2.3 Hyperprior elicitation

When prior information for eliciting the specific values of the parameters $\psi$ and $\lambda$ is not available, a common Bayesian solution is to consider a hyperprior. In the two examples discussed so far, i.e. the binary traits and the Poisson-distributed traits, there is no closed-form characterization for the posterior of $\psi = (\alpha, \beta)$; hence one needs to recut to Markov chain Monte Carlo (MCMC) sampling. The availability of a closed-form and tractable expression for the marginal law in (5.1) is crucial in this regard, as it enables the practical implementation of Metropolis–Hastings steps.

As for $\lambda$, let us assume that $\lambda \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda)$. This choice implies a negative binomial distribution for $M$. The chosen hyperprior yields the following posterior distribution $p(\lambda \,|\, \boldsymbol{Z}) \propto \pi_n(\boldsymbol{a}; \lambda, \psi) p(\lambda; \alpha_\lambda, \beta_\lambda)$, where $\pi_n(\boldsymbol{a}; \lambda, \psi)$ is defined in Theorem 5.1, and $p(\lambda; \alpha_\lambda, \beta_\lambda)$ denotes the density of the gamma distribution with parameters $(\alpha_\lambda, \beta_\lambda)$. It turns out that this hyperprior is conjugate to our model, regardless of the choice of the likelihood $P(\cdot; \theta)$ and the prior $H(\cdot; \psi)$. Indeed, we obtain

$$\lambda \,|\, \boldsymbol{Z} \sim \text{Gamma}\left( \alpha_\lambda + k, \beta_\lambda + 1 - \prod_{q=1}^{d} \int P(0; \theta)^{n_q} H(\mathrm{d}\theta; \psi) \right). \tag{5.13}$$

The above integral simplifies in specific models, such as Examples 5.3 and 5.4. As a result, inference on the number of unseen traits, addressed via $M'$ in Theorem 5.2, is affected. In particular, under the assumptions of $M' \,|\, \lambda' \sim \text{Poisson}(\lambda')$, where $\lambda'$—defined in point (ii) of Theorem 5.2—follows a gamma distribution, we get a marginal negative binomial distribution for $M'$.

## 5.3 Latent class models with an unknown number of traits

### 5.3.1 A mixture model for trait allocations

The grouped modeling framework introduced in Section 5.2 naturally accommodates group-structured data under the assumption of homogeneity within predefined subpopulations. In practice, however, two issues may arise: (i) no prior information on a grouping structure is available, or (ii) the known partition does not adequately capture the underlying organization in the data. In such cases, one may instead seek to learn the group structure directly from the data. In the 'Ndrangheta application, affiliates can be naturally grouped according to their membership in *locali*. Nonetheless, it is important to assess whether this external information truly reflects the structure underlying attendance patterns at meetings. If the clustering inferred from the data diverges from the *locali*, this discrepancy could highlight previously unrecognized collaborations or reveal novel dynamics among members belonging to different *locali*.

To this end, in this section we consider a unknown-groups setting by embedding the model (5.4)–(5.5)–(5.6) within a Bayesian clustering framework, allowing for cluster detection while accounting for unseen traits. In essence, our goal is to estimate a partition of the subjects $Z_1, \dots, Z_n$ rather than relying on predefined groups. Let $\boldsymbol{C} = \{C_1, \dots, C_d\}$ be a partition of the statistical units $\{1, \dots, n\}$, so that $i, i' \in C_q$ if and only if subjects $i$ and $i'$ belong to the same group. We denote by $n_q = |C_q|$ the size of cluster $q$, with $\sum_{q=1}^{d} n_q = n$, where $d$ is the number of clusters. If the partition structure $\boldsymbol{C}$ is known, we recover the known-groups setting of Section 5.2. When instead $\boldsymbol{C}$ is unknown, we can assign a prior distribution. In this way, both the cluster memberships and the number of groups $d$ are random and can be learned from the data.

When the partition is unknown, the observed data take the form of an $n \times k$ matrix $\boldsymbol{A}$ with entries $A_{i\ell}$ and observed labels $X_\ell$, as in the exchangeable case. Moreover, let $\tilde{A}_{ij} \in \{0, 1, 2, \dots\}$ be the abundance of $j$th latent trait $\tilde{X}_j$ in the $i$th subject, as before. Conditionally on a partition $\boldsymbol{C} = \{C_1, \dots, C_d\}$, for a fixed trait $j$ and group $q$ we assume, as in the known-groups case, that

$$\tilde{A}_{ij} \,|\, \theta_{jq} \overset{\text{iid}}{\sim} P(\cdot; \theta_{jq}), \qquad i \in C_q.$$

The random variables are also conditionally independent across traits for $j \geq 1$ and clusters $q = 1, \dots, d$. As in equations (5.5)–(5.6), we let the total number of traits $M$ to be random and the distribution of group-specific parameters $\theta_{jq}$ being equal to

$$M \sim \text{Poisson}(\lambda), \qquad (\theta_{j1}, \dots, \theta_{jd}) \overset{\text{iid}}{\sim} H^{(d)}, \qquad j = 1, \dots, M.$$

Note that the number of groups $d$ is itself a random variable. The main difference, compared to Section 5.2, is that we now specify a prior for the partition $\boldsymbol{C}$. Among the various options, we consider the prior induced by a Pitman–Yor process (Perman et al., 1992; Pitman and Yor, 1997), which stands out for its analytical tractability and whose density is

$$\mathbb{P}(\boldsymbol{C} = \{C_1, \dots, C_d\}) = \frac{\prod_{q=1}^{d-1}(\gamma + q\sigma)}{(\gamma + 1)_{n-1}} \prod_{q=1}^{d} (1 - \sigma)_{n_q - 1}, \tag{5.14}$$

where $\gamma > -\sigma$, $\sigma \in [0, 1)$. Expression (5.14) is referred to as *exchangeable partition probability function* (Pitman, 1996). When $\sigma = 0$, the Pitman–Yor process reduces to the

partition law of a Dirichlet process (Ferguson, 1973). For a general characterization of exchangeable Gibbs-type random partitions, including the Pitman–Yor as a special case, see De Blasi et al. (2015). Recent developments on random partitions with finitely many clusters are discussed in Lijoi et al. (2008, 2020, 2024); De Blasi et al. (2013); Argiento and De Iorio (2022); Colombi et al. (2025).

We can equivalently express this model in the following hierarchical form

$$\begin{aligned} Z_i \,|\, \mu_i &\overset{\text{ind}}{\sim} \text{CP}(\mu_i), \qquad \mu_i \,|\, \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \qquad i \geq 1, \\ \tilde{p} &\sim \mathcal{P}, \end{aligned} \tag{5.15}$$

where $\tilde{p}$ is a discrete random probability measure, and $\mathcal{P}$ denotes its prior distribution. More precisely, we assume that

$$\tilde{p}(\cdot) = \sum_{h \geq 1} \xi_h \delta_{\eta_h}(\cdot),$$

where $(\xi_h)_{h\geq 1}$ is a sequence of random probability weights summing to 1, and $(\eta_h)_{h\geq 1}$ is a sequence of discrete random measures of the form $\eta_h = \sum_{j=1}^{M} \phi_{jh} \delta_{\tilde{X}_j}$; these two sequences are assumed to be independent. The discreteness of $\tilde{p}$ implies that there is a positive probability that $\mu_i$ and $\mu_{i'}$ are be identical, inducing clustering among the observations $Z_1, \ldots, Z_n$. We denote with $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ the $d$ distinct random measures among $\mu_1, \ldots, \mu_n$. The random weights $\xi_h$ represent the clustering probabilities, that is $\xi_h$ is the probability that $\mu_i$ is drawn from $h$th component of $\tilde{p}$, that is $\mu_i = \eta_h$. The weights $\xi_h$ follow the stick-breaking weights of the Pitman-Yor process so that $\xi_h = V_h \prod_{r=1}^{h-1}(1 - V_r)$, with $V_h \overset{\text{ind}}{\sim} \text{Beta}(1 - \sigma, \gamma + h\sigma)$, where we agree that $\xi_1 = V_1$. Moreover, we characterize the law of the sequence of $(\eta_h)_{h\geq 1}$ through its finite-dimensional distributions: we assume that any $d$-dimensional subset of $(\eta_h)_{h\geq 1}$, denoted with $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_d)$, is distributed as $\tilde{\boldsymbol{\eta}} \sim \text{FCRV}(H^{(d)}, \lambda, G_0)$. This means, in particular, that the distinct measures are distributed as $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d) \sim \text{FCRV}(H^{(d)}, \lambda, G_0)$. The representation in (5.15) helps clarify a crucial aspect of the model, as we now discuss. By employing an exchangeable prior for the latent partition, we are implicitly returning to an exchangeable framework

$$\begin{aligned} Z_i \,|\, \tilde{p} &\overset{\text{iid}}{\sim} \int \text{CP}(\mu) \tilde{p}(\mathrm{d}\mu) \\ \tilde{p} &\sim \mathcal{P}, \end{aligned} \tag{5.16}$$

where the subjects $Z_i$ are conditionally i.i.d. draws from a *process of counts mixture*. However, this model is far more flexible than the one considered in Section 5.1, where $Z_i \,|\, \tilde{\mu} \overset{\text{iid}}{\sim} \text{CP}(\tilde{\mu})$. Indeed, under model (5.16) we have that $Z_i(\cdot) = \sum_{j=1}^{M} \tilde{A}_{ij} \delta_{\tilde{X}_j}(\cdot)$ and

$$\mathbb{P}(\tilde{A}_{i1} = \tilde{a}_1, \ldots, \tilde{A}_{iM} = \tilde{a}_M \,|\, \tilde{p}) = \sum_{h \geq 1} \xi_h \prod_{j=1}^{M} P(\tilde{a}_j; \phi_{jh}), \qquad i \geq 1,$$

where the random vectors $(\tilde{A}_{i1}, \ldots, \tilde{A}_{iM})$ for $i \geq 1$ are conditionally i.i.d. from the above law. The prior for $\tilde{p}$ specifies that $M \sim \text{Poisson}(\lambda)$, $\tilde{X}_j \overset{\text{iid}}{\sim} G_0$, the weights $(\xi_h)_{h\geq 1}$ follow the Pitman–Yor stick-breaking distribution, and that the parameters $\phi_{jh}$, under the factorized structure $H^{(d)} = H(\cdot; \psi) \times \cdots \times H(\cdot; \psi)$, are i.i.d. draws $\phi_{jh} \overset{\text{iid}}{\sim} H(\cdot; \psi)$ for $j = 1, \ldots, M$ and $h \geq 1$. Compared to Section 5.1, we note that model (5.16), given the random

probability measure $\tilde{p}$, introduces *dependence* across traits thereby significantly enhancing the flexibility of the likelihood component of the model.

**Remark 5.2.** *The Bayesian model we defined in equation* (5.16) *is related to a classical statistical approach for modeling multivariate discrete data, known as latent class analysis (Lazarsfeld and Henry, 1968; Goodman, 1974; Hagenaars and McCutcheon, 2002). The more recent work of Dunson and Xing (2009) provides a Bayesian nonparametric extension of this framework, making it even more closely related to ours. However, all these approaches assume that $M$ is a known constant, meaning that all traits are known in advance, including those equal to zero for all subjects. While this may be a reasonable assumption in many contexts, such as when the number of variables is fixed, it may be untenable in others. Ignoring unseen traits can skew the analysis and directly affect clustering, as we now show.*

### 5.3.2 Effect of accounting for potentially unseen traits

To assess the impact of accounting for potentially unseen traits in the sample, we compare the predictive allocation probabilities for a generic subject $i$ under the proposed model defined in (5.15), with the Pitman-Yor prior for $\xi_h$, and a naïve model almost identical to (5.15), in which it is assumed that there are no unseen traits, i.e. we suppose $M = k$. Specifically, let

$$p_{iq} = \mathbb{P}(\mu_i = \tilde{\mu}_q \,|\, \boldsymbol{Z}, \boldsymbol{\mu}_{-i}), \qquad p_{i,\text{new}} = \mathbb{P}(\mu_i = \text{“new”} \,|\, \boldsymbol{Z}, \boldsymbol{\mu}_{-i}), \qquad i = 1, \ldots, n,$$

where $\boldsymbol{\mu}_{-i} = (\mu_1, \ldots \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n)$ with $d_{-i}$ distinct values $\tilde{\mu}_1, \ldots, \tilde{\mu}_{d_{-i}}$, for $q = 1, \ldots, d_{-i}$. Thus each $p_{iq}$ denotes the predictive probability that subject $i$ belongs to the $q$th cluster formed by the remaining subjects, and $p_{i,\text{new}}$ is the predictive probability for subject $i$ to form their own cluster. We similarly define $p^*_{iq}$ and $p^*_{i,\text{new}}$ for the naïve model in which it is wrongly assumed that all traits are observed with $M = k$.

**Proposition 5.1.** *The predictive allocation probabilities for a generic subject $i$ under model* (5.16) *and a naïve model in which it is assumed that there are no unseen traits, with $M = k$, satisfy the inequality:*

$$\frac{p_{i,\text{new}}}{p_{iq}} < \frac{p^*_{i,\text{new}}}{p^*_{iq}}, \qquad q = 1, \ldots, d_{-i}, \quad i = 1, \ldots, n.$$

Refer to Section 5.C.1 of the Appendix for the proof. The inequality in Proposition 5.1 shows that, once potentially unseen traits are taken into account, the probability of allocating new clusters decreases compared to the naïve model. The practical implications are substantial. In the naïve model, which ignores unseen traits, the likelihood of assigning new subjects to their own clusters is systematically overestimated, leading to a fragmented view of the data and less interpretable clusters. By contrast, our proposed model incorporates the role of unseen traits, which tends to stabilize cluster allocation probabilities. This adjustment reflects a more cautious and principled approach to inference, acknowledging that the possibility of unseen traits carries essential information. We emphasize that Proposition 5.1 is not limited to the Pitman–Yor specification; rather, it applies to any prior on the set of weights $(\xi_h)_{h\geq 1}$ of the discrete random probability measure $\tilde{p}$ in

model (5.16). In Section 5.4.1, we numerically illustrate a simple scenario where the discrepancy highlighted in Proposition 5.1 has a substantial impact on the inference process, demonstrating the importance of accounting for unseen traits in practical applications.

### 5.3.3 Gibbs sampling and update of the clustering structure

Posterior inference for mixture model (5.16) can be efficiently tackled via a simple marginal Markov Chain Monte Carlo (MCMC) algorithm. The procedure is greatly facilitated by the availability of a closed-form expression for the pETPF, which is given in Theorem 5.1.

The primary goal of the the Gibbs sampling is the approximation of posterior distribution of the latent clustering structure $\boldsymbol{C}$. The probability distribution on $(\xi_h)_{h\geq 1}$ determines the *a priori* predictive allocation probabilities of the sampling model, which describe the stochastic generation of $\mu_1, \ldots, \mu_n$ as well as their clustering structure $\boldsymbol{C}$. Specifically, given the previously observed measures $\mu_1, \ldots, \mu_n$, with $d$ distinct values $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ and frequencies $n_1, \ldots, n_d$, a Pitman–Yor process prescribes that, a priori, we have

$$\mathbb{P}(\mu_{n+1} = \tilde{\mu}_q \,|\, \mu_1, \ldots, \mu_n) = \frac{n_q - \sigma}{n + \gamma}, \qquad \mathbb{P}(\mu_{n+1} = \text{“new”} \,|\, \mu_1, \ldots, \mu_n) = \frac{\gamma + d\sigma}{n + \gamma}.$$

This well-known sequential mechanism induces a partition of the statistical units. In practice, however, we do not explicitly compute the random measures $\mu_1, \ldots, \mu_n$. If the focus is on the clustering structure, we only need to keep track of the labels associated with $\mu_1, \ldots, \mu_n$ and the implied partition. The posterior distribution of the clustering $\boldsymbol{C}$ under the mixture model (5.16) is obtained through a Gibbs sampling procedure, by iteratively updating the labels of $\mu_i \,|\, \boldsymbol{Z}, \boldsymbol{\mu}_{-i}$ for $i = 1, \ldots, n$ according to the previously defined full conditional allocation probabilities $p_{iq} = \mathbb{P}(\mu_i = \tilde{\mu}_q \,|\, \boldsymbol{Z}, \boldsymbol{\mu}_{-i})$ and $p_{i,\text{new}} = \mathbb{P}(\mu_i = \text{“new”} \,|\, \boldsymbol{Z}, \boldsymbol{\mu}_{-i})$, where $\boldsymbol{\mu}_{-i} = (\mu_1, \ldots \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n)$ comprises $d_{-i}$ distinct values $\tilde{\mu}_1, \ldots, \tilde{\mu}_{d_{-i}}$ with frequencies $n_{q,-i}$ for $q = 1, \ldots, d_{-i}$ and forms a partition $\boldsymbol{C}_{-i}$. In addition, the full conditional clustering probabilities $p_{iq}$ and $p_{i,\text{new}}$ have a simple expression, obtained by combining the a priori Pitman–Yor sequential scheme with the marginal likelihood:

$$p_{iq} \propto \frac{n_{q,-i} - \sigma}{n + \gamma - 1}\, \pi_n(\boldsymbol{a}_{iq}; \lambda, \psi), \qquad p_{i,\text{new}} \propto \frac{\gamma + d_{-i}\sigma}{n + \gamma - 1}\, \pi_n(\boldsymbol{a}_{i,\text{new}}; \lambda, \psi), \tag{5.17}$$

where $\boldsymbol{a}_{iq}$ denotes the observed trait values organized according to the partition $\boldsymbol{C}_{-i}$, with subject $i$ allocated to cluster $q$. Similarly, $\boldsymbol{a}_{i,\text{new}}$ corresponds to the trait values organized according to the partition $\boldsymbol{C}_{-i}$ when subject $i$ is assigned to a new cluster. This updating scheme is particularly straightforward thanks to the closed-form expression of $\pi_n(\boldsymbol{a}; \lambda, \psi)$ provided in Theorem 5.1. In a model with binary traits and beta priors described in Example 5.3 and a model with Poisson counts and gamma priors described in Example 5.4, the evaluation of $\pi_n(\boldsymbol{a}; \lambda, \psi)$ is highly efficient, resulting in a fast marginal sampler. Essentially, this procedure can be seen as a variant of Algorithm 3 in Neal (2000).

When hyperpriors are placed on the parameters $\lambda$ and $\psi$, the MCMC scheme described above is coupled with updates of $\lambda$ and $\psi$ from their respective full conditional distributions, i.e. equation (5.13) in the case of $\lambda$. Exact sampling from the full conditional of $\psi$ can also be replaced by a Metropolis-within-Gibbs step for $\psi$.

## 5.4 Simulation studies

### 5.4.1 Assessing the impact of unseen traits in clustering

Before moving to more structured simulation studies, we first provide empirical evidence on the discrepancy highlighted in Proposition 5.1, which may indeed lead to substantial inferential differences. To this end, we consider a simulated binary-outcomes example, where we compare the inference obtained from our proposed model (5.16) with that of a latent class model in which $M = k$ is fixed, thus disregarding the possibility of unseen traits, as in Dunson and Xing (2009). We consider data organized into $d = 5$ groups, with a total of $M = 500$ traits that may potentially be observed. Group-specific probability vectors $\theta_{jq}$ are generated by drawing i.i.d. values from a Beta$(0.1, 10)$ distribution, for $q = 1, \ldots, 5$ and $j = 1, \ldots, 500$. Subject-specific binary vectors are then obtained by independently sampling Bernoulli random variables $\tilde{A}_{ijq}$ with success probabilities $\theta_{jq}$, determined by the group allocation. Samples are drawn from the five groups with different sizes, namely $n_1 = 100$, $n_2 = 60$, $n_3 = 40$, $n_4 = 20$, and $n_5 = 20$, where $n_q$ denotes the sample size of group $q$. In the resulting dataset, $K_n = k = 293$ traits are observed out of the total $M = 500$.

We compare model (5.16) with binary traits, as in Example 5.3, with the naïve model that disregards unseen traits. To ensure a fair comparison, both models are fitted using identical hyperparameters, set to their oracle values. Specifically, the parameters of the Beta prior $H(\cdot; \psi)$ are fixed at $\psi = (0.1, 10)$, while the Poisson prior on $M$ is specified with $\lambda = 500$. For the clustering prior, we adopt a Dirichlet process with concentration parameter 1, corresponding to a Pitman–Yor process with parameters $\sigma = 0$ and $\gamma = 1$. The MCMC algorithm is run for 5,000 iterations, discarding the first 500 as burn-in, and applying thinning every 2 iterations. Figure 5.4.1 displays the posterior distribution of the number of clusters under the two models. Consistent with the analytical result in Proposition 5.1, our proposed model (5.16) yields fewer clusters than the naïve specification that ignores unseen traits. The discrepancy is substantial in this example, with posterior modes equal to 6 and 10, respectively.

### 5.4.2 Synthetic network data

This section evaluates the performance of the latent class model with unseen traits described in Section 5.3. Motivated by our focus on criminal network analysis, we generate synthetic datasets designed to capture a realistically complex structure. The analysis of criminal networks (Legramanti et al., 2022; Lu et al., 2025) typically aims at detecting groups of affiliates who share similar connectivity patterns. These patterns are encoded in a weighted adjacency matrix, which records the frequency of interactions between pairs of individuals. An interaction is defined as the co-attendance of two affiliates at the same meeting. The weighted adjacency matrix thus summarizes the raw data, originally collected as multivariate binary observations, where each affiliate is associated with the list of meetings they attended. Formally, if the dataset comprises $n$ affiliates and $k$ distinct meetings, the attendance information may be arranged in an $n \times k$ binary matrix $\boldsymbol{A}$, with entries $A_{i\ell}$ indicating whether individual $i$ attended meeting $\ell$. The corresponding observed weighted adjacency matrix, denoted by $\boldsymbol{W}$, is then obtained as $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{A}^T$. Its generic

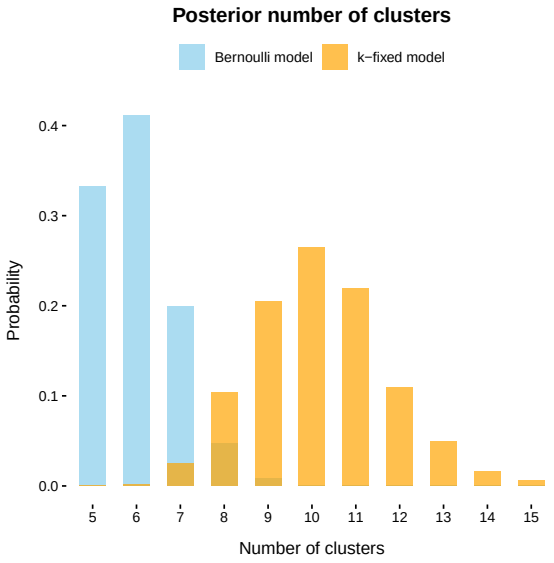



Figure 5.4.1: Posterior distribution of the number of clusters under model (5.16) with binary traits (in skyblue) and Dunson and Xing (2009) model (in orange). The true number of clusters in the generating mechanism is 5.

entry $W_{ii'}$ represents the number of meetings jointly attended by affiliates $i$ and $i'$, that is, $W_{ii'} = \sum_{\ell=1}^{k} A_{i\ell}A_{i'\ell}$ for $i \neq i'$. Our framework induces a probabilistic structure on $\boldsymbol{W}$: each entry $W_{ii'}$ can be interpreted as the realization of a sum of Bernoulli random variables. To assess the flexibility of the proposed methodology in the analysis of criminal networks, we consider two simulated scenarios, each characterized by distinct properties of the resulting weighted adjacency matrix. These scenarios are constructed to produce structures akin to those investigated in Lu et al. (2025), thereby closely reflecting patterns observed in real-world criminal networks. A detailed description of the two scenarios is provided below.

For both scenarios, the analysis is carried out using the binary traits specification of the mixture model (5.16), referred to as the unknown-groups model, where $H(\cdot;\psi)$ is the beta distribution with parameters $(-\alpha, \alpha+\beta)$, with $\psi = (\alpha, \beta)$. Prior distributions for the model parameters $\lambda$, $-\alpha$, and $\alpha+\beta$ are specified as follows. The parameter $\lambda$, governing the Poisson-distributed total number of meetings, is assigned a gamma prior with parameters $(\alpha_\lambda, \beta_\lambda)$, as detailed in Section 5.2.3. The hyperparameters $(\alpha_\lambda, \beta_\lambda)$ are chosen so that the prior expected value of $M$ equals $\hat{M} = 1.5k$, where $k$ denotes the observed number of meetings in the sample, and the prior variance of $M$ is set to $10\hat{M}$. For the parameters $(a, b) = (-\alpha, \alpha+\beta)$ of the beta distribution $H$, independent gamma priors are assumed with hyperparameters $(\alpha_a, \beta_a)$ and $(\alpha_b, \beta_b)$, respectively. Specifically, $(\alpha_a, \beta_a)$ are set to induce a prior mean of 0.2 for $a$ with a large variance, while $(\alpha_b, \beta_b)$ are chosen so that the prior mean of $b$ is 10, also with high variance. Additionally, the clustering structure is modeled using a Dirichlet process with concentration parameter equal to $\gamma = 1$. Posterior inference relies on 10,000 iterations of the MCMC algorithm described in Section 5.3.3, with the first 1,000 samples discarded as burn-in and a thinning interval of 2.

We benchmark our approach against the *negative binomial mixture of* BBs introduced in Chapter 3. This represents the natural competitor for estimating the number of unseen

traits, though it is specifically designed for homogeneous exchangeable settings. Additional details about this competitor, hyperparameter choices and MCMC settings are provided in Section 5.D of the Appendix. Importantly, by construction, the negative binomial mixture of BBs enforces exchangeability, implying that its posterior expectation for $\boldsymbol{W}$ assigns identical values to all entries. As a consequence, the model is unable to capture possible heterogeneity in connectivity patterns that may be revealed by the observed adjacency matrix.

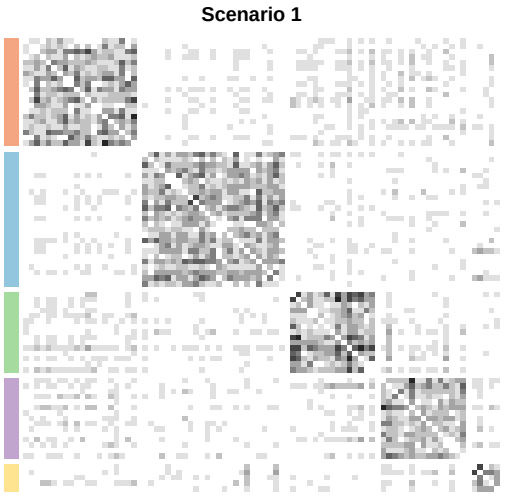
**Scenario 1.** Data are generated with a total of $M = 500$ meetings and $d = 5$ criminal groups. Each group is associated with a core set of 15 meetings, in which affiliates have a higher probability of being observed (0.3). Beyond these core meetings, each group is randomly assigned 300, 125, 50, and 10 additional meetings, with corresponding probabilities of 0.002, 0.01, 0.05, and 0.3. The total number of criminals is $n = 80$, partitioned into groups of size $n_1 = 20$, $n_2 = 25$, $n_3 = n_4 = 15$, and $n_5 = 5$. The number of observed meetings in the sample is $k = 304$. Figure 5.4.2a shows a graphical representation of the resulting adjacency matrix.

Posterior inference under the proposed unknown-groups model demonstrates strong recovery of the underlying structure. The posterior distribution of the number of clusters (not shown) correctly concentrates on the true value $d = 5$, confirming the ability of the model to detect the latent group partition. Beyond validating the clustering estimates, Figure 5.4.2c shows the posterior expectation of the weighted adjacency matrix, which can be visually compared to its simulated counterpart in Figure 5.4.2a. The close alignment between these matrices demonstrates the model's ability to recover the connectivity structure of the data. Notably, posterior estimates of the weighted adjacency matrix are straightforward to compute, leveraging the closed-form posterior expressions in Theorem 5.2 and their specializations for the binary traits model.

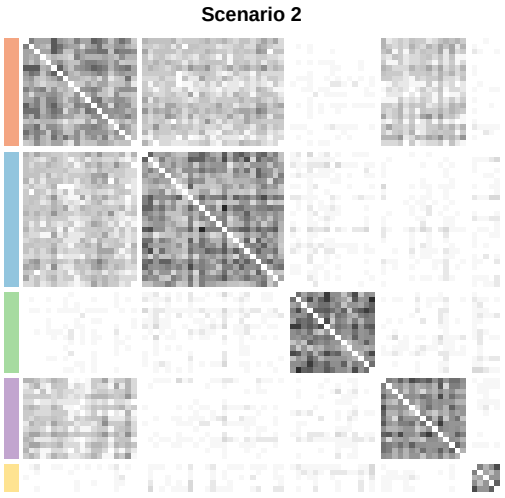
We further compare our methodology to the negative binomial mixture of BBs introduced in Chapter 3. Both models allow inference on the number of unseen meetings $M'$, which in the framework of Theorem 5.2 corresponds to the number of atoms $M'$ in the measures $\mu'_q$. Figure 5.4.3 reports the posterior distributions of $M'$ for the two models. Our proposed model provides an accurate estimate, with uncertainty appropriately quantified. In contrast, the negative binomial mixture of BBs substantially overestimates $M'$, failing to address the unseen features problem in this heterogeneous setting.

Model comparison based on the WAIC supports the same conclusion. The proposed unknown-groups model yields a WAIC of 68546, while the negative binomial mixture of BBs yields a much larger value of 923830.

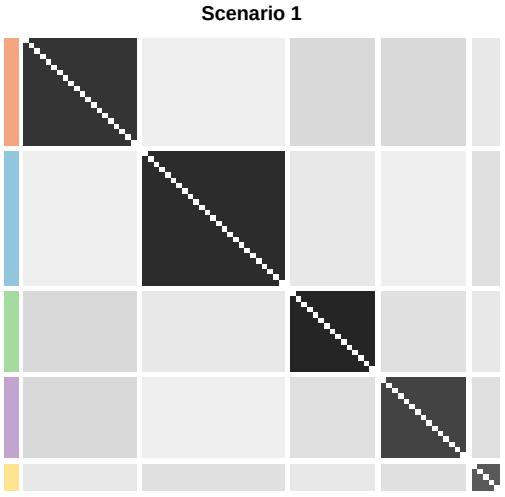
**Scenario 2.** This scenario also considers $M = 500$ meetings and $d = 5$ criminal groups, but introduces more heterogeneous and asymmetric connectivity patterns. Each group is associated with 250, 150, 50, and 50 meetings, with observation probabilities of 0.002, 0.01, 0.05, and 0.4, respectively. To induce additional overlap, 20 meetings have the highest probability (0.4) of being attended by criminals from both groups 1 and 2, while another distinct set of 20 meetings has the highest probability (0.4) for both groups 1 and 4. In general, all high-probability meetings do not overlap between different group pairs. As in Scenario 1, the total number of criminals is $n = 80$, partitioned into groups of size $n_1 = 20$, $n_2 = 25$, $n_3 = n_4 = 15$, and $n_5 = 5$. The number of observed meetings is $k = 340$. Figure 5.4.2b illustrates the resulting adjacency matrix.



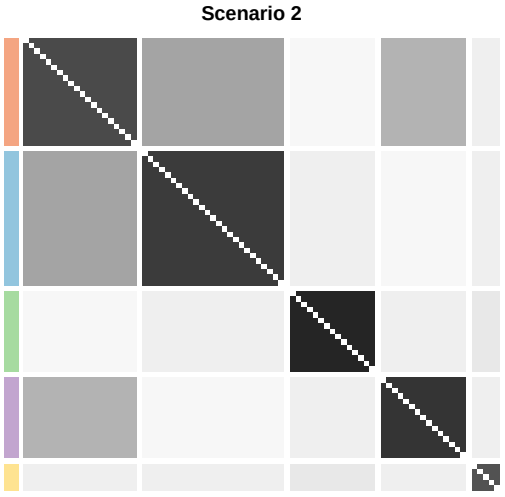

(a) Simulated weighted adjacency matrix in Scenario 1. Side colors represent the true clustering.




(b) Simulated weighted adjacency matrix in Scenario 2. Side colors represent the true clustering.

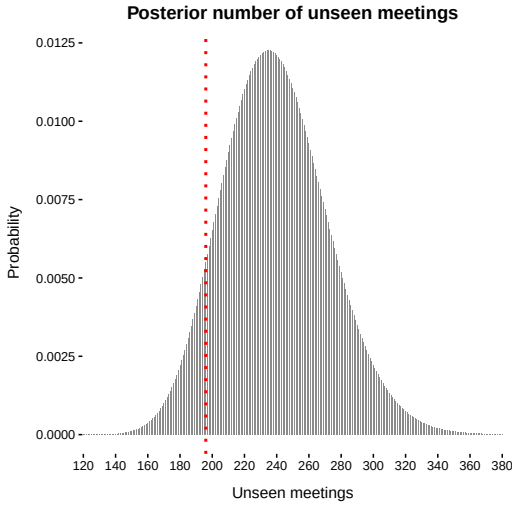



(c) Posterior expectation of the weighted adjacency matrix in Scenario 1. Side colors represent the estimated clustering.

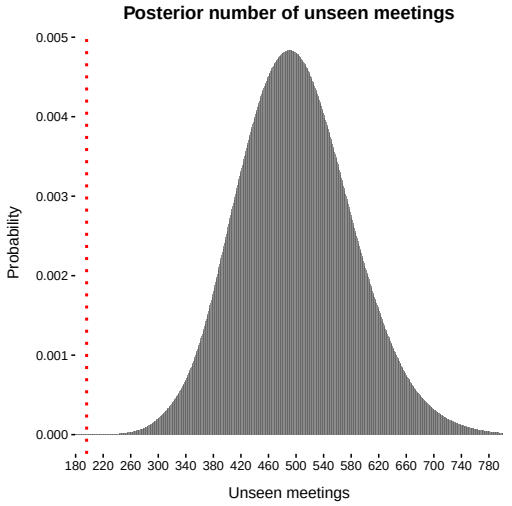



(d) Posterior expectation of the weighted adjacency matrix in Scenario 2. Side colors represent the estimated clustering.

Figure 5.4.2: Synthetic network data. Simulated and estimated weighted adjacency matrices under the two scenarios. Top row: simulated adjacency matrices, with side colors denoting the true clustering. Bottom row: posterior expectations of the adjacency matrices under the proposed model, with side colors denoting the estimated clustering. In all panels, the color of each entry ranges from white to black as the number of co-attended meetings increases.




(a) Unseen meetings under the unknown-groups model.



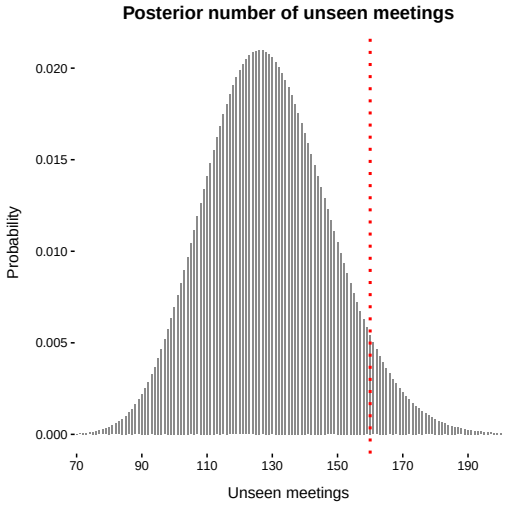

(b) Unseen meetings under the negative binomial mixture of BBs.

Figure 5.4.3: Synthetic network data, Scenario 1. Posterior distribution of the number of unseen meetings under (a) the unknown-groups model and (b) the negative binomial mixture of BBs in Chapter 3. The vertical red dotted lines denote the true number of unseen meetings.

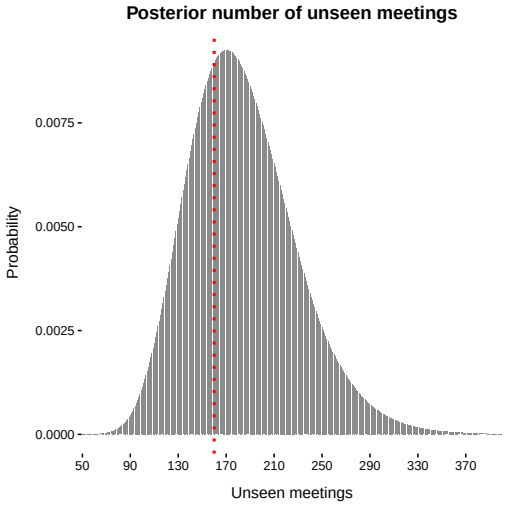
As in Scenario 1, the proposed unknown-groups model correctly identifies the true number of clusters $d = 5$. The corresponding clustering estimates are reported in Figure 5.4.2d, alongside the posterior expectation of the weighted adjacency matrix. A visual comparison with the simulated matrix in Figure 5.4.2b confirms that the method accurately recovers the heterogeneous connectivity patterns.

We next examine inference on the number of unseen meetings $M'$. Figure 5.4.4 presents the posterior distribution of $M'$ under the proposed unknown-groups model and the negative binomial mixture of BBs in Chapter 3. Figure 5.4.4a shows that our unknown-groups model recovers the number of unseen meetings accurately, while also providing a coherent quantification of uncertainty. At first glance, the competitor appears to perform well, as its posterior distribution is centered closer to the true value (red vertical line), albeit with substantially greater dispersion. However, this apparent accuracy is largely coincidental. In fact, because the negative binomial mixture of BBs enforces homogeneity, it cannot accommodate the heterogeneity clearly visible in the simulated adjacency matrix (Figure 5.4.2b). For ease of comparison, Figure 5.4.5 juxtaposes the simulated adjacency matrix, the posterior expectation of the adjacency matrix under our unknown-groups model, and the posterior expectation of the adjacency matrix under the competitor. The figure highlights that while our model reproduces the complex connectivity structure, the competitor collapses to an unrealistic homogeneous pattern.

This conclusion is further supported by a model comparison based on the WAIC. The proposed unknown-groups model achieves a WAIC of $-19107$, while the negative binomial mixture of BBs yields a much larger value of 217142. Such a difference provides strong evidence that the unknown-groups model offers a substantially better fit for these data.

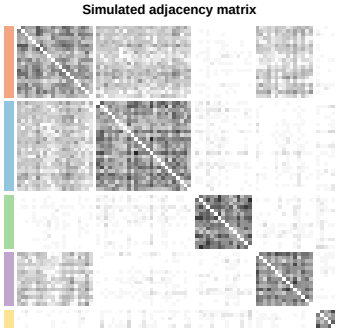



(a) Unseen meetings under the unknown-groups model.

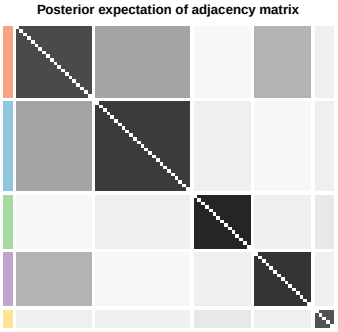



(b) Unseen meetings under the negative binomial mixture of BBs.

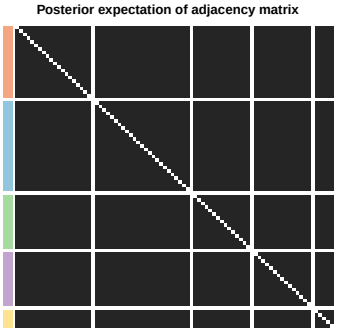
Figure 5.4.4: Synthetic network data, Scenario 2. Posterior distribution of the number of unseen meetings under (a) the unknown-groups model and (b) the negative binomial mixture of BBs in Chapter 3. The vertical red dotted lines denote the true number of unseen meetings.




(a) Simulated weighted adjacency matrix.




(b) Posterior expectation of weighted adjacency matrix under the proposed unknown-groups model.




(c) Posterior expectation of weighted adjacency matrix under the negative binomial mixture of BBs.

Figure 5.4.5: Synthetic network data, Scenario 2. Simulated weighted adjacency matrix and its estimates under the competing models.

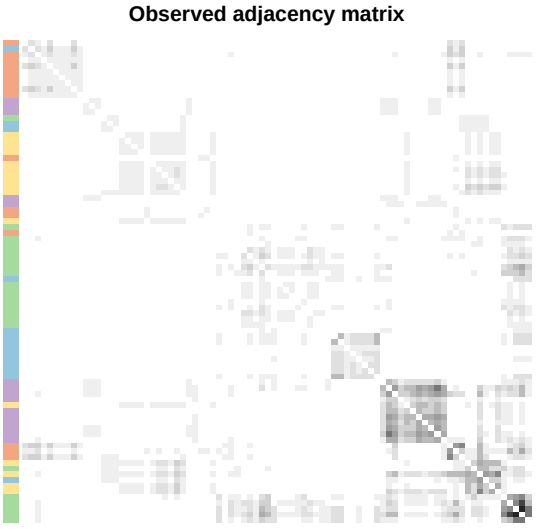
## 5.5 Analysis of the Infinito network

In this section, we analyze the *Infinito* 'Ndrangheta network (Legramanti et al., 2022; Lu et al., 2025). These data stem from *Operazione Infinito* (Calderoni et al., 2017), a six-year-long, large-scale law enforcement operation aimed at monitoring and dismantling the core structure of *La Lombardia*, the highly pervasive branch of the 'Ndrangheta Mafia operating in the Milan area. The raw data collected during the investigation are publicly available at https://sites.google.com/site/ucinetsoftware/datasets/covert-networks in the form of multivariate binary outcome observations, making them well-suited for analysis using the binary traits specification of our proposed model.

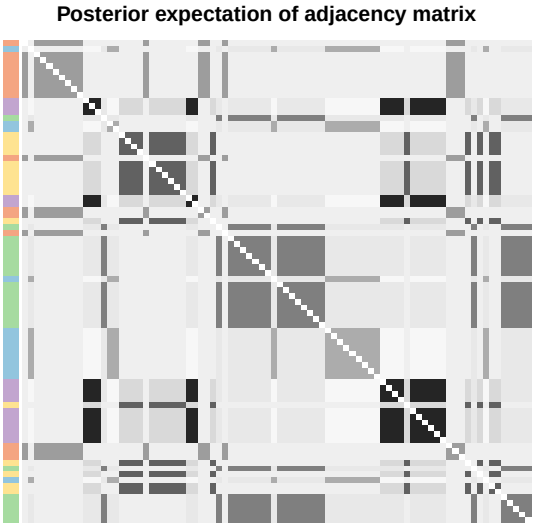
As in the simulation studies, the attendance of the $n$ affiliates at the $k$ distinct meetings can be represented by the $n \times k$ binary matrix $\boldsymbol{A}$, where each entry $A_{i\ell} \in \{0, 1\}$ indicates whether individual $i$ attended meeting $\ell$. A graphical representation of the observed weighted adjacency matrix $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{A}^T$, whose generic entry $W_{ii'}$ represents the number of meetings jointly attended by affiliates $i$ and $i'$, is shown in Figure 5.5.1a. Legramanti et al. (2022); Lu et al. (2025) directly modelled the matrix $\boldsymbol{W}$ without considering the information of the binary traits $A_{i\ell}$. Instead, we directly model $A_{i\ell}$, which will induce a model for $\boldsymbol{W}$ as a by-product, therefore making use of all the available information. We focus on the $d = 5$ most populated *locali*, with sizes $n_1 = 16, n_2 = 14, n_3 = 23, n_4 = 15, n_5 = 16$, of the *Infinito* dataset, for a total of $n = 84$ affiliates and $k = 44$ recorded meetings among them. The affiliates' locale membership provides a natural partition into five groups, making it reasonable to begin the analysis by leveraging this a priori known clustering. This directly motivates the known-groups framework described in Section 5.2. In particular, we consider model (5.4) for binary traits, employing the same hyperparameters and MCMC settings as in the simulation study of Section 5.4.2. The only difference here is that the clustering structure is not inferred, so the sampling reduces to a vanilla MCMC scheme.

Figure 5.5.1b describes the estimated connectivity patterns among different *locali* through the posterior expected weighted adjacency matrix, which we obtained exploiting the closed-form results of Theorem 5.2. As evident by comparing Figure 5.5.1a with Figure 5.5.1b, this estimated adjacency matrix captures the most significant patterns but still partially disagree with the observed one, indicating that the partition induced by *locali* membership is insufficient to describe the underlying connectivity patterns within the criminal organization. This points to more complex connections, extending beyond the *locali* structure. Identifying and analyzing these hidden structures could provide insights into the organizational strategies of the criminal network. For completeness, Figure 5.5.2b reports the inference on the number of meetings that went undetected by investigators; however, the poor fit of the model to data suggests not to rely on this estimate.

All these considerations motivate the necessity of moving to the framework described in Section 5.3, where the clustering structure among the affiliates is learnt from the data. In particular, we consider the binary traits specification of model (5.16), referred to as the unknown-groups model, with the same hyperparameters and MCMC settings as in the simulation study of Section 5.4.2. Figure 5.5.2a displays the posterior distribution of the number of clusters among affiliates. To obtain a point estimate of the clustering structure, we follow Wade and Ghahramani (2018) and select the configuration that minimizes the posterior expectation of the Variation of Information (VI) metric (Meilă, 2003). The re-




(a) Observed adjacency matrix with arbitrary ordering so as to match the ordering in panel (c). The row colors indicate the affiliates' membership to the five analyzed *locali*.

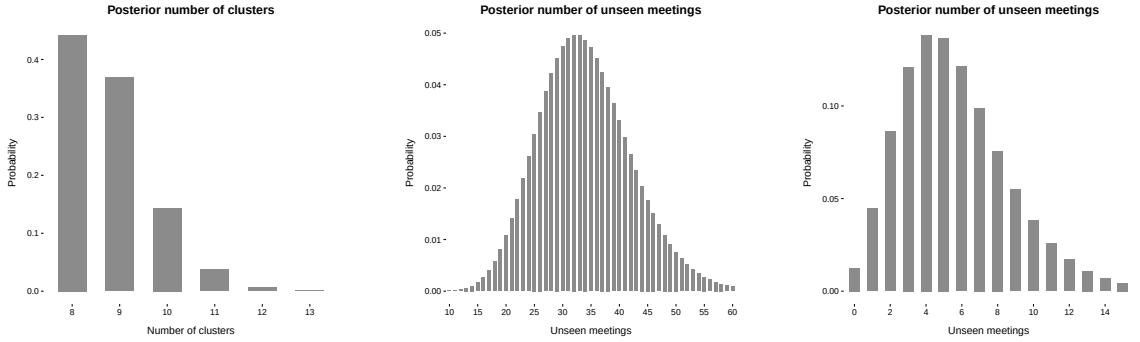



(b) Posterior expectation under a known-groups specification with partition induced by *locali* membership. Ordering matches panels (a) and (c).


(c) Observed adjacency matrix, ordered and partitioned according to the VI clustering. Side colors correspond to the *locali*, with darker/-lighter shades denoting bosses/affiliates.


(d) Posterior expectation under an unknown-groups model with groups learnt from the data. Ordering matches panel (c).

Figure 5.5.1: Comparison of the *Infinito* network with the estimated models. Top row: observed adjacency matrix and posterior expectation under the known-groups model. Bottom row: observed adjacency matrix and posterior expectation under the unknown-groups model. In all panels, the color of each entry ranges from white to black as the number of co-attended meetings increases (observed on the left, expected on the right).




(a) Number of clusters among the affiliates.

(b) Unseen meetings: partition induced by *locali* membership.

(c) Unseen meetings: partition estimated from the data.

Figure 5.5.2: Posterior distributions under the proposed models: (a) number of clusters among the affiliates; (b) number of unseen meetings under a known-groups specification with partition induced by *locali* membership; (c) number of unseen meetings under an unknown-groups model with clustering learnt from the data.

sulting clustering consists of eight unbalanced clusters, represented by the blocks in the colored bar of Figure 5.5.1c. The row colors indicate the affiliates' membership to the five *locali*, as in Figure 5.5.1a, with the additional detail of darker shades denoting bosses and lighter tones representing lower-ranked affiliates, helping for the discussion of the estimated clustering. As evident from Figure 5.5.1, the clustering structure learnt from the data significantly deviates from the one induced by *locali* membership, and the posterior expected weighted adjacency matrix in Figure 5.5.1d nicely aligns with the observed one. These considerations qualitatively support estimating the clustering structure from the data, rather than relying on the partition induced by the *locali* membership. To strengthen this argument, we provide a complementary quantitative assessment. Following standard model selection practices, we compare the WAIC (Watanabe-Akaike Information Criteria) (Watanabe, 2013) between the two models. Specifically, the unknown-groups model yields a WAIC of $-7236$, whereas the known-groups model results in a WAIC of $-6573$, strongly indicating a preference for data-driven clustering estimation.

Consistent with the findings of Lu et al. (2025), the inferred group structures reveal intricate dynamics within this criminal organization. Notably, there is a clear tendency for clusters to form within individual *locali*, though the mechanisms governing group formation vary between affiliates and bosses. This pattern suggests a strategic design aimed at ensuring resilience across different levels of the organization. Furthermore, it highlights that while lower-ranked affiliates can form peripheral groups with relative ease, the formation of core groups, comprising all the bosses, is more constrained, as reflected in their progressively smaller sizes. A deeper examination of the cluster structure uncovers additional meaningful insights, further supporting observations made in Lu et al. (2025). For instance, juridical documents indicate that a specific affiliate from the blue locale was attempting to establish a new locale. This is reflected in our clustering results, where this individual is grouped primarily with lower-ranked members from the red locale (first cluster from the top), suggesting a deliberate strategy of recruiting new affiliates from that locale. Another striking pattern, even more evident in our analysis than in Lu et al. (2025), involves a

cluster composed of bosses from multiple *locali*. Specifically, one green and one blue boss cluster together with two yellow bosses and two lower-ranked yellow affiliates. This aligns with reports that the green and blue bosses supported an unsuccessful attempt to increase the independence of the 'Ndrangheta group in Lombardy from the ruling families in Calabria. Consequently, these individuals sought to distance themselves from their original *locali* and strengthen ties elsewhere. Our inferred clustering suggests that this shift primarily moved in the direction of the yellow locale, a hypothesis supported by the cluster composition. In contrast, Lu et al. (2025) infer this movement indirectly through node positioning in a distance-based network representation, rather than as a direct outcome of their methodology. A further distinction between our results and those in Lu et al. (2025) concerns the clustering of most yellow and purple lower-ranked affiliates. While our model merges them into a single cluster, their approach identifies them as distinct groups. However, their distance-based network representation highlights their close similarity. Finally, it is noteworthy that core groups of bosses frequently include a few lower-ranked affiliates, suggesting that these individuals may hold more central roles than indicated in the juridical documents. As observed in Lu et al. (2025), this is particularly evident in the case of a key affiliate grouped with the green locale bosses. In reality, this individual is a high-ranking member with crucial coordinating responsibilities across multiple *locali*.

From inspecting the posterior expected weighted adjacency matrix in Figure 5.5.1d, additional insights on the connectivity patterns among different groups can be obtained. The darker blocks along the diagonal corresponding to boss-dominated clusters indicate that these small groups tend to meet frequently and primarily within their own ranks. Additionally, the matrix highlights regular meetings among all boss-dominated groups, likely reflecting their need for coordinated decision-making. An exception to this pattern is the core cluster of the blue locale, which interacts only with the core cluster of the green locale. Furthermore, as expected, core groups within each locale exhibit strong connectivity with clusters of lower-ranked affiliates from the same locale, reinforcing internal hierarchies. Notably, our earlier hypothesis regarding the cluster of bosses spanning multiple *locali*, suggesting a gradual shift towards the yellow locale, gains further support. This group exhibits a higher frequency of meetings with the cluster containing all yellow lower-ranked affiliates, indicating its strategic alignment and possible integration into the yellow locale.

A key property of both the known-groups and unknown-groups models is that they explicitly allow for the estimation of the number of meetings that went undetected by investigators. In the analysis of the *Infinito* criminal data, we rely on the inference provided by the unknown-groups model, which was identified as the better-fitting specification based on the comparisons discussed above. Our unknown-groups model suggests the likely presence of a few number of undetected meetings, which aligns with expectations and confirms the need of accounting for unseen binary traits. Figure 5.5.2c displays the posterior distribution of the number of these unseen meetings, providing a probabilistic estimate of their occurrence. This information is useful on its own as it may provide guidance to law enforcements to understand the amount of meetings that went undetected. On the other hand, as formally proved in Section 5.3.2, accounting for unseen meetings directly affects the clustering structure and, indirectly, the estimated adjacency matrices in Figure 5.5.1. The overall good fit of the proposed unknown-groups model to data, highlighted in Figure 5.5.1, also validates the estimate of the number of unseen meetings, which results to

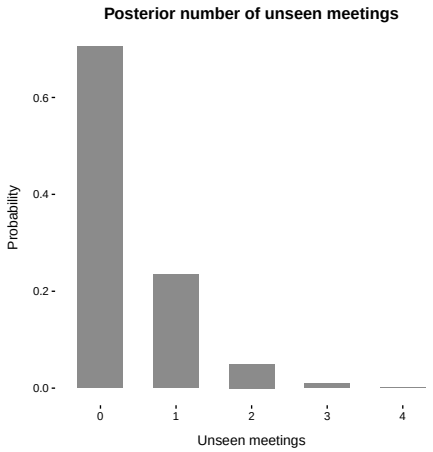



Figure 5.5.3: Posterior distributions of the number $M'$ of unseen meetings under the negative binomial mixture of BBs in Chapter 3.

be much smaller than the estimate provided under the known-groups setting using the clustering induced by *locali* membership, reported in Figure 5.5.2b. For completeness, we also compare these findings with those obtained from the negative binomial mixture of BBs (Chapter 3), the natural benchmark for estimating the number of unseen meetings. We remind that, by construction, this model imposes a homogeneous exchangeable structure, leading to posterior expectations of the adjacency matrix that assign identical values across all entries. As a result, it fails to reproduce the heterogeneity observed in the *Infinito* criminal data (Figure 5.5.1c). Consistent with this limitation, its inference on the number of unseen meetings $M'$ substantially diverges from that of our model. In particular, Figure 5.5.3 shows that the competitor places high posterior probability on the absence of undetected meetings, in contrast with the inference obtained under the unknown-groups model (Figure 5.5.2c). Model comparison via the WAIC further supports the data-driven clustering approach. Indeed, the negative binomial mixture of BBs attains a WAIC of $-5392$, which is worse not only than the unknown-groups model but also than the known-groups specification.

A word of caution is warranted. Estimating the number of unseen meetings is essentially an extrapolation task, which makes it difficult to assess the accuracy of the predictions. Simulation studies confirmed the usefulness of our model under correct specification. However, in real data applications some degree of misspecification is unavoidable. Thus, a pragmatic recommendation is to regard our estimates as lower bounds for the true number of unseen meetings, thereby ensuring a conservative interpretation of the results.

## 5.6 Discussion

In this chapter, we have introduced a new Bayesian nonparametric framework for modeling multivariate count data with an unknown number of traits. The model developed in the main body of the present chapter is based on finite completely random vectors, while results for the more general class of completely random vectors are provided in the Appendix. We focused on the cases where $A_{i\ell} \in \{0, 1\}$ (binary traits) or $A_{i\ell} \in \{0, 1, 2, \dots\}$ (Poisson counts), though our framework naturally extends to general parameter spaces for

$\theta_{jq}$, as discussed in the Appendix. The general theorems established in Section 5.2 were further leveraged to define a mixture model for clustering finite trait allocations. We also emphasized the methodological importance of modeling a random number of traits, as formalized in Proposition 5.1, and demonstrated its practical role in uncovering previously unseen meetings within the 'Ndrangheta dataset.

Although we focused on the analysis of criminal networks, the scope of our approach extends beyond this domain. For instance, we plan to explore its use to cluster microbiome profiles of different subjects, where traits correspond to operational taxonomic units observed with different frequencies in each subject. The importance of trait allocation models for microbiome data analysis has been recently emphasized by James et al. (2025). Another promising area of application is ecology, where our method can be used to cluster sites that exhibit similar patterns in terms of species occurrences (Chapter 3). In this context the traits are the species, observed with varying frequencies in each site.

In addition, our work also serves as a springboard for future methodological developments. First, all results in Section 5.B of the Appendix hold true for general CRVs, offering a theoretical basis for a broad class of models and providing practitioners with new modeling opportunities. Second, the results in Section 5.2 are the basic building blocks for addressing extrapolation problems involving dependent populations of traits. For example, one can estimate the number of unseen traits in an additional sample of subjects, thereby extending the results of Masoero et al. (2022); Camerlenghi et al. (2024) and Chapter 3 when multiple-sample data are available. While Shen et al. (2024) focused on predicting new genetic variants in presence of two dependent populations, our approach offers additional flexibility in ecological applications by allowing for finite species richness, since the number of traits is assumed finite but random. Finally, we believe that it is not difficult to extend our framework to cluster a bunch of trait allocations, rather than individual subjects. This would lead to models akin to nested processes (Rodriguez et al., 2008) within the trait allocation context. The aforementioned applications and methodological developments are left for future research.

# Appendix

## Organization of the Appendix

The Appendix is organized as follows. In Section 5.A, we review in detail the general class of completely random vectors, with special focus on the notable subclass of finite completely random vectors. Section 5.B presents the general distribution theory under completely random vector priors, which is then specialized to derive the results in Section 5.2.2. Finally, in Section 5.C, we report the proof of Proposition 5.1.

## 5.A Account on completely random vectors

In this section, we provide an account on completely random vectors (CRVs), with the goal of illustrating the connection between the processes presented in Section 5.2.1 and completely random vectors. The class of CRVs can be considered as a generalization of the class of completely random measures (CRMs), introduced by Kingman (1967), to the multivariate setting and played a leading role in many successful stories in the BNP literature. Indeed many authors have focused on CRVs to model the prior opinion in presence of partially exchangeable data, i.e., data organized in groups (Griffin and Leisen, 2017; Lijoi et al., 2014; Camerlenghi et al., 2019). The use of CRVs also allows the quantification of dependence induced by the prior across the different groups of observations (Catalano et al., 2021, 2024).

To mathematically introduce the class of CRVs, consider a Polish space $\mathbb{X}$ equipped with the corresponding Borel $\sigma$-algebra $\mathscr{X}$; consistently with the rest of the chapter, $\mathbb{X}$ represents the space of trait labels. Denote by $(\mathsf{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$ the space of boundedly finite (non-negative) measures on $\mathbb{X}$ equipped with the corresponding Borel $\sigma$-algebra generated by the topology of weak$^{\#}$-convergence (Daley and Vere-Jones, 2008). We also indicate by $\mathsf{M}_{\mathbb{X}}^d$ the $d$-fold product space of $\mathsf{M}_{\mathbb{X}}$, which is naturally endowed with the product $\sigma$-algebra. Remind that a random vector of measures $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ is a measurable function from a suitable probability space, say $(\Omega, \mathscr{A}, \mathsf{P})$, taking values in $(\mathsf{M}_{\mathbb{X}}^d, \mathscr{M}_{\mathbb{X}}^d)$. In the sequel, we will use the notation $\tilde{\boldsymbol{\mu}}(A) = (\tilde{\mu}_1(A), \ldots, \tilde{\mu}_d(A))$ to indicate the random vector evaluated on a set $A \in \mathscr{X}$. Then, CRVs are defined as follows.

**Definition 7** (Completely random vector)**.** *A random vector of measures $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ is a CRV if, for any $n \in \mathbb{N}$ and for any collection of disjoint sets $A_1, \ldots, A_n \in \mathscr{X}$, the evaluations $\tilde{\boldsymbol{\mu}}(A_1), \ldots, \tilde{\boldsymbol{\mu}}(A_n)$ are independent random vectors on $(0, \infty)^d$.*

Note that this definition entails that the marginals of a CRV are CRMs, since they have independent increments. Moreover, CRVs satisfy a similar decomposition as the one that holds true for CRMs. Indeed, a CRV consists of three main components: (i) a deterministic

drift $\boldsymbol{u}$, i.e., a deterministic measure with $d$ entries; (ii) a part with random $d$-dimensional jumps $(\boldsymbol{\beta}_k)_{k\geq 1}$ at fixed locations $(y_k)_{k\geq 1}$; (iii) a part with random $d$-dimensional jumps $(\boldsymbol{\theta}_j)_{j\geq 1}$ at random locations $(\tilde{X}_j)_{j\geq 1}$. Summing up, the following representation holds true

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{u} + \sum_{k\geq 1} \boldsymbol{\beta}_k \delta_{y_k} + \sum_{j\geq 1} \boldsymbol{\theta}_j \delta_{\tilde{X}_j}. \tag{5.18}$$

See (Kallenberg, 2017, Theorem 3.19). As a common practice in BNP models, here we remove the deterministic drift and the part with fixed locations, thus working only with component (iii); in other words we consider CRVs of the type:

$$\tilde{\boldsymbol{\mu}} = \sum_{j\geq 1} \boldsymbol{\theta}_j \delta_{\tilde{X}_j} = \sum_{j\geq 1} (\theta_{j1}, \ldots, \theta_{jd}) \delta_{\tilde{X}_j}.$$

Similarly to the CRM case, any CRV can be expressed as a functional of a Poisson point process (Baccelli et al., 2020) on the space $(0,\infty)^d \times \mathbb{X}$. Indeed, consider the Laplace functional of $\tilde{\boldsymbol{\mu}}$, which is defined as

$$\mathcal{L}_{\tilde{\boldsymbol{\mu}}}(g_1, \ldots, g_d) := \mathsf{E}\left( e^{-\sum_{q=1}^d \int_{\mathbb{X}} g_q(x) \tilde{\mu}_q(\mathrm{d}x)} \right)$$

for all measurable functions $g_1, \ldots, g_d : \mathbb{X} \to (0,\infty)$. The Laplace functional of $\tilde{\boldsymbol{\mu}}$ admits the following Lévy-Khintchine representation

$$\mathcal{L}_{\tilde{\boldsymbol{\mu}}}(g_1, \ldots, g_d) = \exp\left\{ -\int_{(0,\infty)^d \times \mathbb{X}} \left( 1 - \exp\left\{ -\sum_{q=1}^d g_q(x)\theta_q \right\} \right) \nu(\mathrm{d}\theta_1 \ \ldots \ \mathrm{d}\theta_d \, \mathrm{d}x) \right\}, \tag{5.19}$$

for all measurable functions $g_1, \ldots, g_d : \mathbb{X} \to (0,\infty)$, where $\nu(\,\cdot\,)$ is a measure on $(0,\infty)^d \times \mathbb{X}$ referred to as the Lévy intensity of $\tilde{\boldsymbol{\mu}}$. The measure $\nu(\,\cdot\,)$ must satisfy the following conditions

$$\nu((0,\infty)^d \times \{x\}) = 0, \ \forall x \in \mathbb{X}, \qquad \text{and} \qquad \int_{(0,\infty)^d \times B} \min\{||\boldsymbol{\theta}||, 1\} \, \nu(\mathrm{d}\theta_1 \ \ldots \ \mathrm{d}\theta_d \, \mathrm{d}x) < \infty,$$

for any bounded Borel set $B \in \mathscr{X}$, where $||\boldsymbol{\theta}||$ denotes the Euclidean norm of the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. In the present work we consider homogeneous CRVs, namely CRVs whose Lévy intensity satisfies $\nu(\mathrm{d}\theta_1 \ \ldots \ \mathrm{d}\theta_d \, \mathrm{d}x) = \rho_d(\mathrm{d}\theta_1 \ \ldots \ \mathrm{d}\theta_d)\alpha(\mathrm{d}x)$. In particular, assume that $\alpha(\,\cdot\,)$ is a finite non-atomic measure on $\mathbb{X}$ with total mass $\lambda$, thus set $G_0(\,\cdot\,) := \alpha(\,\cdot\,)/\lambda$. In the sequel, we use the notation $\tilde{\boldsymbol{\mu}} \sim \text{CRV}(\rho_d, \lambda, G_0)$ to indicate the distribution of a homogeneous CRV. We refer to Kallenberg (2017) for additional details on CRVs.

### 5.A.1 Finite completely random vectors

In this work, we focus on a special class of CRVs which we refer to as finite completely random vectors (FCRVs). Due to the purposes of the chapter, we restrict the attention to the homogeneous case; extension to non-homogeneous CRVs is straightforward. We define (homogeneous) FCRVs as follows.

**Definition 8** (Finite completely random vector)**.** *Let $\tilde{\boldsymbol{\mu}} \sim \text{CRV}(\rho_d, \lambda, G_0)$. If $\rho_d$ is a finite measure, then $\tilde{\boldsymbol{\mu}}$ is a FCRV.*

When $\tilde{\boldsymbol{\mu}}$ is a FCRV, we write $\tilde{\boldsymbol{\mu}} \sim \text{FCRV}(\rho_d, \lambda, G_0)$, where $\rho_d$ is necessarily finite. Without loss of generality, assume that $\rho_d = H^{(d)}$ is a probability measure and adjust $\lambda$ to account for it. As we will discuss in the next proposition, the specification *finite* CRV is motivated by the fact that FCRVs can be seen as functionals of finite Poisson point processes (Baccelli et al., 2020, Section 4.3). In particular, the random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are supported on a (finite) Poisson random number of common atoms.

**Proposition 5.2.** *If $\tilde{\boldsymbol{\mu}} \sim \text{FCRV}(H^{(d)}, \lambda, G_0)$, then $\tilde{\boldsymbol{\mu}}$ can be represented as*

$$\tilde{\boldsymbol{\mu}} = \sum_{j=1}^{M} (\theta_{j1}, \ldots, \theta_{jd}) \delta_{\tilde{X}_j}, \tag{5.20}$$

*where $M$ is a Poisson random variable with parameter $\lambda > 0$. Moreover, conditionally to $M$, the vectors $(\theta_{j1}, \ldots, \theta_{jd})$ are i.i.d. draws from $H^{(d)}$, namely $(\theta_{j1}, \ldots, \theta_{jd}) \overset{iid}{\sim} H^{(d)}$, and $\tilde{X}_j \overset{iid}{\sim} G_0$, as $j = 1, \ldots, M$.*

*Proof.* In order to show the result, we prove that the Laplace functional of $\tilde{\boldsymbol{\mu}}$, as defined in (5.20), coincides with the one of a $\text{FCRV}(H^{(d)}, \lambda, G_0)$. To this end, consider the measurable functions $g_1, \ldots, g_d : \mathbb{X} \to (0, \infty)$ and note that, conditionally on $M$, we have

$$\begin{aligned}
&\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\} \mid M\right] = \mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\sum_{j=1}^{M}\theta_{jq}\, g_q(\tilde{X}_j)\right\} \mid M\right] \\
&\quad = \mathsf{E}\left[\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\sum_{j=1}^{M}\theta_{jq}\, g_q(\tilde{X}_j)\right\} \mid M, \tilde{X}_1, \ldots, \tilde{X}_M\right] \mid M\right] \\
&\quad = \mathsf{E}\left[\prod_{j=1}^{M}\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\theta_{jq}\, g_q(\tilde{X}_j)\right\} \mid M, \tilde{X}_j\right] \mid M\right],
\end{aligned}$$

where we first exploit the tower property of conditional expectations and then we use the independence of the random vectors, conditionally on $M$. Observing that, conditionally on $M$, the vectors $(\theta_{j1}, \ldots, \theta_{jd})$ are i.i.d. from distribution $H^{(d)}$ and $\tilde{X}_j \overset{\text{iid}}{\sim} G_0$, for $j = 1, \ldots, M$, the previous expression equals

$$\begin{aligned}
&\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\} \mid M\right] \\
&\quad = \mathsf{E}\left[\prod_{j=1}^{M}\int_{(0,\infty)^d} \exp\left\{-\sum_{q=1}^{d}\theta_q\, g_q(\tilde{X}_j)\right\} H^{(d)}(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) \mid M\right] \\
&\quad = \prod_{j=1}^{M}\int_{\mathbb{X}}\int_{(0,\infty)^d} \exp\left\{-\sum_{q=1}^{d}\theta_q\, g_q(x)\right\} H^{(d)}(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) G_0(\mathrm{d}x) \\
&\quad = \left(\int_{(0,\infty)^d\times\mathbb{X}} \exp\left\{-\sum_{q=1}^{d}\theta_q\, g_q(x)\right\} H^{(d)}(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) G_0(\mathrm{d}x)\right)^{M}.
\end{aligned} \tag{5.21}$$

We can now integrate out $M$ in the final expression of (5.21) to obtain the Laplace functional of $\tilde{\boldsymbol{\mu}}$ defined as in (5.20). Specifically, we have

$$\begin{aligned}
\mathcal{L}_{\tilde{\boldsymbol{\mu}}}(g_1,\dots,g_d) &= \mathsf{E}\left[\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\} \mid M\right]\right] \\
&= \mathsf{E}\left[\left(\int_{(0,\infty)^d\times\mathbb{X}} \exp\left\{-\sum_{q=1}^{d}\theta_q\, g_q(x)\right\} H^{(d)}(\mathrm{d}\theta_1\,\dots\,\mathrm{d}\theta_d)G_0(\mathrm{d}x)\right)^{M}\right] \\
&= e^{-\lambda}\sum_{j\geq 0}\frac{1}{j!}\left[\lambda\int_{(0,\infty)^d\times\mathbb{X}} \exp\left\{-\sum_{q=1}^{d}\theta_q\, g_q(x)\right\} H^{(d)}(\mathrm{d}\theta_1\,\dots\,\mathrm{d}\theta_d)G_0(\mathrm{d}x)\right]^{j} \\
&= e^{-\lambda}\exp\left\{\lambda\int_{(0,\infty)^d\times\mathbb{X}} \exp\left\{-\sum_{q=1}^{d}\theta_q\, g_q(x)\right\} H^{(d)}(\mathrm{d}\theta_1\,\dots\,\mathrm{d}\theta_d)G_0(\mathrm{d}x)\right\} \\
&= \exp\left\{-\lambda\left(1-\int_{(0,\infty)^d\times\mathbb{X}} \exp\left\{-\sum_{q=1}^{d}\theta_q\, g_q(x)\right\} H^{(d)}(\mathrm{d}\theta_1\,\dots\,\mathrm{d}\theta_d)G_0(\mathrm{d}x)\right)\right\} \\
&= \exp\left\{-\lambda\int_{(0,\infty)^d\times\mathbb{X}}\left(1-\exp\left\{-\sum_{q=1}^{d}\theta_q\, g_q(x)\right\}\right) H^{(d)}(\mathrm{d}\theta_1\,\dots\,\mathrm{d}\theta_d)G_0(\mathrm{d}x)\right\},
\end{aligned} \tag{5.22}$$

where the first equality follows from the tower property, expression (5.21) has been exploited in the second line, and finally the expectation has been evaluated relying on the fact that $M$ is a Poisson random variable with parameter $\lambda$. From (5.19), we note that the equation in (5.22) corresponds to the Laplace functional of a FCRV$(H^{(d)},\lambda,G_0)$, and the thesis follows. □

Based on Proposition 5.2, the connection between the prior distribution for $\tilde{\boldsymbol{\mu}}$ defined in equations (5.5)–(5.6), Section 5.2.1, and the class of FCRVs is evident. Indeed, the construction in equations (5.5)–(5.6) precisely corresponds to the formulation in (5.20), thus $\tilde{\boldsymbol{\mu}}$ can be equivalently described as $\tilde{\boldsymbol{\mu}} \sim$ FCRV$(H^{(d)},\lambda,G_0)$.

## 5.B General distribution theory under CRV priors

In this section, we provide a complete distribution theory for partially exchangeable trait allocation models (5.4), when the prior on $\tilde{\boldsymbol{\mu}}$ is a generic CRV. This encompasses both finite- and infinite-dimensional trait models. The results presented here are very general and can be specialized to a variety of different prior specifications. Moreover, they can also be applied to face prediction in presence of multiple dependent populations. In the last part of this section, we will focus on the finite trait models of Section 5.2.1, corresponding to the case of $\tilde{\boldsymbol{\mu}}$ distributed as a FCRV. In particular, we will comment on how the results presented in Section 5.2.2 follow from these general theorems.

First, we recall the general model we deal with:

$$\begin{aligned}
Z_{iq}\,|\,\tilde{\mu}_q &\overset{\text{ind}}{\sim} \text{CP}(\tilde{\mu}_q), \qquad i\geq 1, \quad q=1,\dots,d, \\
\tilde{\boldsymbol{\mu}} &= (\tilde{\mu}_1,\dots,\tilde{\mu}_d) \sim \text{CRV}(\rho_d,\lambda,G_0),
\end{aligned} \tag{5.23}$$

where $\rho_d$ is a measure on $(0,\infty)^d$, $\lambda > 0$ and $G_0$ is a non-atomic probability measure on $\mathbb{X}$. Refer to Section 5.A for the definition of $\text{CRV}(\rho_d, \lambda, G_0)$. Clearly, model (5.4) with prior distribution given by equations (5.5)–(5.6) is obtained by considering $\rho_d = H^{(d)}$, with $H^{(d)}$ being a probability law on $(0,\infty)^d$, that is $\tilde{\boldsymbol{\mu}} \sim \text{FCRV}(H^{(d)}, \lambda, G_0)$.

We first focus on the marginal distribution of a sample $\boldsymbol{Z} = (Z_{iq} : i = 1, \ldots, n_q; q = 1, \ldots, d)$. As explained in the main body, with marginal distribution of $\boldsymbol{Z}$, we refer to the probabilities of the events $(\boldsymbol{A} = \boldsymbol{a}, K_n = k)$, where $K_n$ is the observed number of distinct traits and $\boldsymbol{A}$ collects the observed counts, as introduced for Theorem 5.1, Section 5.2.2. The following theorem describes the marginal probability for the event $(\boldsymbol{A} = \boldsymbol{a}, K_n = k)$.

**Theorem 5.3** (Marginal distribution, pETPF)**.** *Let $\boldsymbol{Z}$ be a sample from the statistical model* (5.23)*. The probability that $\boldsymbol{Z}$ displays $K_n = k$ distinct traits with counts $\boldsymbol{A} = \boldsymbol{a}$ equals*

$$\begin{aligned}\pi_n(\boldsymbol{a}) = \frac{\lambda^k}{k!} \exp\left\{-\lambda \int_{(0,\infty)^d} \left(1 - \prod_{q=1}^{d} P(0;\theta_q)^{n_q}\right) \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d)\right\} \\ \times \prod_{\ell=1}^{k} \int_{(0,\infty)^d} \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(a_{i\ell q};\theta_q) \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d)\end{aligned} \tag{5.24}$$

*where $n = \sum_{q=1}^{d} n_q$ and $\boldsymbol{n} = (n_1, \ldots, n_d)$ are the sample sizes.*

*Proof.* Consider the event $(\boldsymbol{A} = \boldsymbol{a}, K_n = k)$, corresponding to $\boldsymbol{Z}$ displaying $K_n = k$ distinct traits, with associated counts described by $\boldsymbol{A} = \boldsymbol{a}$. Though already specified in the definition of $\boldsymbol{A}$, it worth stressing that this event $(\boldsymbol{A} = \boldsymbol{a}, K_n = k)$ refers to one specific ordering of the traits among all the possible ones, when a uniform distribution on the orderings is assumed. For the sake of exposition, it is convenient to observe that $\boldsymbol{A} = \boldsymbol{a}$ describes the presence and absence of the $k$ observed traits in the subjects. In particular, for each subpopulation $q = 1, \ldots, d$ and each observed trait $\ell = 1, \ldots, k$, define $\mathcal{B}_{\ell q} = \{(i,q) : a_{i\ell q} > 0, i = 1, \ldots, n_q\}$, which might be empty.

The probability of the event under study $(\boldsymbol{A} = \boldsymbol{a}, K_n = k)$ is indicated by $\pi_n(\boldsymbol{a})$ and generalizes the well-known notion of partially Exchangeable Partition Probability Function (pEPPF) to the setting of trait allocation models, thus we refer to it as partially Exchangeable Trait Probability Function (pETPF). In order to evaluate $\pi_n(\boldsymbol{a})$, we apply a limiting argument, which is based on the evaluation of the joint distribution of $(\boldsymbol{A}, K_n)$ and the associated trait labels $X_1, \ldots, X_{K_n}$. To this end, we consider $k$ small balls $B_\varepsilon(X_\ell)$ centered at $X_\ell$ with radius $\varepsilon$, where $\varepsilon > 0$ is sufficiently small so that all the balls are disjoint. Moreover, we set $\mathbb{X}^* := \mathbb{X} \setminus \cup_{\ell=1}^{k} B_\varepsilon(X_\ell)$. First, focus on the probability of the event $\mathscr{E}$

$$\begin{aligned}\mathscr{E} = \{\, \exists j \geq 1 :\ & \tilde{A}_{ijq} = a_{i\ell q} \text{ with } \tilde{X}_j \in B_\varepsilon(X_\ell),\ \forall (i,q) \in \mathcal{B}_{\ell q}, \\ & \qquad\qquad q = 1, \ldots, d,\ \ell = 1, \ldots, k; \\ & Z_{iq}(B_\varepsilon(X_\ell)) = 0,\ \forall (i,q) \notin \mathcal{B}_{\ell q},\ q = 1, \ldots, d,\ \ell = 1, \ldots, k; \\ & Z_{iq}(\mathbb{X}^*) = 0,\ \forall i = 1, \ldots, n_q,\ q = 1, \ldots, d \,\},\end{aligned} \tag{5.25}$$

that is to say the *infinitesimal* probability of observing $k$ distinct traits with associated counts $\boldsymbol{A} = \boldsymbol{a}$ and trait labels belonging to the small balls $B_\varepsilon(X_1), \ldots, B_\varepsilon(X_k)$.

To evaluate the probability of $\mathscr{E}$, by an application of the tower property, we have

$$\mathsf{P}(\mathscr{E}) = \mathsf{E}\left[\mathsf{P}(\mathscr{E} \mid \tilde{\boldsymbol{\mu}})\right]. \tag{5.26}$$

Define the following three disjoint events, whose union equals $\mathscr{E}$, as

$$\begin{aligned}
\mathscr{E}_1 &:= \left\{\exists j \geq 1 :\ \tilde{A}_{ijq} = a_{i\ell q} \text{ with } \tilde{X}_j \in B_\varepsilon(X_\ell),\ \forall (i,q) \in \mathcal{B}_{\ell q},\ q = 1,\dots,d,\ \ell = 1,\dots,k\right\},\\
\mathscr{E}_2 &:= \left\{Z_{iq}(B_\varepsilon(X_\ell)) = 0,\ \forall (i,q) \notin \mathcal{B}_{\ell q},\ q = 1,\dots,d,\ \ell = 1,\dots,k\right\},\\
\mathscr{E}_3 &:= \left\{Z_{iq}(\mathbb{X}^*) = 0,\ \forall i = 1,\dots,n_q,\ q = 1,\dots,d\right\}.
\end{aligned}$$

Conditionally to $\tilde{\boldsymbol{\mu}}$, the randomness of the observations $Z_{iq}$ from model (5.23) is only in the independent random variables $\tilde{A}_{ijq}$. Thus, the three events $\mathscr{E}_1, \mathscr{E}_2, \mathscr{E}_3$ are independent, conditionally to $\tilde{\boldsymbol{\mu}}$, because they relate to independent $\tilde{A}_{ijq}$, and their probabilities can be evaluated separately as

$$\mathsf{P}(\mathscr{E}) = \mathsf{E}\left[\mathsf{P}(\mathscr{E}_1 \mid \tilde{\boldsymbol{\mu}})\,\mathsf{P}(\mathscr{E}_2 \mid \tilde{\boldsymbol{\mu}})\,\mathsf{P}(\mathscr{E}_3 \mid \tilde{\boldsymbol{\mu}})\right]. \tag{5.27}$$

Hereafter, by a slight abuse of notation, we write $i \in \mathcal{B}_{\ell q}$ to indicate that $(i,q) \in \mathcal{B}_{\ell q}$. For the first conditional probability in (5.27), it is easy to see that

$$\begin{aligned}
\mathsf{P}(\mathscr{E}_1 \mid \tilde{\boldsymbol{\mu}}) &= \prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i \in \mathcal{B}_{\ell q}} \mathsf{P}\left(\exists j \geq 1 :\ \tilde{A}_{ijq} = a_{i\ell q} \text{ with } \tilde{X}_j \in B_\varepsilon(X_\ell) \mid \tilde{\boldsymbol{\mu}}\right)\\
&= \prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i \in \mathcal{B}_{\ell q}} \left(1 - \prod_{j \geq 1}(1 - P(a_{i\ell q}; \theta_{jq}))^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))}\right).
\end{aligned}$$

For the second conditional probability in (5.27), we observe that

$$\begin{aligned}
\mathsf{P}(\mathscr{E}_2 \mid \tilde{\boldsymbol{\mu}}) &= \prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i \notin \mathcal{B}_{\ell q}} \mathsf{P}\left(Z_{iq}(B_\varepsilon(X_\ell)) = 0 \mid \tilde{\mu}_q\right)\\
&= \prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i \notin \mathcal{B}_{\ell q}}\prod_{j \geq 1} P(0; \theta_{jq})^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))}.
\end{aligned}$$

Analogous considerations hold true for the conditional probability of $\mathscr{E}_3$ in (5.27). Thus, putting everything together, we have

$$\begin{aligned}
\mathsf{P}(\mathscr{E}) = \mathsf{E}\Bigg[&\prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i \notin \mathcal{B}_{\ell q}}\prod_{j \geq 1} P(0; \theta_{jq})^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))}\\
&\times \prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i \in \mathcal{B}_{\ell q}} \left(1 - \prod_{j \geq 1}(1 - P(a_{i\ell q}; \theta_{jq}))^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))}\right)\\
&\times \prod_{q=1}^{d}\prod_{j \geq 1}\prod_{i=1}^{n_q} P(0; \theta_{jq})^{\delta_{\tilde{X}_j}(\mathbb{X}^*)}\Bigg].
\end{aligned} \tag{5.28}$$

By observing that, for any $\ell = 1, \dots, k$ and for any $q = 1, \dots, d$,

$$\prod_{i \in \mathcal{B}_{\ell q}} \left( 1 - \prod_{j \geq 1} (1 - P(a_{i\ell q}; \theta_{jq}))^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))} \right) = \sum_{\substack{v_{i\ell q} \in \{0,1\} \\ i \in \mathcal{B}_{\ell q}}} \prod_{i \in \mathcal{B}_{\ell q}} (-1)^{v_{i\ell q}} \prod_{j \geq 1} (1 - P(a_{i\ell q}; \theta_{jq}))^{v_{i\ell q} \delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))},$$

the expected value in (5.28) boils down to the following sum

$$\begin{aligned} \mathsf{P}(\mathscr{E}) = \sum_{\substack{v_{i\ell q} \in \{0,1\} \\ i \in \mathcal{B}_{\ell q} \\ q=1,\dots,d \\ \ell = 1, \dots, k}} \left( \prod_{q=1}^{d} \prod_{\ell=1}^{k} \prod_{i \in \mathcal{B}_{\ell q}} (-1)^{v_{i\ell q}} \right) \mathsf{E} \Bigg[ & \prod_{q=1}^{d} \prod_{\ell=1}^{k} \prod_{i \notin \mathcal{B}_{\ell q}} \prod_{j \geq 1} P(0; \theta_{jq})^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))} \\ & \times \prod_{q=1}^{d} \prod_{\ell=1}^{k} \prod_{i \in \mathcal{B}_{\ell q}} \prod_{j \geq 1} (1 - P(a_{i\ell q}; \theta_{jq}))^{v_{i\ell q} \delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))} \\ & \times \prod_{q=1}^{d} \prod_{j \geq 1} \prod_{i=1}^{n_q} P(0; \theta_{jq})^{\delta_{\tilde{X}_j}(\mathbb{X}^*)} \Bigg]. \end{aligned} \tag{5.29}$$

At this point, we need to evaluate the expectation in (5.29). Since the sets $B_\varepsilon(X_1), \dots, B_\varepsilon(X_k)$ and $\mathbb{X}^*$ are disjoint, the independence property of the CRV $\tilde{\boldsymbol{\mu}}$ implies that the expected value in (5.29) equals the following product

$$\prod_{\ell=1}^{k} E_\ell \times E_*, \tag{5.30}$$

where we have set:

$$\begin{aligned} E_\ell &:= \mathsf{E} \left\{ \prod_{j \geq 1} \prod_{q=1}^{d} \left[ \prod_{i \notin \mathcal{B}_{\ell q}} P(0; \theta_{jq})^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))} \cdot \prod_{i \in \mathcal{B}_{\ell q}} (1 - P(a_{i\ell q}; \theta_{jq}))^{v_{i\ell q} \delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))} \right] \right\}, \\ E_* &:= \mathsf{E} \left[ \prod_{q=1}^{d} \prod_{j \geq 1} \prod_{i=1}^{n_q} P(0; \theta_{jq})^{\delta_{\tilde{X}_j}(\mathbb{X}^*)} \right]. \end{aligned}$$

We now concentrate on the evaluation of $E_*$ in (5.30):

$$\begin{aligned} E_* &= \mathsf{E} \left[ \prod_{q=1}^{d} \prod_{j \geq 1} \prod_{i=1}^{n_q} P(0; \theta_{jq})^{\delta_{\tilde{X}_j}(\mathbb{X}^*)} \right] \\ &= \mathsf{E} \left[ \exp \left\{ \sum_{j \geq 1} \log \left( \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(0; \theta_{jq}) \right) \delta_{\tilde{X}_j}(\mathbb{X}^*) \right\} \right] \\ &= \exp \left\{ - \int_{(0,\infty)^d \times \mathbb{X}^*} \left( 1 - \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(0; \theta_q) \right) \rho_d(\mathrm{d}\theta_1 \, \dots \, \mathrm{d}\theta_d) \, \lambda G_0(\mathrm{d}x) \right\}, \end{aligned} \tag{5.31}$$

where we exploit the available expression of the Laplace functional for the CRVs (or analogously for Poisson processes). As for the general term $E_\ell$ in (5.30), the Lévy-Khintchine

representation implies again

$$E_\ell = \mathsf{E}\left[\exp\left\{\sum_{j\geq 1}\log\left(\prod_{q=1}^{d}\left[\prod_{i\notin\mathcal{B}_{\ell q}}P(0;\theta_{jq})\prod_{i\in\mathcal{B}_{\ell q}}(1-P(a_{i\ell q};\theta_{jq}))^{v_{i\ell q}}\right]\right)\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))\right\}\right]$$
$$= \exp\left\{-\lambda G_0(B_\varepsilon(X_\ell))\right.$$
$$\left.\times\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\left[\prod_{i\notin\mathcal{B}_{\ell q}}P(0;\theta_q)\prod_{i\in\mathcal{B}_{\ell q}}(1-P(a_{i\ell q};\theta_q))^{v_{i\ell q}}\right]\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}.$$

By defining

$$I_{\ell,v_{i\ell q}} := \int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\left[\prod_{i\notin\mathcal{B}_{\ell q}}P(0;\theta_q)\prod_{i\in\mathcal{B}_{\ell q}}(1-P(a_{i\ell q};\theta_q))^{v_{i\ell q}}\right]\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d),$$

we get

$$E_\ell = \exp\left\{-\lambda G_0(B_\varepsilon(X_\ell))I_{\ell,v_{i\ell q}}\right\}. \tag{5.32}$$

We now substitute the expressions (5.31)–(5.32) in (5.30) to evaluate the expected value in (5.29), which equals

$$\exp\left\{-\lambda G_0(\mathbb{X}^*)\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\prod_{i=1}^{n_q}P(0;\theta_q)\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}$$
$$\times\prod_{\ell=1}^{k}\exp\left\{-\lambda G_0(B_\varepsilon(X_\ell))I_{\ell,v_{i\ell q}}\right\}$$
$$= \exp\left\{-\lambda G_0(\mathbb{X}^*)\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\prod_{i=1}^{n_q}P(0;\theta_q)\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}$$
$$\times\prod_{\ell=1}^{k}\left(1-\lambda G_0(B_\varepsilon(X_\ell))I_{\ell,v_{i\ell q}}+o(G_0(B_\varepsilon(X_\ell)))\right).$$

We finally substitute the previous expression in (5.29) to get the following final expression for the probability of the event $\mathscr{E}$,

$$\mathsf{P}(\mathscr{E}) = \sum_{\substack{v_{i\ell q}\in\{0,1\}\\ i\in\mathcal{B}_{\ell q}\\ q=1,\ldots,d\\ \ell=1,\ldots,k}}\left\{\left(\prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i\in\mathcal{B}_{\ell q}}(-1)^{v_{i\ell q}}\right)\prod_{\ell=1}^{k}\left(1-\lambda G_0(B_\varepsilon(X_\ell))I_{\ell,v_{i\ell q}}+o(G_0(B_\varepsilon(X_\ell)))\right)\right\}$$
$$\times\exp\left\{-\lambda G_0(\mathbb{X}^*)\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\prod_{i=1}^{n_q}P(0;\theta_q)\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}.$$

Handling the last expression, we focus on the sum over the $v_{i\ell q}$ and we get

$$\begin{aligned}
\mathsf{P}(\mathscr{E}) &= \exp\left\{-\lambda G_0(\mathbb{X}^*)\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\prod_{i=1}^{n_q}P(0;\theta_q)\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}\\
&\quad\times\prod_{\ell=1}^{k}\Big(\sum_{\substack{v_{iq}\in\{0,1\}\\ i\in\mathcal{B}_{\ell q}\\ q=1,\ldots,d}}\Big(\prod_{q=1}^{d}\prod_{i\in\mathcal{B}_{\ell q}}(-1)^{v_{iq}}\Big)(1-\lambda G_0(B_\varepsilon(X_\ell))I_{\ell,v_{iq}}+o(G_0(B_\varepsilon(X_\ell))))\Big)\\
&= \exp\left\{-\lambda G_0(\mathbb{X}^*)\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\prod_{i=1}^{n_q}P(0;\theta_q)\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}\\
&\quad\times\prod_{\ell=1}^{k}\Big(\sum_{\substack{v_{iq}\in\{0,1\}\\ i\in\mathcal{B}_{\ell q}\\ q=1,\ldots,d}}\Big(\prod_{q=1}^{d}\prod_{i\in\mathcal{B}_{\ell q}}(-1)^{v_{iq}}\Big)(-\lambda G_0(B_\varepsilon(X_\ell))I_{\ell,v_{iq}}+o(G_0(B_\varepsilon(X_\ell))))\Big)
\end{aligned}$$

where we use the fact that

$$\sum_{\substack{v_{iq}\in\{0,1\}\\ i\in\mathcal{B}_{\ell q}\\ q=1,\ldots,d}}\Big(\prod_{q=1}^{d}\prod_{i\in\mathcal{B}_{\ell q}}(-1)^{v_{iq}}\Big)=0.$$

By exploiting the definition of $I_{\ell,v_{iq}}$ and applying similar arguments as above, we obtain

$$\begin{aligned}
\mathsf{P}(\mathscr{E}) &= \exp\left\{-\lambda G_0(\mathbb{X}^*)\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\prod_{i=1}^{n_q}P(0;\theta_q)\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}\\
&\quad\times\prod_{\ell=1}^{k}\left\{\lambda G_0(B_\varepsilon(X_\ell))\int_{(0,\infty)^d}\prod_{q=1}^{d}\Big(\prod_{i\notin\mathcal{B}_{\ell q}}P(0;\theta_q)\sum_{\substack{v_i\in\{0,1\}\\ i\in\mathcal{B}_{\ell q}}}\prod_{i\in\mathcal{B}_{\ell q}}(P(a_{i\ell q};\theta_q)-1)^{v_i}\Big)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}\\
&\quad+o\Big(\prod_{\ell=1}^{k}G_0(B_\varepsilon(X_\ell))\Big).
\end{aligned}$$

It is now easy to solve the summation over the $v_i$ to get the following expression:

$$\begin{aligned}
\mathsf{P}(\mathscr{E}) &= \exp\left\{-\lambda G_0(\mathbb{X}^*)\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\prod_{i=1}^{n_q}P(0;\theta_q)\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}\\
&\quad\times\prod_{\ell=1}^{k}\left\{\lambda G_0(B_\varepsilon(X_\ell))\int_{(0,\infty)^d}\prod_{q=1}^{d}\Big(\prod_{i\notin\mathcal{B}_{\ell q}}P(0;\theta_q)\prod_{i\in\mathcal{B}_{\ell q}}P(a_{i\ell q};\theta_q)\Big)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}\\
&\quad+o\Big(\prod_{\ell=1}^{k}G_0(B_\varepsilon(X_\ell))\Big).
\end{aligned}$$

Finally, to obtain the targeted probability $\pi_n(\boldsymbol{a})$, namely the pETPF of model (5.23), we need to link it with the just computed probability of the event $\mathscr{E}$. For sufficiently small $\varepsilon$,

$B_\varepsilon(X_1), \ldots, B_\varepsilon(X_k)$ are disjoint, thus it holds

$$\mathsf{P}(\mathscr{E}) = k! \prod_{\ell=1}^{k} G_0(B_\varepsilon(X_\ell)) \cdot \pi_n(\boldsymbol{a}),$$

where $k!$ discounts for the specific ordering of the traits which is implicit in $\pi_n(\boldsymbol{a})$, as detailed in the initial comments of the proof. Then,

$$\begin{aligned}
\pi_n(\boldsymbol{a}) &= \lim_{\varepsilon \to 0} \frac{\mathsf{P}(\mathscr{E})}{k! \prod_{\ell=1}^{k} G_0(B_\varepsilon(X_\ell))} \\
&= \frac{\lambda^k}{k!} \exp\left\{ -\lambda \int_{(0,\infty)^d} \left( 1 - \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(0;\theta_q) \right) \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) \right\} \\
&\quad \times \prod_{\ell=1}^{k} \int_{(0,\infty)^d} \prod_{q=1}^{d} \Big( \prod_{i \notin \mathcal{B}_{\ell q}} P(0;\theta_q) \prod_{i \in \mathcal{B}_{\ell q}} P(a_{i\ell q};\theta_q) \Big) \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d),
\end{aligned}$$

which can be rewritten as in the statement of the proposition. □

Secondly, we move to the characterization of the posterior distribution of $\tilde{\boldsymbol{\mu}}$ in (5.23), conditionally to a sample $\boldsymbol{Z}$.

**Theorem 5.4** (Posterior distribution)**.** *Let $\boldsymbol{Z}$ be a sample from the statistical model* (5.23)*. If $\boldsymbol{Z}$ displays $K_n = k$ distinct traits labeled $X_1, \ldots, X_k$, with associated counts $\boldsymbol{a}$, then the posterior distribution of $\tilde{\boldsymbol{\mu}}$ satisfies the distributional equality*

$$(\tilde{\mu}_1, \ldots, \tilde{\mu}_d) \,|\, \boldsymbol{Z} \stackrel{d}{=} (\mu_1^*, \ldots, \mu_d^*) + (\mu_1', \ldots, \mu_d'), \tag{5.33}$$

*where $\boldsymbol{\mu}^* := (\mu_1^*, \ldots, \mu_d^*)$ and $\boldsymbol{\mu}' := (\mu_1', \ldots, \mu_d')$ are independent random vectors such that*

*(i) the components of the vector $\boldsymbol{\mu}^*$ are defined as $\mu_q^*(\cdot) = \sum_{\ell=1}^{k} \theta_{\ell q}^* \delta_{X_\ell}(\cdot)$, for $q = 1, \ldots, d$, and the random vectors $(\theta_{\ell 1}^*, \ldots, \theta_{\ell d}^*)$ are independent across $\ell = 1, \ldots, k$, with distribution*

$$H_{\ell q}(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) \propto \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(a_{i\ell q};\theta_q) \cdot \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d); \tag{5.34}$$

*(ii) the process $(\mu_1', \ldots, \mu_d')$ is a CRV$(\rho_d', \lambda, G_0)$, with*

$$\rho_d'(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) = \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(0;\theta_q) \cdot \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d).$$

*Proof.* In order to characterize the posterior distribution of $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$, we show that its Laplace functional coincides with the one of the sum on the right-hand side of (5.33). To this end, consider $d$ measurable functions $g_q : \mathbb{X} \to (0, \infty)$, as $q = 1, \ldots, d$ and focus on the Laplace functional of the vector $\tilde{\boldsymbol{\mu}}$, conditionally to $\boldsymbol{Z}$,

$$\mathcal{L}_{\tilde{\boldsymbol{\mu}} \,|\, \boldsymbol{Z}}(g_1, \ldots, g_d) = \mathsf{E}\left[ \exp\left\{ -\sum_{q=1}^{d} \int_{\mathbb{X}} g_q(x) \tilde{\mu}_q(\mathrm{d}x) \right\} \,|\, \boldsymbol{Z} \right].$$

This conditional Laplace functional may be evaluated as

$$\mathcal{L}_{\tilde{\boldsymbol{\mu}}\,|\,\boldsymbol{Z}}(g_1,\ldots,g_d) = \lim_{\varepsilon\downarrow 0}\frac{1}{\mathsf{P}(\mathscr{E})}\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\}\cdot\mathbb{1}_{\mathscr{E}}\right], \tag{5.35}$$

where $\mathscr{E}$ is the event defined in (5.25), for proof of Theorem 5.3, which depends on $\varepsilon$, and $\mathbb{1}_{\mathscr{E}}$ denotes the corresponding indicator function. The denominator in (5.35) has already been computed in the proof of Theorem 5.3; a similar argument may be applied to find an expression for the expected value in (5.35). First of all, the tower property of the conditional expectation implies

$$\begin{aligned}\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\}\cdot\mathbb{1}_{\mathscr{E}}\right] &= \mathsf{E}\left[\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\}\cdot\mathbb{1}_{\mathscr{E}}\;|\;\tilde{\boldsymbol{\mu}}\right]\right]\\ &= \mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\}\cdot\mathsf{P}(\mathscr{E}\,|\,\tilde{\boldsymbol{\mu}})\right].\end{aligned}$$

The probability $\mathsf{P}(\mathscr{E}\,|\,\tilde{\boldsymbol{\mu}})$ has already been computed in the proof of Theorem 5.3, while the exponential term can be explicitly written as a function of the a.s. discrete random measures $\tilde{\mu}_q$, which leads to

$$\begin{aligned}&\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\}\cdot\mathbb{1}_{\mathscr{E}}\right]\\ &\quad= \mathsf{E}\left[\prod_{q=1}^{d}\prod_{j\geq 1} e^{-\theta_{jq}g_q(\tilde{X}_j)}\cdot\prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i\notin\mathcal{B}_{\ell q}}\prod_{j\geq 1} P(0;\theta_{jq})^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))}\right.\\ &\qquad\left.\times\prod_{q=1}^{d}\prod_{\ell=1}^{k}\prod_{i\in\mathcal{B}_{\ell q}}\Big(1-\prod_{j\geq 1}(1-P(a_{i\ell q};\theta_{jq}))^{\delta_{\tilde{X}_j}(B_\varepsilon(X_\ell))}\Big)\cdot\prod_{q=1}^{d}\prod_{j\geq 1}\prod_{i=1}^{n_q} P(0;\theta_{jq})^{\delta_{\tilde{X}_j}(\mathbb{X}^*)}\right].\end{aligned}$$

Along similar lines as in the proof of Theorem 5.3, we end up with:

$$\begin{aligned}&\mathsf{E}\left[\exp\left\{-\sum_{q=1}^{d}\int_{\mathbb{X}} g_q(x)\tilde{\mu}_q(\mathrm{d}x)\right\}\cdot\mathbb{1}_{\mathscr{E}}\right]\\ &\quad= \exp\left\{-\int_{\mathbb{X}^*}\int_{(0,\infty)^d}\left(1-\prod_{q=1}^{d}\left(e^{-\theta_q g_q(x)}\prod_{i=1}^{n_q}P(0;\theta_q)\right)\right)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\,\lambda G_0(\mathrm{d}x)\right\}\\ &\qquad\times\prod_{\ell=1}^{k}\left\{\int_{B_\varepsilon(X_\ell)}\int_{(0,\infty)^d}\prod_{q=1}^{d}\Big(e^{-\theta_q g_q(x)}\prod_{i\notin\mathcal{B}_{\ell q}}P(0;\theta_q)\right.\\ &\qquad\qquad\left.\times\prod_{i\in\mathcal{B}_{\ell q}}P(a_{i\ell q};\theta_q)\Big)\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\,\lambda G_0(\mathrm{d}x)\right\}+o\Big(\prod_{\ell=1}^{k}G_0(B_\varepsilon(X_\ell))\Big).\end{aligned}$$

By using the just derived expression and $\mathsf{P}(\mathscr{E})$, available in the proof of Theorem 5.3, we can compute the limit in (5.35). Since $G_0$ is non-atomic, standard limiting arguments lead

to

$$\begin{aligned}\mathcal{L}_{\tilde{\boldsymbol{\mu}}\,|\,\boldsymbol{Z}}(g_1,\ldots,g_d) &= \exp\left\{-\int_{(0,\infty)^d\times\mathbb{X}}\left(1-\prod_{q=1}^{d}e^{-\theta_q g_q(x)}\right)\rho'_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\,\lambda G_0(\mathrm{d}x)\right\} \\ &\quad\times\prod_{\ell=1}^{k}\int_{(0,\infty)^d}\prod_{q=1}^{d}e^{-\theta_q g_q(X_\ell)}\cdot H_{\ell q}(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d),\end{aligned} \tag{5.36}$$

where $\rho'_d(\cdot)$ and $H_{\ell q}(\cdot)$ have been defined in the statement of the present theorem. First, the exponential term in (5.36) corresponds to the Laplace functional of a CRV, say $\boldsymbol{\mu}'$, having Lévy intensity measure given by

$$\nu'(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d\,\mathrm{d}x)=\prod_{q=1}^{d}\prod_{i=1}^{n_q}P(0;\theta_q)\cdot\rho_d(\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\,\lambda G_0(\mathrm{d}x).$$

Second, the product over $\ell=1,\ldots,k$ in (5.36) corresponds to the Laplace functional of the vector of random measures

$$\boldsymbol{\mu}^*(\cdot)=\sum_{\ell=1}^{k}(\theta^*_{\ell 1},\ldots,\theta^*_{\ell d})\,\delta_{X_\ell}(\cdot),$$

where the vectors $(\theta^*_{\ell 1},\ldots,\theta^*_{\ell d})$ are independent across $\ell=1,\ldots,k$, with distribution $H_{\ell q}(\cdot)$. Thus, the thesis follows. □

For completeness, we finally provide the predictive distribution of a vector of new observations $(Z_{(n_1+1)1},\ldots,Z_{(n_d+1)d})$, conditionally to $\boldsymbol{Z}$.

**Theorem 5.5** (Predictive distribution)**.** *Let $\boldsymbol{Z}$ be a sample from the statistical model* (5.23)*. If $\boldsymbol{Z}$ displays $K_n=k$ distinct traits labeled $X_1,\ldots,X_k$, with associated counts $\boldsymbol{a}$, then the predictive distribution of $(Z_{(n_1+1)1},\ldots,Z_{(n_d+1)d})$ satisfies the distributional equality*

$$(Z_{(n_1+1)1},\ldots,Z_{(n_d+1)d})\,|\,\boldsymbol{Z}\overset{d}{=}(Z^*_{(n_1+1)1},\ldots,Z^*_{(n_d+1)d})+(Z'_{(n_1+1)1},\ldots,Z'_{(n_d+1)d}), \tag{5.37}$$

*where the vectors on the right-hand side are independent, and in addition:*

*(i) the generic component $Z^*_{(n_q+1)q}$ is defined as $Z^*_{(n_q+1)q}(\cdot)=\sum_{\ell=1}^{k}A^*_{(n_q+1)\ell q}\delta_{X_\ell}(\cdot)$, for $q=1,\ldots,d$, and the random variables $A^*_{(n_q+1)\ell q}$ are independent with distribution $P(\mathrm{d}a;\theta^*_{\ell q})$, where the $\theta^*_{\ell q}$ are given in point (i) of Theorem 5.4;*

*(ii) the generic component $Z'_{(n_q+1)q}$ is defined as $Z'_{(n_q+1)q}\,|\,\boldsymbol{\mu}'\overset{ind}{\sim}\mathrm{CP}(\mu'_q)$, for $q=1,\ldots,d$, where $\boldsymbol{\mu}'$ is the FCRV defined in point (ii) of Theorem 5.4.*

*Proof.* The thesis follows by a straightforward application of Bayes formula and Theorem 5.4. □

To conclude this section, we provide some details about how Theorems 5.1-5.2 follow as simple corollaries of Theorems 5.3-5.4, respectively. Indeed, as already commented in Section 5.A.1, the FCRVs analyzed in the main body are special examples of CRVs such that

$\rho_d = H^{(d)}$, where $H^{(d)}(\cdot)$ is a probability distribution on $(0,\infty)^d$. In the particular case of Theorems 5.1-5.2, we focus on the special choice where $H^{(d)}(\cdot\,;\psi) = H(\cdot\,;\psi)\times\cdots\times H(\cdot\,;\psi)$, with $H(\cdot\,;\psi)$ any probability distribution on $(0,\infty)$. Thus, Theorems 5.1-5.2 follow by specializing Theorems 5.3-5.4 to this situation.

**Remark 5.3.** *In addition to the example of FCRVs discussed in the main body, there are many other tractable classes of CRVs that can be analyzed based on the general theory of the present section, and that are of potential interest in applications. For example, Theorems 5.3-5.4-5.5 may be applied to obtain marginal, posterior, and predictive distributions of additive CRMs introduced by Lijoi et al. (2014). Another example of interest that can be addressed with the provided theory relates to compound random measures (Griffin and Leisen, 2017).*

### 5.B.1 Extension of the modeling framework to general parameter space

As highlighted in Remark 5.1, throughout the chapter we assume that the parametric distribution $P(\cdot\,;\theta)$ is governed by a single positive parameter $\theta > 0$. This is related to the use of the technical tool of CRVs for the general proofs in Section 5.B. However, this assumption can be relaxed with basically no additional cost, allowing for a general parameter space $\mathbb{S}$, just by moving from CRVs to Poisson processes. To be rigorous, this shift only calls for a slight change of notation, which we describe next. Starting from the exchangeable trait allocation model expressed by (5.1)-(5.2)-(5.3), it is here convenient to organize the parameters $(\theta_j)_{j\geq 1}$ in a point process $\tilde{\Psi}(\cdot) = \sum_{j\geq 1}\delta_{(\tilde{X}_j,\theta_j)}(\cdot)$ on $\mathbb{X}\times\mathbb{S}$, i.e., a random (locally finite) counting measure on $\mathbb{X}\times\mathbb{S}$, with $\mathbb{S}$ a Polish space. Remarkably, note that $\tilde{\Psi}$ and the discrete measure $\tilde{\mu}$ in (5.3) contain the exact same information. In the partially exchangeable trait allocation models in (5.4), which are the main focus of the general theory, the group-specific parameters of interest can be similarly collected in the point processes $\tilde{\Psi}_q(\cdot) = \sum_{j\geq 1}\delta_{(\tilde{X}_j,\theta_{jq})}(\cdot)$, for $q = 1,\ldots,d$, instead of using the discrete measures $\tilde{\mu}_q$. With a slight abuse of notation, we can generalize model (5.4) by writing

$$\begin{aligned} Z_{iq}\,|\,\tilde{\mu}_q &\overset{\text{ind}}{\sim} \text{CP}(\tilde{\Psi}_q), \qquad i\geq 1, \quad q = 1,\ldots,d, \\ (\tilde{\Psi}_1,\ldots,\tilde{\Psi}_d) &\sim \mathcal{Q}_d, \end{aligned} \tag{5.38}$$

where $\mathcal{Q}_d$ denotes here the de Finetti measure of the vector of point processes $(\tilde{\Psi}_1,\ldots,\tilde{\Psi}_d)$. The generalization with respect to model (5.4) is only in that $\theta_{jq}\in\mathbb{S}$ instead of $\theta_{jq} > 0$. With a slight abuse of notation, we will use without distinction the vector $(\tilde{\Psi}_1,\ldots,\tilde{\Psi}_d)$ of point processes on $\mathbb{X}\times\mathbb{S}$ and the point process $\tilde{\boldsymbol{\Psi}} = \sum_{j\geq 1}\delta_{(\tilde{X}_j,\theta_{j1},\ldots,\theta_{jd})}$ on $\mathbb{X}\times\mathbb{S}^d$. The natural choice for the prior distribution of $\tilde{\boldsymbol{\Psi}}$ in this setting is the class of Poisson point processes on $\mathbb{X}\times\mathbb{S}^d$. Indeed, in the special case where $\mathbb{S} = (0,\infty)$, the class of Poisson point processes induces the class of CRV priors for the vector of random measures $\tilde{\boldsymbol{\mu}}$ in (5.23). Therefore, model (5.38) with $\tilde{\boldsymbol{\Psi}}$ distributed as a Poisson point process on $\mathbb{X}\times\mathbb{S}^d$ generalizes the main model (5.23), for a general parameter space $\mathbb{S}$.

A point process $\tilde{\boldsymbol{\Psi}}$ on $\mathbb{X}\times\mathbb{S}^d$, like any random measure, is uniquely characterized by its Laplace functional

$$\mathcal{L}_{\tilde{\boldsymbol{\Psi}}}(f) := \mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{S}^d} f(x,\theta_1,\ldots,\theta_d)\tilde{\boldsymbol{\Psi}}(\mathrm{d}x\,\mathrm{d}\theta_1\,\ldots\,\mathrm{d}\theta_d)\right\}\right],$$

for any measurable function $f : \mathbb{X} \times \mathbb{S}^d \to (0, \infty)$. The class of Poisson point processes is characterized by the following representation of the Laplace functional,

$$\mathcal{L}_{\tilde{\boldsymbol{\Psi}}}(f) = \exp\left\{ -\int_{\mathbb{X}\times\mathbb{S}^d} \left(1 - \exp\left\{-f(x, \theta_1, \ldots, \theta_d)\right\}\right) \nu(\mathrm{d}x\, \mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) \right\},$$

where $\nu(\,\cdot\,)$ is a locally finite measure on $\mathbb{X} \times \mathbb{S}^d$ referred to as the intensity of $\tilde{\boldsymbol{\Psi}}$. To link with the general theory and notation of CRVs in Section 5.B, assume that the intensity measure factorizes as $\nu(\mathrm{d}x\, \mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) = \lambda \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) G_0(\mathrm{d}x)$, where $\lambda > 0$ and $G_0(\,\cdot\,)$ is a non-atomic probability measure on $\mathbb{X}$. We write $\tilde{\boldsymbol{\Psi}} \sim \text{PP}(\rho_d, \lambda, G_0)$. Then, Theorems 5.3-5.4-5.5 hold identically with the only replacement of $\mathbb{S}$ instead of $(0, \infty)$ as parameter space. For completeness, we report here the most general formulation of Theorems 5.3-5.4 for the general parameter space $\mathbb{S}$.

**Theorem 5.6** (Marginal distribution, PETPF)**.** *Let $\boldsymbol{Z}$ be a sample from the statistical model* (5.38)*, with $\tilde{\boldsymbol{\Psi}} \sim \text{PP}(\rho_d, \lambda, G_0)$. The probability that $\boldsymbol{Z}$ displays $K_n = k$ distinct traits with counts $\boldsymbol{A} = \boldsymbol{a}$ equals*

$$\pi_n(\boldsymbol{a}) = \frac{\lambda^k}{k!} \exp\left\{ -\lambda \int_{\mathbb{S}^d} \left(1 - \prod_{q=1}^{d} P(0; \theta_q)^{n_q}\right) \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) \right\} \times \prod_{\ell=1}^{k} \int_{\mathbb{S}^d} \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(a_{i\ell q}; \theta_q) \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d)$$

*where $n = \sum_{q=1}^{d} n_q$ and $\boldsymbol{n} = (n_1, \ldots, n_d)$ are the sample sizes.*

**Theorem 5.7** (Posterior distribution)**.** *Let $\boldsymbol{Z}$ be a sample from the statistical model* (5.38)*, with $\tilde{\boldsymbol{\Psi}} \sim \text{PP}(\rho_d, \lambda, G_0)$. If $\boldsymbol{Z}$ displays $K_n = k$ distinct traits labeled $X_1, \ldots, X_k$, with associated counts $\boldsymbol{a}$, then the posterior distribution of $\tilde{\boldsymbol{\Psi}}$ satisfies the distributional equality*

$$(\tilde{\Psi}_1, \ldots, \tilde{\Psi}_d) \,|\, \boldsymbol{Z} \stackrel{d}{=} (\Psi_1^*, \ldots, \Psi_d^*) + (\Psi_1', \ldots, \Psi_d'),$$

*where $(\Psi_1^*, \ldots, \Psi_d^*)$ and $(\Psi_1', \ldots, \Psi_d')$ are independent random vectors such that*

*(i) the generic $\Psi_q^*(\cdot)$ is defined as $\Psi_q^*(\cdot) = \sum_{\ell=1}^{k} \delta_{(X_\ell, \theta_{\ell q}^*)}(\cdot)$, for $q = 1, \ldots, d$, and the random vectors $(\theta_{\ell 1}^*, \ldots, \theta_{\ell d}^*)$ are independent across $\ell = 1, \ldots, k$, with distribution*

$$H_{\ell q}(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) \propto \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(a_{i\ell q}; \theta_q) \cdot \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d);$$

*(ii) the vector of point processes $(\Psi_1', \ldots, \Psi_d')$ is a $\text{PP}(\rho_d', \lambda, G_0)$, with*

$$\rho_d'(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d) = \prod_{q=1}^{d} \prod_{i=1}^{n_q} P(0; \theta_q) \cdot \rho_d(\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_d).$$

The computations we followed in Section 5.B to prove Theorems 5.3-5.4 hold for the more general model (5.38), with $\tilde{\boldsymbol{\Psi}} \sim \text{PP}(\rho_d, \lambda, G_0)$, with trivial modifications such as the use of Laplace functionals of Poisson point processes instead of the ones of CRVs. The finite-dimensional trait models analyzed in Section 5.2, corresponding to $\tilde{\boldsymbol{\mu}} \sim \text{FCRV}(H^{(d)}, \lambda, G_0)$,

are generalized to a generic parameter space via finite Poisson point processes, i.e., $\tilde{\boldsymbol{\Psi}} \sim \text{PP}(\rho_d, \lambda, G_0)$ where $\rho_d = H^{(d)}$ and $H^{(d)}$ is a probability distribution on $\mathbb{S}^d$. As in Section 5.2, we commonly assume $H^{(d)}(\cdot\,;\psi) = H(\cdot\,;\psi) \times \cdots \times H(\cdot\,;\psi)$, with $H(\cdot\,;\psi)$ a probability distribution on $\mathbb{S}$, parameterized by $\psi$.

To conclude this discussion on the extension of the proposed modeling framework to general parameter space $\mathbb{S}$ for the count distribution $P(\cdot;\theta)$, $\theta \in \mathbb{S}$, we provide an example. Besides specifying $P(\cdot;\theta)$, we derive the marginal of a sample $\boldsymbol{Z}$ and the posterior distribution of $\tilde{\boldsymbol{\Psi}}$, assuming $\tilde{\boldsymbol{\Psi}} \sim \text{PP}(H^{(d)}, \lambda, G_0)$, and $H^{(d)}(\cdot\,;\psi) = H(\cdot\,;\psi) \times \cdots \times H(\cdot\,;\psi)$. This results are derived by specializing the previous general theorems.

**Example 5.7** (Zero-inflated shifted negative binomial counts)**.** In the setting of count data with support on $\{0, 1, 2, \dots\}$, we propose to consider a zero-inflated shifted negative binomial (ZI-SNB) distribution. More precisely, let $X$ be a negative binomial random variable with parameters $c > 0$ (number of trials) and $p \in (0, 1]$ (success probability). We say that $Y := X + 1$ has a shifted negative binomial random distribution, whose probability mass function $p_{\text{sNB}}(\,\cdot\,; c, p)$, depending on the parameters $(c, p)$, is supported by the set of natural numbers. Then, we specify the distribution $P(\,\cdot\,;\theta)$ as $P(\,\cdot\,;\theta) = (1-w)\delta_0(\,\cdot\,) + w\, p_{\text{sNB}}(\,\cdot\,; c, p)$, with $\theta = (c, w, p)$ and $w \in (0, 1)$. This choice allows to independently model the presence of a trait and its associated measurement. That is, the pattern of occurrence of traits is captured as in the binary traits setting. On top of that, when a trait is displayed, the shifted negative binomial distribution models the magnitude of the associated expression. The shifted version of the negative binomial is considered to correctly identify the actual expression of a trait.

To complete the specification of this model, we consider $H(\cdot\,;\psi)$ such that $c$ is fixed, with $w$ and $p$ following independent beta laws with parameters $(a_w, b_w)$ and $(a_p, b_p)$, respectively, so that $\psi = (c, a_w, b_w, a_p, b_p)$. In this case, the marginal distribution of $\boldsymbol{Z}$ depends on the whole collection of counts $\boldsymbol{a}$ and equals

$$\begin{aligned}\pi_n(\boldsymbol{a}; \lambda, \psi) = \frac{\lambda^k}{k!} \exp\left\{-\lambda\left[1 - \prod_{q=1}^{d} \frac{B(a_w, b_w + n_q)}{B(a_w, b_w)}\right]\right\} \\ \times \prod_{\ell=1}^{k}\prod_{q=1}^{d}\prod_{i:a_{i\ell q}>1} \binom{a_{i\ell q} + c - 2}{a_{i\ell q} - 1} \frac{b_{\ell q}^{(p)}}{B(a_p, b_p)} \frac{b_{\ell q}^{(w)}}{B(a_w, b_w)},\end{aligned} \tag{5.39}$$

where $b_{\ell q}^{(p)} = B(a_p + c\, m_{\ell q}, b_p + \sum_{i=1}^{n_q} a_{i\ell q} - m_{\ell q})$ and $b_{\ell q}^{(w)} = B(a_w + m_{\ell q}, b_w + n_q - m_{\ell q})$.

Moving to the posterior distribution of $\tilde{\boldsymbol{\Psi}}$, each $\theta_{\ell q}^*$ in Theorem 5.7 is a vector of three component $(c, w_{\ell q}^*, p_{\ell q}^*)$, where $w_{\ell q}^*$ has a beta distribution with parameters $(a_w + m_{\ell q}, b_w + n_q - m_{\ell q})$, and $p_{\ell q}^*$ is again a beta with parameters $(a_p + c\, m_{\ell q}, b_p + \sum_{i=1}^{n_q} a_{i\ell q} - m_{\ell q})$, further independent between them. In addition, for each $q = 1, \dots, d$,

$$\Psi_q'(\cdot) = \sum_{j=1}^{M'} \delta_{(\tilde{X}_j', \theta_{jq}')}(\cdot), \qquad \theta_{jq}' \overset{\text{iid}}{\sim} H_q', \qquad \tilde{X}_j' \overset{\text{iid}}{\sim} G_0, \qquad j = 1, \dots, M',$$

where $H_q'$ is such that $c$ is fixed, $w_{jq}'$ has a beta distribution with parameters $(a_w, b_w + n_q)$,

independent of $p'_{jq}$, which follows the prior beta law. Moreover, $M' \sim \text{Poisson}(\lambda')$ with

$$\lambda' = \lambda \prod_{q=1}^{d} B(a_w, b_w + n_q)/B(a_w, b_w).$$

## 5.C Proofs of Section 5.3

The main theoretical result of Section 5.3 is Proposition 5.1, which illustrates the effect on clustering estimation of accounting for potentially unseen traits by comparing the proposed model in (5.15) with and the naïve model almost identical to (5.15), in which it is assumed that there are no unseen traits, i.e. we suppose $M = k$. Before showing the proof of Proposition 5.1, we recall a lemma which is instrumental for it.

**Lemma 5.1.** *Let $X$ be an almost surely non-negative random variable and $n = 1, 2, \ldots$. The following holds true:*

$$\mathsf{E}(X^{n+1}) \geq \mathsf{E}(X^n)\mathsf{E}(X).$$

*Proof.* For any $n = 1, 2, \ldots$, by Holder inequality, it holds that

$$\mathsf{E}(X^n) \leq \mathsf{E}(X^{n+1})^{\frac{n}{n+1}} = \mathsf{E}(X^{n+1})^{1-\frac{1}{n+1}},$$

and consequently

$$\mathsf{E}(X^n)\mathsf{E}(X^{n+1})^{\frac{1}{n+1}} \leq \mathsf{E}(X^{n+1}). \tag{5.40}$$

By Jensen inequality,

$$\mathsf{E}(X^{n+1}) \geq \mathsf{E}(X)^{n+1}$$

and, from (5.40), we get

$$\mathsf{E}(X^{n+1}) \geq \mathsf{E}(X^n)\mathsf{E}(X^{n+1})^{\frac{1}{n+1}} \geq \mathsf{E}(X^n)\mathsf{E}(X).$$

The thesis is proven. □

### 5.C.1 Proof of Proposition 5.1

In Proposition 5.1, we compare the predictive allocation probabilities for a generic subject $i$ under the proposed model defined in (5.15), with a Pitman-Yor prior for $\xi_h$, and a naïve model almost identical to (5.15), in which it is assumed that there are no unseen traits, i.e. we suppose $M = k$. Recalling the notation used in the statement, let

$$p_{iq} = \mathbb{P}(\mu_i = \tilde{\mu}_q \,|\, \boldsymbol{Z}, \boldsymbol{\mu}_{-i}), \qquad p_{i,\text{new}} = \mathbb{P}(\mu_i = \text{“new”} \,|\, \boldsymbol{Z}, \boldsymbol{\mu}_{-i}), \qquad i = 1, \ldots, n,$$

where $\boldsymbol{\mu}_{-i} = (\mu_1, \ldots \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n)$ with $d_{-i}$ distinct values $\tilde{\mu}_1, \ldots, \tilde{\mu}_{d_{-i}}$, for $q = 1, \ldots, d_{-i}$. Here, $p_{iq}$ denotes the predictive probability that subject $i$ belongs to the $q$th cluster formed by the remaining subjects, and $p_{i,\text{new}}$ is the predictive probability for subject $i$ to form its own cluster. Similarly, we define $p^*_{iq}$ and $p^*_{i,\text{new}}$ for the naïve model.

For the proposed model defined in (5.15), the predictive allocation probabilities $p_{iq}$ and $p_{i,\text{new}}$ are expressed in (5.17). For the associated naïve model, since this is identical to (5.15) with the specification $M = k$, the predictive allocation probabilities $p^*_{iq}$ and $p^*_{i,\text{new}}$ are obtained from (5.17) by replacing $\pi_n(\boldsymbol{a}; \lambda, \psi)$ with $\pi_n(\boldsymbol{a}; M = k, \psi)$ in (5.8).

Consider the ratio between the probability that subject $i$ creates a new cluster and the probability that it is allocated to the $q$-th existing cluster. For the proposed model defined in (5.15), this is equal to

$$\frac{p_{i,\text{new}}}{p_{iq}} = \frac{\pi_n(\boldsymbol{a}_{i,\text{new}}; \lambda, \psi)(\gamma + d_{-i}\sigma)/(n+\gamma-1)}{\pi_n(\boldsymbol{a}_{iq}; \lambda, \psi)(n_{q,-i} - \sigma)/(n+\gamma-1)},$$

while the same quantity, for the naïve model, is equal to

$$\frac{p^*_{i,\text{new}}}{p^*_{iq}} = \frac{\pi_n(\boldsymbol{a}_{i,\text{new}}; M=k, \psi)(\gamma + d_{-i}\sigma)/(n+\gamma-1)}{\pi_n(\boldsymbol{a}_{iq}; M=k, \psi)(n_{q,-i} - \sigma)/(n+\gamma-1)}.$$

Denote with $\boldsymbol{n}_{i,\text{new}}$ the sample sizes associated with $\boldsymbol{a}_{i,\text{new}}$, and with $\boldsymbol{n}_{iq}$ the sample sizes associated with $\boldsymbol{a}_{iq}$. By comparing the previous two ratios, we obtain

$$\begin{aligned}
\frac{p_{i,\text{new}}/p_{iq}}{p^*_{i,\text{new}}/p^*_{iq}} &= \frac{\pi_n(\boldsymbol{a}_{i,\text{new}}; \lambda, \psi)}{\pi_n(\boldsymbol{a}_{iq}; \lambda, \psi)} \cdot \frac{\pi_n(\boldsymbol{a}_{iq}; M=k, \psi)}{\pi_n(\boldsymbol{a}_{i,\text{new}}; M=k, \psi)} \\
&= \frac{\exp\left\{-\lambda\left(1 - \prod_{h=1}^{d_{-i}+1}\int_{(0,\infty)} P(0;\theta)^{(\boldsymbol{n}_{i,\text{new}})_h} H(\mathrm{d}\theta;\psi)\right)\right\}}{\exp\left\{-\lambda\left(1 - \prod_{h=1}^{d_{-i}}\int_{(0,\infty)} P(0;\theta)^{(\boldsymbol{n}_{iq})_h} H(\mathrm{d}\theta;\psi)\right)\right\}} \\
&= \exp\left\{-\lambda\left[\prod_{h=1}^{d_{-i}}\int_{(0,\infty)} P(0;\theta)^{(\boldsymbol{n}_{iq})_h} H(\mathrm{d}\theta;\psi) - \prod_{h=1}^{d_{-i}+1}\int_{(0,\infty)} P(0;\theta)^{(\boldsymbol{n}_{i,\text{new}})_h} H(\mathrm{d}\theta;\psi)\right]\right\} \\
&= \exp\left\{-\lambda \prod_{\substack{h=1\\h\neq q}}^{d_{-i}}\int_{(0,\infty)} P(0;\theta)^{n_{h,-i}} H(\mathrm{d}\theta;\psi) \times \left[\int_{(0,\infty)} P(0;\theta)^{n_{q,-i}+1} H(\mathrm{d}\theta;\psi)\right.\right. \\
&\qquad \left.\left. - \int_{(0,\infty)} P(0;\theta) H(\mathrm{d}\theta;\psi) \int_{(0,\infty)} P(0;\theta)^{n_{q,-i}} H(\mathrm{d}\theta;\psi)\right]\right\}.
\end{aligned}$$

Using the inequality in Lemma 5.1, it holds that

$$\int_{(0,\infty)} P(0;\theta)^{n_{q,-i}+1} H(\mathrm{d}\theta;\psi) - \int_{(0,\infty)} P(0;\theta) H(\mathrm{d}\theta;\psi) \int_{(0,\infty)} P(0;\theta)^{n_{q,-i}} H(\mathrm{d}\theta;\psi) \geq 0,$$

and consequently

$$\frac{p_{i,\text{new}}}{p_{iq}} < \frac{p^*_{i,\text{new}}}{p^*_{iq}}.$$

The proof is complete. Note that the result is not restricted to the Pitman-Yor prior for $\xi_h$; the same argument holds for any arbitrary prior on $\xi_h$.

## 5.D Additional details about the natural competitor

The *negative binomial mixture of* BBs, proposed in Chapter 3, is the natural competitor for estimating the number of unseen features (here, traits), though it is specifically designed for homogeneous exchangeable settings. It belongs to the class of Example 5.1, and in fact corresponds to the binary traits specification of model (5.4) with prior given in equations (5.5)–(5.6), under the special case $d = 1$. In particular, $H(\cdot;\psi)$ is the beta distribution with parameters $(-\alpha, \alpha+\beta)$. The beta parameters $(a, b) = (-\alpha, \alpha+\beta)$ are

equipped with independent gamma priors with hyperparameters $(\alpha_a, \beta_a)$ and $(\alpha_b, \beta_b)$, respectively. The model parameter $\lambda$, governing the Poisson-distributed total number of meetings, is assigned a gamma prior with parameters $(\alpha_\lambda, \beta_\lambda)$, as detailed in Section 5.2.3.

For all the comparisons we provide in the chapter, hyperparameters are fixed following Chapter 3, with the prior variance for $M$ chosen to be comparable to that used in the specific applications of Sections 5.4.2 and 5.5. Posterior inference relies on 100,000 iterations, with the first 10,000 samples discarded as burn-in and a thinning interval of 5.

# 6 Palm distributions of superposed point processes for statistical inference

This final chapter builds on a novel probabilistic result in point process theory. Unlike the preceding chapters, its focus differs from the main theme of the thesis, namely feature allocation models, and is primarily centered on the theoretical development of this result. The chapter then explores several statistical applications, including one that naturally connects to feature allocation models. In particular, the novel result enables inference for a specific instance of an *extended feature model* within the class introduced in Chapter 4. This application can thus be viewed as an additional original example complementing those discussed in Chapter 4, Section 4.4. It should be noted that this application is still under development, and simulation studies and real-data applications are left for future work.

## 6.1 Introduction

Real-world point patterns often combine several structured components, some regular or inhomogeneous, others explicitly clustered, and may also include random noise. Semiconductor wafer defect maps (Borgoni et al., 2021), disease-case locations in epidemiology (Meyer et al., 2017), cellular network base-station layouts (Choi et al., 2017), mixed age tree stands in ecology (Ngo Bieng et al., 2011), and earthquake aftershock sequences (Ogata, 1998) can all be regarded as superpositions of two (or more) independent point processes. While the superposition operation is trivial at the model level, from an inferential standpoint, superposed point processes are notoriously awkward. Standard tools such as minimum contrast estimation rely on closed-form expressions for second-order summaries, such as the $J$ function, Ripley's $K$-function, and Besag's $L$ function. However, those summaries remain unknown for a generic superposition (Møller and Waagepetersen, 2003; Diggle, 2013). Hence, practitioners need to rely on complex algorithms often developed on a case-by-case basis (e.g., Tanaka and Ogata, 2014; Xu et al., 2018).

In this chapter, we focus on the Palm distributions of the superposition of two independent point processes. Palm distributions (Baccelli et al., 2020) are key mathematical objects in the study of point processes, describing the conditional behavior of a process given the location of one or more of its points or *atoms*. We establish a mixture representation for the Palm distributions of the superposed process, in which the mixture components are sums of the two original processes and their Palm versions, weighted by their respective mean measures. Our analysis easily extends to the superposition of more than two independent processes.

We demonstrate the practical usefulness of our result for statistical inference in several contexts. First, we consider fitting a determinantal point process (DPP, Lavancier et al., 2015) contaminated by random background noise via minimum contrast estima-

tion. Indeed, through the Palm distributions, it is straightforward to obtain Ripley's $K$ function, enabling robust and fast inference via minimum contrast. This methodology is particularly relevant in applied contexts such as the analysis of spatial defect structures in semiconductor manufacturing (Borgoni et al., 2021).

Second, we address parameter estimation for the shot noise Cox process (SNCP, Møller, 2003), a prominent class of cluster processes. Cluster processes are routinely employed in several applied areas, such as astronomy, materials science, and plant ecology; see, e.g., Illian et al. (2008); Møller and Waagepetersen (2003). Despite their generality and usefulness, statistical inference for general SNCP models is lacking, with most of the contributions focusing on simple sub-classes (see Wang et al., 2023, for a recent contribution) or on simulation of the process with fixed parameters (Møller and Waagepetersen, 2003). An application of our main result yields the higher-order Palm distributions of the SNCP, previously unknown in the literature. Building on this result, we obtain an explicit expression for the Janossy density, which in turn enables maximum likelihood estimation via an expectation-maximization algorithm. We demonstrate the effectiveness of such an approach on the gamma SNCPs with Thomas kernel (Møller and Waagepetersen, 2003). We also disprove a conjecture by Coeurjolly et al. (2017) showing that the higher-order Palm distributions of the SNCP do not match the law of a SNCP.

Finally, we outline an application of the SNCP within the framework of extended feature allocation models introduced in Chapter 4. This formulation allows for clustering features according to their similarity, while simultaneously retaining the ability to estimate unseen features. This extension may be relevant in spatial contexts, such as the application discussed in Section 4.5, where, for instance, trees can be grouped based on the proximity of their locations. This connection highlights the broader relevance of the proposed theory to the recurring themes of this thesis.

Although not explored here, our theoretical analysis is helpful in several other statistical contexts. For example, in Bayesian nonparametrics, the superposition of point processes is used to define a prior distribution when data is subdivided into groups (Lijoi et al., 2014; Griffin et al., 2013). Our main result can be leveraged for posterior analysis and numerical computations.

## 6.2 Superposition of point processes

### 6.2.1 Background and notation for point processes

Let $\mathbb{X}$ be a Polish space equipped with the corresponding Borel $\sigma$-algebra $\mathscr{X}$. A point process $\Phi$ on $\mathbb{X}$ can be represented as $\Phi = \sum_{j\geq 1} \delta_{\tilde{X}_j}$, where $(\tilde{X}_j)_{j\geq 1}$ is a sequence of random variables (*atoms*) taking values in $\mathbb{X}$, and $\delta_x$ denotes the Dirac delta mass at $x$. The probability distribution of $\Phi$ is denoted with $\mathbf{P}_\Phi$. The number of atoms in $\Phi$ could be either finite or infinite. In the sequel, we will follow the approach of (Baccelli et al., 2020), where $\Phi$ is regarded as a random counting measure. See Appendix 6.A for further mathematical details, including the $\sigma$-algebra on the space of counting measures.

Let $M_\Phi(B) = \mathsf{E}[\Phi(B)]$, for $B \in \mathscr{X}$, be the mean measure of $\Phi$, and define the $k$-th factorial moment measure $M_\Phi^{(k)}$ as the mean measure of the $k$-th factorial power of $\Phi$, i.e.,

of the point process $\Phi^{(k)}$ defined as:

$$\Phi^{(k)} := \sum_{(j_1,\dots,j_k)}^{\neq} \delta_{(\tilde{X}_{j_1},\dots,\tilde{X}_{j_k})},$$

where the symbol $\neq$ means that the sum is extended over all pairwise distinct indexes.

To introduce the notion of *Palm distributions*, let us define the Campbell measure of $\Phi$, namely $\mathscr{C}_\Phi(B \times L) := \mathsf{E}[\Phi(B)\mathbb{1}(\Phi \in L)]$, where $B \in \mathscr{X}$ and $L$ is an element of the appropriate $\sigma$-algebra on the space of random counting measures (see Appendix 6.A). Under the assumption that $M_\Phi$ is $\sigma$-finite, it can be shown that $\mathscr{C}_\Phi$ admits the following representation

$$\mathscr{C}_\Phi(B \times L) = \int_B \mathbf{P}_\Phi^x(L) M_\Phi(\mathrm{d}x),$$

where $\{\mathbf{P}_\Phi^x\}_{x\in\mathbb{X}}$ is the almost surely (a.s.) unique disintegration probability kernel of $\mathscr{C}_\Phi$ with respect to $M_\Phi$, and it is referred to as the Palm kernel of $\Phi$. For fixed $x \in \mathbb{X}$, $\mathbf{P}_\Phi^x$ is a probability distribution over the space of counting measures, termed the Palm distribution of $\Phi$ at $x$, and thus it can be identified with the law of a point process $\Phi_x \sim \mathbf{P}_\Phi^x$, which is consequently called a *Palm version* of $\Phi$ at $x$. By Proposition 3.1.12 in (Baccelli et al., 2020), for $M_\Phi$-almost all $x \in \mathbb{X}$, the point process $\Phi_x$ contains the atom $x$ with probability 1. This justifies the interpretation of the Palm distribution of $\Phi$ at $x$ as the law of $\Phi$ conditionally to $\Phi$ having an atom at $x$. In addition, the point process $\Phi_x^! := \Phi_x - \delta_x$ is well-defined, and $\Phi_x^!$ is called a *reduced Palm version* of $\Phi$ at $x$. Finally, it is possible to extend the definition of Palm distributions to multiple conditioning points $\boldsymbol{x} = (x_1, \dots, x_k) \in \mathbb{X}^k$. In this case, the Palm distribution of $\Phi$ at $\boldsymbol{x}$ is interpreted as the probability distribution of $\Phi$ conditionally to $\Phi$ having $k$ atoms at locations $x_1, \dots, x_k$. See Appendix 6.A.

### 6.2.2 Palm distributions of the superposition of independent processes

Consider independent point processes $\Phi_i$, $i = 1, \dots, n$. Their superposition $\Phi$ is defined as $\Phi = \sum_{i=1}^n \Phi_i$, i.e., the union of all the point patterns. The following theorem states the main theoretical result of the chapter, which characterizes the Palm distributions of the superposition of two independent point processes.

**Theorem 6.1.** *Let $\Phi_1$ and $\Phi_2$ be two independent point processes on $\mathbb{X}$. Then, for any $x \in \mathbb{X}$, the Palm version $(\Phi_1 + \Phi_2)_x$ can be expressed as the following mixture:*

$$(\Phi_1 + \Phi_2)_x \overset{d}{=} \begin{cases} \Phi_{1x} + \Phi_2 & \textit{with probability equal to } \frac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_\Phi}(x) \\ \Phi_1 + \Phi_{2x} & \textit{with probability equal to } \frac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_\Phi}(x). \end{cases}$$

Theorem 6.1 admits an intuitive interpretation about the Palm distributions of the superposed process. Conditioning the superposed process $\Phi = \Phi_1 + \Phi_2$ on having a point at $x$, its distribution depends on whether $x$ originates from $\Phi_1$ or from $\Phi_2$, thus reflecting the respective contributions of each source. The mixture structure in the Palm version $(\Phi_1 + \Phi_2)_x$ explicitly accounts for the two mutually exclusive scenarios. The mixing probabilities correspond precisely to the probabilities that point $x$ originates from one or the other process.

**Remark 6.1** (Reduced Palm version of superposition)**.** *Exploiting the relation* $\Phi_{ix} \overset{d}{=} \Phi_{ix}^{!} + \delta_x$*, for* $M_{\Phi_i}$*-almost all* $x$*,* $i = 1, 2$*, we can replace all the Palm versions appearing in the statement of Theorem 6.1 with their reduced counterparts.*

Theorem 6.1 can be extended in several directions. In particular, in Appendix 6.B.2, we characterize the Palm distributions of the superposition of more than two independent point processes, as well as the Palm distributions given multiple conditioning points.

## 6.3 Inference for corrupted determinantal process via minimum contrast

### 6.3.1 Minimum contrast estimation

Minimum contrast estimimation methods (MCE, Møller and Waagepetersen, 2003) are a class of techniques for fitting parametric point process models to observed point patterns. In several common settings where the likelihood of the point process is intractable, MCE is the only viable option to perform inference. Let $\Phi$ be a point process whose distribution $\mathbf{P}_\Phi$ is parametrized by a set of parameters $\theta$. Let $s(r;\theta)$, $r \geq 0$, be a functional summary statistics of $\Phi$, for which an analytical expression is available, and let $\hat{s}(r)$ be a nonparametric estimate of $s$ based on the observed data. The MCE for $\theta$ is obtained as

$$\hat{\theta} = \operatorname{argmin}_\theta \int_{r_l}^{r_u} |\hat{s}(r)^q - s(r;\theta)^q|^p \mathrm{d}r. \tag{6.1}$$

Common choices consider $r_l = 0$, $q = 1/2$, $p = 2$ (Diggle, 2013). Among the various options for $s(r;\theta)$, we focus here on Ripley's $K$-function (Ripley, 1976), a canonical choice for stationary point processes; see, e.g., Baddeley et al. (2015) and Illian et al. (2008). Assuming that $\Phi$ is stationary with intensity $M_\Phi \equiv \rho > 0$, the $K$-function of $\Phi$ is defined as

$$K_\Phi(r) = \frac{1}{\rho} \mathsf{E}[\Phi_o^{!}(B(o,r))], \quad \text{for } r \geq 0,$$

where $o$ denotes a generic point of $\mathbb{X}$ and $B(o,r)$ is the ball of radius $r$ centered at $o$, given that $B(o,r) \subset \mathbb{X}$. Thanks to stationariety, $K_\Phi$ is invariant to the choice of point $o$, which is called the *typical point* of $\Phi$: the quantity $\rho K_\Phi(r)$ represents the expected number of points that are $r$-close to a generic point $o$, given that $\Phi$ has an atom in $o$.

An application of Theorem 6.1 yields the $K$-function for the superposition of two independent and stationary point processes $\Phi_i$, $i = 1, 2$ with intensities $M_{\Phi_i} \equiv \rho_i$. Specifically, letting $\rho = \rho_1 + \rho_2$,

$$K_{\Phi_1+\Phi_2}(r) = \frac{1}{\rho}\left[\{K_{\Phi_1}(r)\rho_1 + \rho_2|B(o,r)|\}\frac{\rho_1}{\rho} + \{\rho_1|B(o,r)| + K_{\Phi_2}(r)\rho_2\}\frac{\rho_2}{\rho}\right], \tag{6.2}$$

where $|B(o,r)|$ denotes the volume of the ball in $\mathbb{X}$. Plugging (6.2) into (6.1), we need to solve an optimization problem to estimate the parameters of $\Phi_1$ and $\Phi_2$. Below, we describe an application where $\Phi_1$ is a repulsive point pattern and $\Phi_2$ is a random background noise. Specifically, $\Phi_1$ is assumed to be a DPP; the class of DPPs is briefly recalled in the next section.

### 6.3.2 Determinantal point processes

DPPs are a class of repulsive point processes that enjoy a high degree of analytical tractability; they have been used to model inhibitory structure in spatial datasets and for diversity-promoting sampling in machine learning. See, e.g., Lavancier et al. (2015); Kulesza and Taskar (2012) and references therein.

Let $R \subset \mathbb{R}^2$ be a compact set. A DPP $\xi$ on $R$ is specified by a covariance kernel $C : R \times R \to \mathbb{C}$, such that the $k$-th factorial moment measure $M_\xi^{(k)}$ equals

$$M_\xi^{(k)}(\mathrm{d}x_1 \dots x_k) = \det\{C(x_h, x_w)_{h,w=1,\dots,k}\}\mathrm{d}x_1 \dots \mathrm{d}x_k, \qquad x_1, \dots, x_k \in R,$$

where $C(x_h, x_w)_{h,w=1,\dots,k}$ is the $k \times k$ matrix with entries $C(x_h, x_w)$. In particular, we assume that $\xi$ follows a Gaussian DPP (Lavancier et al., 2015) with kernel $C(x, y) = \rho_\xi \exp\{-\|(x-y)/\alpha\|^2\}$ parametrized by $(\rho_\xi, \alpha)$, with $\rho_\xi < (\pi\alpha^2)^{-1}$ to ensure the process is well-defined. From Lavancier et al. (2015), Ripley's $K$-function of $\xi$ is

$$K_\xi(r) = \pi r^2 - \frac{\pi\alpha^2}{2}\left(1 - \mathrm{e}^{-2r^2/\alpha^2}\right). \tag{6.3}$$

While DPPs allow for likelihood-based inference as discussed in Lavancier et al. (2015), this is typically numerically cumbersome due to a Fourier series expansion and the determinant of large matrices involved in the likelihood. Hence, in the `spatstat` package (Baddeley and Turner, 2005), the default way of fitting a DPP is through MCE based on the $K$-function.

### 6.3.3 Fitting a corrupted determinantal point process via minimum contrast

Assume now to observe a realization of $\xi$ corrupted by a background noise, independent of $\xi$. Let $\Phi_2$ be a homogeneous Poisson point process on $R$ with intensity $\omega$, which models the corrupting noise. This setting naturally fits within our proposed modeling framework. Plugging (6.3) into (6.2), and recalling that for the homogeneous Poisson process $\Phi_2$ we have $K_{\Phi_2}(r) = \pi r^2$, the $K$-function for $\Phi = \xi + \Phi_2$ equals

$$K_{\xi+\Phi_2}(r) = \pi r^2 - \frac{\rho_\xi^2}{(\rho_\xi + \omega)^2}\frac{\pi\alpha^2}{2}\left(1 - \mathrm{e}^{-2r^2/\alpha^2}\right). \tag{6.4}$$

Let $\boldsymbol{x} = (x_1, \dots, x_k) \in R^k$ denote the observed point pattern. The goal is to estimate the parameters of $\Phi = \xi + \Phi_2$, namely the intensity $\rho_\xi$ of the signal process, the repulsion parameter $\alpha$, and the noise intensity $\omega$. As customary, we estimate the overall intensity $\rho = \rho_\xi + \omega$ of $\Phi$ by $\hat{\rho} = k/|R|$, so that we are left with estimating only $\rho_\xi$ and $\alpha$. Following the discussion above, we minimize the objective function in (6.1), where $\hat{s}$ is the edge-corrected nonparametric estimator discussed by Ripley (1976) (see also Møller and Waagepetersen (2003)), and $s$ is the $K$-function in (6.4). Moreover, assuming that we observe data on a rectangular region $R = [a, b] \times [c, d]$, $r_u$ is set to one quarter of the smallest side length of $R$, following Diggle (2013). For numerical purposes, we approximate the integral via numerical quadrature using Simpson's rule. The minimization is performed via the BFGS algorithm using the `Julia` programming language.

We show an illustrative simulation to assess the performance of the MCE approach based on the Ripley's $K$-function in estimating the parameters of $\Phi$. We generate a

| True parameters | | | Gaussian DPP | | Gaussian DPP + Poisson noise | |
|---|---|---|---|---|---|---|
| $\rho_\xi$ | $\alpha$ | $u$ | $\hat{\rho}_\xi$ | $\hat{\alpha}$ | $\hat{\rho}_\xi$ | $\hat{\alpha}$ |
| 50 | 0.06 | 0.2 | 60.59 (6.99) | 0.05 (0.02) | 53.74 (11.56) | 0.06 (0.03) |
| 50 | 0.06 | 0.35 | 68.20 (7.82) | 0.04 (0.02) | 56.72 (14.20) | 0.06 (0.03) |
| 50 | 0.02 | 0.2 | 60.10 (8.19) | 0.02 (0.02) | 46.48 (16.69) | 0.03 (0.04) |
| 50 | 0.02 | 0.35 | 67.64 (9.20) | 0.02 (0.02) | 51.49 (19.40) | 0.03 (0.04) |
| 100 | 0.05 | 0.2 | 120.61 (9.35) | 0.04 (0.01) | 107.25 (15.73) | 0.05 (0.01) |
| 100 | 0.05 | 0.35 | 135.71 (10.53) | 0.03 (0.01) | 111.83 (21.14) | 0.05 (0.01) |
| 100 | 0.025 | 0.2 | 120.18 (11.15) | 0.02 (0.01) | 98.30 (28.35) | 0.03 (0.03) |
| 100 | 0.025 | 0.35 | 135.22 (12.54) | 0.02 (0.01) | 104.31 (33.28) | 0.03 (0.03) |

Table 6.3.1: Mean and standard deviation (in brackets) of the estimates $(\hat{\rho}_\xi, \hat{\alpha})$ of the Gaussian DPP over $1,000$ independent replicated datasets, for different combinations of the true parameters reported in the left column. Middle column: parameter estimates when fitting only the Gaussian DPP. Right column: parameter estimates when fitting the superposition of the Gaussian DPP and the background noise Poisson process.

point pattern from a Gaussian DPP on the unit-square $R$ with parameters $(\rho_\xi, \alpha) \in \{(50, 0.06), (50, 0.02), (100, 0.05), (100, 0.025)\}$, and perturb the point pattern adding a realization from a homogeneous Poisson process with intensity $\omega = u\rho_\xi$ for $u \in \{0.2, 0.35\}$. The two rightmost columns of Table 6.3.1 show the mean and standard deviation of the estimated values $(\hat{\rho}_\xi, \hat{\alpha})$ over $1,000$ independent replicated datasets, for each combination of the true parameters. For comparison (two middle columns), we fit Gaussian DPPs to the same (contaminated) point patterns, neglecting the presence of the background noise, via MCE using Ripley's $K$-function as implemented in the `spatstat` package. It is clear how ignoring the contamination leads to substantial bias in the estimates for $\rho_\xi$ and $\alpha$, which is mitigated by our approach. However, since our approach requires estimating the parameters of two processes instead of one, our estimators exhibit larger variances.

## 6.4 Statistical inference via shot noise Cox processes

### 6.4.1 Shot noise Cox process and its Palm distributions

SNCPs are a class of general models for clustered point patterns. To define a SNCP, let $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)$ be a locally finite diffuse intensity measure on $\mathbb{X} \times \mathbb{R}_+$. Let $\{\kappa(\cdot;\theta)\}_{\theta\in\mathbb{X}}$ be a family of parametric probability density functions on $\mathbb{X}$, called the kernel of the SNCP, where the parameter space is $\mathbb{X}$ itself. Assuming that $\int_{\mathbb{X}\times\mathbb{R}_+} \gamma\kappa(x;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma) < \infty$ for any $x \in \mathbb{X}$, then $\Phi$ is a SNCP directed by $\nu$ with kernel $\kappa$ if

$$\Phi \,|\, \Lambda \sim \mathrm{PP}\left(\int_{\mathbb{X}\times\mathbb{R}_+} \gamma\kappa(x;\theta)\Lambda(\mathrm{d}\theta\,\mathrm{d}\gamma)\mathrm{d}x\right), \qquad \Lambda \sim \mathrm{PP}(\nu), \tag{6.5}$$

where $\mathrm{PP}(\omega)$ denotes the law of a Poisson point process with intensity measure $\omega$. We write $\Phi \sim \mathrm{SNCP}(\kappa, \nu)$. A SNCP is a *cluster process*: the coloring theorem of Poisson processes entails that $\Phi$ equals in distribution the sum $\sum_{i\geq 1} \Phi_i$, where $\Phi_i \,|\, \Lambda \overset{\text{ind}}{\sim} \mathrm{PP}(\gamma_i\kappa(x;\theta_i)\mathrm{d}x)$, and $\Lambda = \sum_{i\geq 1} \delta_{(\theta_i,\gamma_i)}$. Then, for each point $\tilde{X}_j$ of $\Phi$, it is possible to introduce a latent cluster allocation variable $T_j$ such that $\Phi_i(\{\tilde{X}_j\}) = 1$ iff $T_j = i$ and $\Phi_i(\{\tilde{X}_j\}) = 0$ otherwise, i.e., $T_j = i$ if $\tilde{X}_j$ arose from $\Phi_i$. The number of distinct values across $\boldsymbol{T} := (T_j)_{j\geq 1}$, denoted with $|\boldsymbol{T}|$, represents the number of clusters among the points. Wang et al. (2023) exploit

this latter construction to draw a connection with Bayesian mixtures with finite number of components (Lijoi et al., 2008; De Blasi et al., 2013), also known as mixtures of finite mixtures (Miller and Harrison, 2018), when $\Lambda$ is a finite Poisson process, and the $\gamma_i$'s are independent and gamma-distributed random variables.

The Palm distributions of a SNCP at a single point $x$ was obtain in Møller (2003), but higher-order Palm and reduced Palm distributions for SNCPs are not known in the literature. We fill this gap with the theorem below, which follows from a recursive application of Theorem 6.1. Define $\eta(x_1, \dots, x_p) = \int_{\mathbb{X}\times\mathbb{R}_+} \gamma^p \prod_{j=1}^p \kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)$.

**Theorem 6.2.** *Let $\Phi \sim \mathrm{SNCP}(\kappa, \nu)$ and $\boldsymbol{x} = (X_1, \dots, X_k) \in \mathbb{X}^k$. Then, the reduced Palm version of $\Phi$ at $\boldsymbol{x}$ admits the following representation*

$$\Phi^!_{\boldsymbol{x}} \,|\, \boldsymbol{T} \stackrel{d}{=} \Phi + \sum_{h=1}^{|\boldsymbol{T}|} \Phi_{\zeta_{\boldsymbol{x}_h}}, \qquad \mathsf{P}(\boldsymbol{T} = \boldsymbol{t}) \propto \prod_{h=1}^{|\boldsymbol{t}|} \eta(\boldsymbol{x}_h), \tag{6.6}$$

*where $\boldsymbol{T} := (T_1, \dots, T_k)$ are latent indicators describing a partition of $\boldsymbol{x}$ into $|\boldsymbol{T}|$ clusters and*

$$\Phi_{\zeta_{\boldsymbol{x}_h}} \,|\, \zeta_{\boldsymbol{x}_h} = (\theta_{\boldsymbol{x}_h}, \gamma_{\boldsymbol{x}_h}) \sim \mathrm{PP}(\gamma_{\boldsymbol{x}_h}\kappa(x;\theta_{\boldsymbol{x}_h})\,\mathrm{d}x)$$
$$\zeta_{\boldsymbol{x}_h} \sim f_{\boldsymbol{x}_h}(\mathrm{d}\theta\,\mathrm{d}\gamma) \propto \gamma^{n_h} \prod_{X\in\boldsymbol{x}_h} \kappa(X;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma),$$

*where $\boldsymbol{x}_h = (X_\ell : T_\ell = h)$ and $n_h$ is the cardinality of $\boldsymbol{x}_h$. Finally, the processes $\Phi$ and $\Phi_{\zeta_{\boldsymbol{x}_h}}$, $h = 1, \dots, |\boldsymbol{T}|$, are mutually independent conditionally to $\boldsymbol{T}$.*

See Appendix 6.D.1 for the precise definition of the space where $\boldsymbol{T}$ takes value. Remarkably, from Theorem 6.2, $\Phi^!_{\boldsymbol{x}}$ is not a SNCP itself, disproving a conjecture by Coeurjolly et al. (2017).

### 6.4.2 Maximum likelihood for shot noise Cox processes

Let now $\mathbb{X} = \mathbb{R}^2$ and assume that $\Phi$ has a finite number of points almost surely. Following Daley and Vere-Jones (2003), we define the likelihood of $\Phi$ as its Janossy density seen as a function of the parameters of $\Phi$. Briefly, we recall that for a regular finite point process $\Phi$ with associated Janossy density $j_k : \mathbb{X}^k \to \mathbb{R}_+$, $j_k(x_1, \dots, x_k)\mathrm{d}x_1 \dots \mathrm{d}x_k$ represents the probability that $\Phi$ consists of exactly $k$ points located in infinitesimal neighborhoods of $x_\ell$, $\ell = 1, \dots, k$. Under suitably regularity conditions, the family of Janossy densities $j_k(\cdot)$, $k \geq 0$, characterizes its probability distribution (Proposition 5.3.II, Daley and Vere-Jones, 2003).

The next theorem, derived from Theorem 6.2, gives an explicit expression for the Janossy densities when $\Phi$ is a general finite SNCP. We remark that a sufficient condition for the finiteness of $\Phi$ is $\int_{\mathbb{X}\times\mathbb{R}_+} \gamma\nu(\mathrm{d}\theta\,\mathrm{d}\gamma) < \infty$. For simplicity, we report here only the case when $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma) = \rho(\mathrm{d}\gamma)G_0(\mathrm{d}\theta)$; see Appendix 6.D.3 for the general statement and the proof.

**Theorem 6.3.** *Let $\Phi \sim \mathrm{SNCP}(\kappa, \nu)$ such that $\Phi(\mathbb{X}) < \infty$ almost surely and $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma) = \rho(\mathrm{d}\gamma)G_0(\mathrm{d}\theta)$. Then*

$$j_k(x_1, \dots, x_k) = k!\,\mathsf{P}(\Phi(\mathbb{X}) = k)\,\mathsf{E}\left[\prod_{h=1}^{|\boldsymbol{T}|} \int_{\mathbb{X}} \prod_{\ell:T_\ell=h} \kappa(x_\ell;\theta)G_0(\mathrm{d}\theta)\right], \tag{6.7}$$

| True parameter | | Estimates: median (IQR) | | |
|---|---|---|---|---|
| $\tau$ | Number of points ($c$) | $\hat{\tau}$ | $\hat{\alpha}$ | $\hat{\sigma_0}$ |
| 1 | 100 (0.01) | 1.192 (0.650) | 0.497 (0.055) | 0.912 (0.380) |
| 1 | 200 (0.005) | 1.508 (0.801) | 0.498 (0.049) | 0.932 (0.308) |
| 5 | 100 (0.05) | 5.584 (2.381) | 0.510 (0.140) | 0.951 (0.253) |
| 5 | 200 (0.025) | 6.463 (2.155) | 0.499 (0.071) | 0.955 (0.256) |

Table 6.4.1: Median and interquantile range (IQR, in brackets) of the estimates $(\hat{\tau}, \hat{\alpha}, \hat{\sigma_0})$ of the Thomas-gamma SNCP for different combinations of the true parameters

*where the expectation is taken with respect to the indicators* $\boldsymbol{T} = (T_1, \ldots, T_k)$ *with distribution*

$$\mathsf{P}(\boldsymbol{T} = \boldsymbol{t}) \propto \prod_{h=1}^{|\boldsymbol{t}|} \int_{\mathbb{R}_+} e^{-\gamma} \gamma^{n_h} \rho(\mathrm{d}\gamma). \tag{6.8}$$

Maximum likelihood estimation for $\Phi \sim \text{SNCP}(\kappa, \nu)$ consists in selecting the parameters by maximizing $j_k$ in (6.7)–(6.8). In particular, Theorem 6.3 suggests an expectation-maximization algorithm to perform such a maximization, which is discussed in Appendix 6.D.5.

**Numerical illustration on the Thomas-gamma SNCP.** Let $\kappa(x; \theta)$ be the density of the bivariate Gaussian distribution with mean $\theta \in \mathbb{R}^2$ and covariance $\alpha^2 I$, $\alpha > 0$, where $I$ is the $2 \times 2$ identity matrix. Let $\nu(\mathrm{d}\theta\, \mathrm{d}\gamma) = \rho(\mathrm{d}\gamma) G_0(\mathrm{d}\theta)$, where $\rho(\mathrm{d}\gamma) = \tau \gamma^{-1} e^{-c\gamma} \mathrm{d}\gamma$, for $\tau, c > 0$, is the Lévy intensity of the gamma process, and $G_0$ is the bivariate Gaussian distribution with mean $(0, 0)$ and covariance matrix ${\sigma_0}^2 I$, $\sigma_0 > 0$. We say that $\Phi$ is a *Thomas-gamma SNCP*. Observe that, even if $\Lambda$ has an a.s. infinite number of points, $\Phi$ is a.s. finite. In particular, $\Phi(\mathbb{X})$ is a negative binomial random variable with parameters $(\tau, c/(c+1))$. Moreover, the law of the indicators $\boldsymbol{T}$ corresponds to the celebrated Chinese restaurant process with concentration parameter $\tau$ (Blackwell and MacQueen, 1973).

We explore different choices for the parameters of the Thomas-gamma SNCP. Specifically, we consider $\tau \in \{1, 5\}$ and, for each $\tau$, the parameter $c$ is chosen such that the expected number of points of the SNCP equals $\mathsf{E}(M) \in \{100, 200\}$. Moreover, we set $\alpha = 0.5$ and $\sigma_0 = 1$. For each of the combinations of the parameters, we generate a point pattern imposing that its number of points $M$ coincide with its expected value $\mathsf{E}(M)$. Given the produced point pattern, we compute the maximum likelihood estimates $(\hat{\tau}, \hat{\alpha}, \hat{\sigma_0})$, while $c$ is fixed at its true value, by maximizing the Janossy density in (6.7)–(6.8), which simplifies for the Thomas-gamma SNCP. Table 6.4.1 reports the median and interquantile range of the estimates $(\hat{\tau}, \hat{\alpha}, \hat{\sigma_0})$ over 100 independent replicated datasets, for each combination of the true parameters. Overall, the maximum likelihood estimates perform well, accurately recovering the true parameter values. It is worth noting that, to the best of our knowledge, maximum likelihood estimation has not previously been available in this setting, where estimation methods have traditionally relied on sampling-based procedures.

### 6.4.3 Clustering labels in extended feature models via shot noise Cox processes

In the general setting of *extended feature allocation models*, introduced in Chapter 4, observations $Z_i$'s contain collections of feature labels, which, in the spatial context, are points in the space. That is, each $Z_i$ is a point pattern, which can be think of as representing tree locations. In particular, suppose that all tree locations are collected in the point process $\Phi = \sum_{j\geq 1} \delta_{\tilde{X}_j}$, and each tree $\tilde{X}_j$ is observed in the generic survey $Z_i$ with probability $\tilde{q}_j \in (0,1]$. The use of shot noise Cox processes enables the clustering of the trees based on their positions, while still allowing the estimation of the locations of the unseen trees. This can be naturally obtained by assuming that $\Phi$ is distributed as a shot noise Cox process and, recalling the notation used in Chapter 4, the model writes as

$$Z_i \,|\, \tilde{\mu} \overset{\text{iid}}{\sim} \text{BP}(\tilde{\mu}), \tag{6.9}$$

where $\tilde{\mu}(B) = \int_{\mathbb{X}\times(0,1]} s\mathbb{1}_B(x)\Psi(\mathrm{d}x\,\mathrm{d}s)$, $B \in \mathscr{X}$, and $\Psi = \sum_{j\geq 1} \delta_{(\tilde{X}_j,\tilde{q}_j)}$ is an independently marked shot noise Cox process with ground process $\Phi = \sum_{j\geq 1} \delta_{\tilde{X}_j} \sim \text{SNCP}(\kappa,\nu)$ and mark kernel $H(\cdot \,|\, x)$. The general theorems in Chapter 4 and Theorem 6.2 allow us to derive the marginal distribution of a sample $\boldsymbol{Z} = (Z_i : i = 1,\ldots,n)$ and the posterior distribution of $\tilde{\mu}$, given a sample of size $n$. The first result concerns the marginal distribution of $\boldsymbol{Z}$.

**Proposition 6.1.** *Let $\boldsymbol{Z}$ be a sample from the statistical model (6.9), where $\tilde{\mu}$ is the functional of the independently marked shot noise Cox process $\Psi$ described above. The marginal distribution of the sample $\boldsymbol{Z}$, namely the probability of observing $k$ features labeled $\boldsymbol{x} = (X_1,\ldots,X_k)$ with corresponding vector of frequency counts $\boldsymbol{m} := (m_1,\ldots,m_k)$, equals*

$$\begin{aligned}
&\exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left[1-e^{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\right]\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\}\prod_{\ell=1}^{k}\int_{(0,1]}(1-s)^{n-m_\ell}s^{m_\ell}H(\mathrm{d}s\,|\,X_\ell)\\
&\quad\times\sum_{\boldsymbol{t}}\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}\times\mathbb{R}_+}e^{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\gamma^{n_h}\prod_{\ell:t_\ell=h}\kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma),
\end{aligned}$$

*where $\boldsymbol{t} := (t_1,\ldots,t_k)$ are indicators describing a partition of $\boldsymbol{x}$ into $|\boldsymbol{t}|$ clusters and $n_h$ is the cardinality of $\{\ell : t_\ell = h\}$, for $h = 1,\ldots,|\boldsymbol{t}|$. Moreover, let $\beta_n(x) = \int_{(0,1]}(1-s)^n H(\mathrm{d}s\,|\,x)$.*

The second result presents the posterior distribution of $\tilde{\mu}$ given a sample $\boldsymbol{Z}$ of size $n$.

**Proposition 6.2.** *Let $\boldsymbol{Z}$ be a sample from the statistical model (6.9), where $\tilde{\mu}$ is the functional of the independently marked shot noise Cox process $\Psi$ described above. Let $\boldsymbol{x} = (X_1,\ldots,X_k)$. The posterior distribution of $\tilde{\mu}$ satisfies the distributional equality*

$$\tilde{\mu}\,|\,\boldsymbol{Z} \overset{d}{=} \sum_{\ell=1}^{k} q_\ell\delta_{X_\ell} + \mu', \tag{6.10}$$

*where the weights $q_\ell$'s are independent random variables, further independent of $\mu'$, with marginal density $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell}H(\mathrm{d}s\,|\,X_\ell)$, as $\ell = 1,\ldots,k$. Moreover, $\mu'$ in (6.10) is an independently marked point process with ground process $\Phi'$ and mark kernel*

$H'(\mathrm{d}s \,|\, x) \propto (1-s)^n H(\mathrm{d}s \,|\, x)$. *The ground process* $\Phi'$ *can be represented as*

$$\begin{aligned} \Phi' \,|\, \boldsymbol{T} &\stackrel{d}{=} \Phi^{(0)} + \sum_{h=1}^{|\boldsymbol{T}|} \Phi^{(\boldsymbol{x}_h)}, \\ \mathsf{P}(\boldsymbol{T} = \boldsymbol{t}) &\propto \prod_{h=1}^{|\boldsymbol{t}|} \int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma \int_{\mathbb{X}} (1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x} \gamma^{n_h} \prod_{\ell: t_\ell = h} \kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma); \end{aligned} \tag{6.11}$$

*where* $\boldsymbol{T} := (T_1, \ldots, T_k)$ *are latent indicators describing a partition of* $\boldsymbol{x}$ *into* $|\boldsymbol{T}|$ *clusters and*

*(a)* $\Phi^{(0)} \sim \mathit{SNCP}(\kappa^{(0)}, \nu^{(0)})$, *where*

$$\kappa^{(0)}(x;\theta) = \beta_n(x)\kappa(x;\theta), \quad \nu^{(0)}(\mathrm{d}\theta\,\mathrm{d}\gamma) = e^{-\gamma \int_{\mathbb{X}} (1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x} \nu(\mathrm{d}\theta\,\mathrm{d}\gamma);$$

*(b) for each* $h = 1, \ldots, |\boldsymbol{T}|$, *the process* $\Phi^{(\boldsymbol{x}_h)}$ *is such that*

$$\begin{aligned} \Phi^{(\boldsymbol{x}_h)} \,|\, \zeta_{\boldsymbol{x}_h} = (\theta_{\boldsymbol{x}_h}, \gamma_{\boldsymbol{x}_h}) &\sim \mathit{PP}\left(\gamma_{\boldsymbol{x}_h} \beta_n(x) \kappa(x;\theta_{\boldsymbol{x}_h})\,\mathrm{d}x\right), \\ \zeta_{\boldsymbol{x}_h} \sim f^{(\boldsymbol{x}_h)}(\mathrm{d}\theta\,\mathrm{d}\gamma) &\propto e^{-\gamma \int_{\mathbb{X}} (1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x} \gamma^{n_h} \prod_{\ell: t_\ell = h} \kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma), \end{aligned}$$

*where* $\boldsymbol{x}_h = (X_\ell : T_\ell = h)$ *and* $n_h$ *is the cardinality of* $\boldsymbol{x}_h$. *Finally, the processes* $\Phi^{(0)}$ *and* $\Phi^{(\boldsymbol{x}_h)}$, $h = 1, \ldots, |\boldsymbol{T}|$, *are mutually independent conditionally to* $\boldsymbol{T}$.

The proofs of Propositions 6.1 and 6.2 are reported in Appendix 6.D.4. The posterior characterization of $\tilde{\mu}$ in (6.10) is both expressive and interpretable. In particular, beyond the *old* features, whose occurrence probabilities follow the usual form for an independently marked point process prior, the distinctive behavior concerns the *unseen* features encoded by $\mu'$. Specifically, the peculiarity lies in the distribution of the unseen feature labels $\tilde{X}'_j$ associated with $\mu'$. These labels arise from multiple sources. A first, less interesting component corresponds to labels generated from a SNCP that depends on the observed sample $\boldsymbol{Z}$ only through its size $n$. More interestingly, the observed labels $\boldsymbol{x}$ directly influence the generation of unseen features. Indeed, the joint distribution of $\boldsymbol{T}$ and $\Phi^{(\boldsymbol{x}_h)}$, $h = 1, \ldots, |\boldsymbol{T}|$, in (6.11) induces the following mechanism: the observed labels $\boldsymbol{x}$ tend to form clusters according to their similarities, analogous to a mixture model structure, and, for each cluster, new unseen features are generated with locations concentrated around the observed features assigned to that cluster.

# Appendix

## 6.A Mathematical background on point processes

Here, we provide the formal definition of a point process. A rigorous treatment can be found in Daley and Vere-Jones (2008) or Baccelli et al. (2020). See also Kallenberg (1984) for some further intuition. Following the notation in Baccelli et al. (2020), let $\mathbb{X}$ be a Polish space equipped with the corresponding Borel $\sigma$-algebra $\mathscr{X}$. Let $\mathbb{M}_{\mathbb{X}}$ be the set of counting measures on $(\mathbb{X}, \mathscr{X})$, that is $\nu \in \mathbb{M}_{\mathbb{X}}$ if (i) $\nu$ is a locally finite measure, i.e., $\nu(B) < \infty$ for all relative compact sets $B \in \mathscr{X}$, and (ii) $\nu(B) \in \{0, 1, \ldots\}$ for all relatively compact sets $B \in \mathscr{X}$. Let $\mathcal{M}_{\mathbb{X}}$ be the smallest $\sigma$-algebra which makes the mappings $\nu \mapsto \nu(B)$ measurable, for any $B \in \mathscr{X}$. By definition, a point process $\Phi$ on the space $\mathbb{X}$ is a measurable map from an underlying probability space $(\Omega, \mathscr{A}, \mathsf{P})$ taking values in $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$. Its probability distribution is given by $\mathbf{P}_\Phi = \mathsf{P} \circ \Phi^{-1}$. Notably, any point process $\Phi$ can be represented as $\Phi = \sum_{j \geq 1} \delta_{\tilde{X}_j}$, where $(\tilde{X}_j)_{j \geq 1}$ is a sequence of random variables taking values in $\mathbb{X}$, and $\delta_x$ denotes the Dirac delta mass at $x$. To describe the probabilty distribution of $\Phi$, the most essential summary is its mean measure $M_\Phi$, which is defined as $M_\Phi(B) = \mathsf{E}[\Phi(B)]$ for any $B \in \mathscr{X}$. In general, the $k$-th order factorial moment measure $M_\Phi^{(k)}$ of $\Phi$ is the mean measure of the $k$-th factorial power of $\Phi$, i.e., of the point process $\Phi^{(k)}$ defined as:

$$\Phi^{(k)} := \sum_{(j_1, \ldots, j_k)}^{\neq} \delta_{(X_{j_1}, \ldots, X_{j_k})},$$

where the symbol $\neq$ means that the sum is over all pairwise distinct indexes. Informally, $M_\Phi^{(k)}(\mathrm{d}x_1 \ldots \mathrm{d}x_k)$ corresponds to the probability that $\Phi$ has atoms in infinitesimal neighborhoods of $x_j$, $j = 1, \ldots, k$. See Baccelli et al. (2020) for a detailed account.

To define the *Palm distributions*, we introduce the Campbell measure of $\Phi$ by $\mathscr{C}_\Phi(B \times L) = \mathsf{E}[\Phi_B \mathbb{1}(\Phi \in L)]$, for $B \in \mathscr{X}$ and $L \in \mathcal{M}_{\mathbb{X}}$. Under the assumption that $M_\Phi$ is $\sigma$-finite, then $\mathscr{C}_\Phi$ admits a disintegration

$$\mathscr{C}_\Phi(B \times L) = \int_B \mathbf{P}_\Phi^x(L) M_\Phi(\mathrm{d}x),$$

where $\{\mathbf{P}_\Phi^x\}_{x \in \mathbb{X}}$ is the (a.s.) unique disintegration probability kernel of $\mathscr{C}_\Phi$ with respect to $M_\Phi$, and it is referred to as the Palm kernel of $\Phi$. For fixed $x \in \mathbb{X}$, $\mathbf{P}_\Phi^x$ is a probability distribution over $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$, named the Palm distribution of $\Phi$ at $x$, and thus it can be identified with the law of a point process $\Phi_x \sim \mathbf{P}_\Phi^x$, which is consequently called a *Palm version* of $\Phi$ at $x$. Since $\mathsf{P}(\Phi_x(\{x\}) \geq 1) = 1$, i.e., $x$ is a trivial atom of $\Phi_x$, one can define the point process $\Phi_x^! := \Phi_x - \delta_x$, which is called the *reduced Palm version of* $\Phi$ *at* $x \in \mathbb{X}$, whose associated reduced Palm kernel is indicated by $\mathbf{P}_\Phi^{!x}$.

In a similar fashion, under the assumption that the $k$-th factorial moment measure $M_\Phi^{(k)}$ is $\sigma$-finite, it is possible to construct the family of $k$-th order Palm distributions $\{\mathbf{P}_\Phi^{\boldsymbol{x}}\}_{\boldsymbol{x}\in\mathbb{X}^k}$, and the generic probability measure $\mathbf{P}_\Phi^{\boldsymbol{x}}$ can be interpreted as the distribution of $\Phi$ given that $\boldsymbol{x} = (x_1, \ldots, x_k)$ are atoms of $\Phi$. Again, by removing the trivial atoms $(x_1, \ldots, x_k)$, we obtain the reduced Palm distributions $\mathbf{P}_\Phi^{!\boldsymbol{x}}$, namely the probability law of

$$\Phi_{\boldsymbol{x}}^! := \Phi_{\boldsymbol{x}} - \sum_{j=1}^{k} \delta_{x_j}.$$

## 6.B Proof of the main result and extensions

### 6.B.1 Proof of Theorem 6.1

Let $\Phi = \Phi_1 + \Phi_2$, with $\Phi_1$ independent of $\Phi_2$. By (Baccelli et al., 2020, Proposition 3.2.1), the Palm distribution of a point process $\Phi$ is uniquely characterized by the following relation: for all measurable $f, g : \mathbb{X} \to \mathbb{R}_+$ such that $g$ is $M_\Phi$-integrable

$$\frac{\partial}{\partial t}\mathcal{L}_\Phi(f + tg)|_{t=0} = -\mathsf{E}\left[\Phi(g)e^{-\Phi(f)}\right] = -\int_{\mathbb{X}} g(x)\mathcal{L}_{\Phi_x}(f)M_\Phi(\mathrm{d}x), \tag{6.12}$$

where $\mathcal{L}_\Phi(f)$ denotes the Laplace functional of $\Phi$ evaluated at $f$, i.e., $\mathcal{L}_\Phi(f) = \mathsf{E}\left[e^{-\int_{\mathbb{X}} f(x)\Phi(\mathrm{d}x)}\right]$, and $\Phi(f) := \int_{\mathbb{X}} f(x)\Phi(\mathrm{d}x)$. Therefore, we have

$$\begin{aligned}
\mathsf{E}\left[\Phi(g)e^{-\Phi(f)}\right] &= \mathsf{E}\left[\int_{\mathbb{X}} g(x)\mathrm{e}^{-\int_{\mathbb{X}} f(y)(\Phi_1+\Phi_2)(\mathrm{d}y)}\Phi_1(\mathrm{d}x)\right] \\
&\quad + \mathsf{E}\left[\int_{\mathbb{X}} g(x)\mathrm{e}^{-\int_{\mathbb{X}} f(y)(\Phi_1+\Phi_2)(\mathrm{d}y)}\Phi_2(\mathrm{d}x)\right] \\
&= \mathsf{E}\left[\Phi_1(g)\mathrm{e}^{-\Phi_1(f)}\right]\mathsf{E}[\mathrm{e}^{-\Phi_2(f)}] + \mathsf{E}\left[\Phi_2(g)\mathrm{e}^{-\Phi_2(f)}\right]\mathsf{E}[\mathrm{e}^{-\Phi_1(f)}]
\end{aligned}$$

where the last equality follows from the independence of $\Phi_1$ and $\Phi_2$. Then, applying again (Baccelli et al., 2020, Proposition 3.2.1), we obtain

$$\mathsf{E}\left[\Phi(g)e^{-\Phi(f)}\right] = \left[\int_{\mathbb{X}} g(x)\mathcal{L}_{\Phi_{1x}}(f)M_{\Phi_1}(\mathrm{d}x)\right]\mathcal{L}_{\Phi_2}(f) + \left[\int_{\mathbb{X}} g(x)\mathcal{L}_{\Phi_{2x}}(f)M_{\Phi_2}(\mathrm{d}x)\right]\mathcal{L}_{\Phi_1}(f).$$

Since $M_\Phi(\mathrm{d}x) = M_{\Phi_1}(\mathrm{d}x) + M_{\Phi_2}(\mathrm{d}x)$, and $M_{\Phi_i}$ is absolutely contintuous with respect to $M_\Phi$, we write

$$\mathsf{E}\left[\Phi(g)e^{-\Phi(f)}\right] = \int_{\mathbb{X}} g(x)\left\{\mathcal{L}_{\Phi_2}(f)\mathcal{L}_{\Phi_{1x}}(f)\frac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_\Phi}(x) + \mathcal{L}_{\Phi_1}(f)\mathcal{L}_{\Phi_{2x}}(f)\frac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_\Phi}(x)\right\}M_\Phi(\mathrm{d}x).$$

Therefore, from (6.12), it holds that

$$\mathcal{L}_{\Phi_x}(f) = \mathcal{L}_{\Phi_2}(f)\mathcal{L}_{\Phi_{1x}}(f)\frac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_\Phi}(x) + \mathcal{L}_{\Phi_1}(f)\mathcal{L}_{\Phi_{2x}}(f)\frac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_\Phi}(x),$$

which is an alternative writing for the statement of the theorem.

### 6.B.2 Extensions of Theorem 6.1

In this section, we present and discuss two natural extensions to the main theoretical result of the main body in Theorem 6.1. Firstly, we provide the characterization of the Palm distributions of the superposition of multiple independent point processes. This is shown in the next corollary.

**Corollary 6.1.** *Let $\Phi_1, \dots, \Phi_d$ be independent point processes on $\mathbb{X}$, then*

$$\left(\sum_{i=1}^{d} \Phi_i\right)_x \overset{d}{=} \Phi_{ix} + \sum_{q \neq i} \Phi_q \quad \textit{with probability proportional to } M_{\Phi_i}(\mathrm{d}x), \quad \textit{for } i = 1, \dots, d.$$

*Moreover, the same distributional equivalence holds true for the corresponding reduced Palm versions.*

*Proof.* The proof is analogous to the one of Theorem 6.1. See Appendix 6.B.1. □

Secondly, the Palm distributions of the superposition of two independent processes under multiple conditioning points can be formally addressed by applying the Palm algebra. Under some technical conditions on the factorial moment measures of $\Phi$ (see Section 3.3.2 in Baccelli et al. (2020)), Palm algebra allows to write

$$\left(\Phi^!_{\boldsymbol{x}}\right)^!_{\boldsymbol{y}} \overset{\mathrm{d}}{=} \Phi^!_{(\boldsymbol{x},\boldsymbol{y})}, \quad (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X}^k \times \mathbb{X}^\ell. \tag{6.13}$$

Here, we investigate the special case of the Palm distributions of the superposition of two independent processes under two conditioning points. The following corollary holds true.

**Corollary 6.2.** *Let $\Phi_1$ and $\Phi_2$ be two independent point processes, and $\Phi = \Phi_1 + \Phi_2$ the corresponding superposition. For any $(x, y) \in \mathbb{X}^2$, we have*

$$(\Phi_1 + \Phi_2)^!_{(x,y)} \overset{d}{=} \begin{cases} (\Phi_1)^!_{(x,y)} + \Phi_2 & \textit{with probability equal to } \dfrac{\mathrm{d}M^{(2)}_{\Phi_1}}{\mathrm{d}M^{(2)}_{\Phi}}(x,y) \\ (\Phi_1)^!_x + (\Phi_2)^!_y & \textit{with probability equal to } \dfrac{\mathrm{d}M_{\Phi_1} \times M_{\Phi_2}}{\mathrm{d}M^{(2)}_{\Phi}}(x,y) \\ (\Phi_1)^!_y + (\Phi_2)^!_x & \textit{with probability equal to } \dfrac{\mathrm{d}M_{\Phi_1} \times M_{\Phi_2}}{\mathrm{d}M^{(2)}_{\Phi}}(y,x) \\ \Phi_1 + (\Phi_2)^!_{(x,y)} & \textit{with probability equal to } \dfrac{\mathrm{d}M^{(2)}_{\Phi_2}}{\mathrm{d}M^{(2)}_{\Phi}}(x,y) \end{cases}$$

*where $M_{\Phi_1} \times M_{\Phi_2}$ denotes the product measure.*

*Proof.* By Palm algebra in (6.13), we have that

$$(\Phi_1 + \Phi_2)^!_{(x,y)} = ((\Phi_1 + \Phi_2)^!_x)^!_y.$$

From Theorem 6.1, $(\Phi_1 + \Phi_2)^!_x$ is characterized by the following mixture of point processes,

$$(\Phi_1 + \Phi_2)^!_x \overset{\mathrm{d}}{=} \begin{cases} (\Phi_1)^!_x + \Phi_2 & \text{with probability proportional to } M_{\Phi_1}(\mathrm{d}x) \\ \Phi_1 + (\Phi_2)^!_x & \text{with probability proportional to } M_{\Phi_2}(\mathrm{d}x). \end{cases} \tag{6.14}$$

For notational clarity in the following of the proof, let $\Psi$ denote $(\Phi_1 + \Phi_2)^!_x$, let $\Psi_1$ denote $(\Phi_1)^!_x + \Phi_2$ and let $\Psi_2$ denote $\Phi_1 + (\Phi_2)^!_x$. The goal is then to characterize the distribution of $\Psi^!_y$.

First, observe that the mixture representation in (6.14) can be written in terms of the probability laws of the processes as

$$\mathbf{P}_\Psi = \frac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_\Phi}(x)\mathbf{P}_{\Psi_1} + \frac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_\Phi}(x)\mathbf{P}_{\Psi_2}.$$

Second, the law of $\Psi_y^!$ is characterized by applying Proposition 6.3. Specifically, it holds that

$$\mathbf{P}_{\Psi}^{!y} = \frac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_{\Phi}}(x)\frac{\mathrm{d}M_{\Psi_1}}{\mathrm{d}M_{\Psi}}(y)\mathbf{P}_{\Psi_1}^{!y} + \frac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_{\Phi}}(x)\frac{\mathrm{d}M_{\Psi_2}}{\mathrm{d}M_{\Psi}}(y)\mathbf{P}_{\Psi_2}^{!y}. \tag{6.15}$$

Now, we need to determine the laws of $(\Psi_1)_y^!$ and $(\Psi_2)_y^!$. Still applying Theorem 6.1, we have

$$\mathbf{P}_{\Psi_1}^{!y} = \frac{\mathrm{d}M_{(\Phi_1)_x^!}}{\mathrm{d}M_{\Psi_1}}(y)\mathbf{P}_{(\Phi_1)^!_{(x,y)}+\Phi_2} + \frac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_{\Psi_1}}(y)\mathbf{P}_{(\Phi_1)_x^!+(\Phi_2)_y^!},$$

and

$$\mathbf{P}_{\Psi_2}^{!y} = \frac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_{\Psi_2}}(y)\mathbf{P}_{(\Phi_1)_y^!+(\Phi_2)_x^!} + \frac{\mathrm{d}M_{(\Phi_2)_x^!}}{\mathrm{d}M_{\Psi_2}}(y)\mathbf{P}_{\Phi_1+(\Phi_2)^!_{(x,y)}}.$$

Finally, plugging in the last two expressions in (6.15), we obtain

$$\begin{aligned}\mathbf{P}_{\Psi}^{!y} = {} & \frac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_{\Phi}}(x)\left[\frac{\mathrm{d}M_{(\Phi_1)_x^!}}{\mathrm{d}M_{\Psi}}(y)\mathbf{P}_{(\Phi_1)^!_{(x,y)}+\Phi_2} + \frac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_{\Psi}}(y)\mathbf{P}_{(\Phi_1)_x^!+(\Phi_2)_y^!}\right] \\ & + \frac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_{\Phi}}(x)\left[\frac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_{\Psi}}(y)\mathbf{P}_{(\Phi_1)_y^!+(\Phi_2)_x^!} + \frac{\mathrm{d}M_{(\Phi_2)_x^!}}{\mathrm{d}M_{\Psi}}(y)\mathbf{P}_{\Phi_1+(\Phi_2)^!_{(x,y)}}\right],\end{aligned}$$

which is equivalent to write

$$\Psi_y^! \stackrel{\mathrm{d}}{=} \begin{cases} (\Phi_1)^!_{(x,y)} + \Phi_2 & \text{with probability } \dfrac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_{\Phi}}(x)\dfrac{\mathrm{d}M_{(\Phi_1)_x^!}}{\mathrm{d}M_{\Psi}}(y) \\ (\Phi_1)_x^! + (\Phi_2)_y^! & \text{with probability } \dfrac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_{\Phi}}(x)\dfrac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_{\Psi}}(y) \\ (\Phi_1)_y^! + (\Phi_2)_x^! & \text{with probability } \dfrac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_{\Phi}}(x)\dfrac{\mathrm{d}M_{\Phi_1}}{\mathrm{d}M_{\Psi}}(y) \\ \Phi_1 + (\Phi_2)^!_{(x,y)} & \text{with probability } \dfrac{\mathrm{d}M_{\Phi_2}}{\mathrm{d}M_{\Phi}}(x)\dfrac{\mathrm{d}M_{(\Phi_2)_x^!}}{\mathrm{d}M_{\Psi}}(y) \end{cases}$$

The thesis follows from noticing that $M^{(2)}_{\Phi_1}(\mathrm{d}x\,\mathrm{d}y) = M_{(\Phi_1)_x^!}(\mathrm{d}y)M_{\Phi_1}(\mathrm{d}x)$ and that $M_{\Phi}(\mathrm{d}x)M_{\Psi}(\mathrm{d}y) = M_{\Phi}(\mathrm{d}x)M_{\Phi_x^!}(\mathrm{d}y) = M^{(2)}_{\Phi}(\mathrm{d}x\,\mathrm{d}y)$, thanks to (Baccelli et al., 2020, Proposition 3.3.9). □

## 6.C General results on Palm distributions of point processes

In the present section we prove two general results concenring point process theory that will be useful in the sequel but that could be also of independent interest. First, we characterize the reduced Palm distribution of a mixture of point processes, which is a mixture of the reduced Palm distribution of the individual components.

**Proposition 6.3** (Palm distribution of mixtures of point processes)**.** *Let $\mathbf{P}_{\Phi_i}$, $i = 1, \dots, d$ be probability distributions over $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ and $p_1, \dots, p_d$ such that $p_i \geq 0$, $\sum_{i=1}^d p_i = 1$. Define $\Phi$ such that $\mathbf{P}_{\Phi} = \sum_{i=1}^d p_i \mathbf{P}_{\Phi_i}$ and $\Phi_i$ with law $\mathbf{P}_{\Phi_i}$. Then*

$$\mathbf{P}_{\Phi}^{!x} = \sum_{i=1}^{d} w_i(x)\mathbf{P}_{\Phi_i}^{!x}, \qquad w_i(x) = p_i\frac{\mathrm{d}M_{\Phi_i}}{\mathrm{d}M_{\Phi}}(x).$$

*Proof.* By the CLM formula

$$\mathsf{E}\int f(x, \Phi - \delta_x)\Phi(\mathrm{d}x) = \sum_{i=1}^{d} p_i \int f(x,\varphi)\mathbf{P}_{\Phi_i}^{!x}(\mathrm{d}\varphi)M_{\Phi_i}(\mathrm{d}x).$$

Since $M_{\Phi}$ clearly dominates all the $M_{\Phi_i}$'s we have $M_{\Phi_i}(\mathrm{d}x) = \frac{\mathrm{d}M_{\Phi_i}}{\mathrm{d}M_{\Phi}}(x)M_{\Phi}(\mathrm{d}x)$ and the proof follows. □

In the second result, we provide a general description of the Janossy measures of finite point processes in terms of their Palm distributions and their factorial moment measures. Remind that, for any finite point process $\Phi$, the family of Janossy measures $J_k(\cdot)$, $k \geq 0$, characterizes its probability distribution (Proposition 5.3.II, Daley and Vere-Jones, 2003).

**Theorem 6.4** (Janossy measures of finite point processes)**.** *Let $\Phi$ be a finite point process on $\mathbb{X}$. Then, its Janossy measures satisfy*

$$J_k(\mathrm{d}x_1 \ldots \mathrm{d}x_k) = \mathsf{P}(\Phi^!_{(x_1,\ldots,x_k)}(\mathbb{X} \setminus \{x_1,\ldots,x_k\}) = 0) M^{(k)}_\Phi(\mathrm{d}x_1 \ldots \mathrm{d}x_k).$$

*Moreover, if $\Phi$ is a simple point process, it simplifies as*

$$J_k(\mathrm{d}x_1 \ldots \mathrm{d}x_k) = \mathsf{P}(\Phi^!_{(x_1,\ldots,x_k)}(\mathbb{X}) = 0) M^{(k)}_\Phi(\mathrm{d}x_1 \ldots \mathrm{d}x_k).$$

*Proof.* Let $x_1,\ldots,x_k$ be a generic set of points and denote with $\mathrm{d}x_j$ the ball of radius $\epsilon$ and center $x_j$, as $j = 1,\ldots,k$. Take the radius $\epsilon$ small enough so that the balls are disjoint. Denote with $\tilde{\mathbb{X}} = \mathbb{X} \setminus \cup_{j=1}^k \mathrm{d}x_j$. In this case, the Janossy measure correspond to

$$\begin{aligned} J_k(\mathrm{d}x_1 \ldots \mathrm{d}x_k) &= \mathsf{P}\left(\Phi(\mathrm{d}x_1) = 1, \ldots, \Phi(\mathrm{d}x_k) = 1, \Phi(\tilde{\mathbb{X}}) = 0\right) \\ &= \mathsf{E}\left[\prod_{j=1}^k \mathbb{1}_{\{1\}}(\Phi(\mathrm{d}x_j)) \times \mathbb{1}_{\{0\}}(\Phi(\tilde{\mathbb{X}}))\right]. \end{aligned}$$

Proceed with the following computations:

$$\begin{aligned} J_k(\mathrm{d}x_1 \ldots \mathrm{d}x_k) &= \mathsf{E}\left[\prod_{j=1}^k \mathbb{1}_{\{1\}}(\Phi(\mathrm{d}x_j)) \times \mathbb{1}_{\{0\}}(\Phi(\tilde{\mathbb{X}}))\right] \\ &= \mathsf{E}\left[\int_{\mathbb{X}^k} \mathbb{1}_{\{0\}}(\Phi(\tilde{\mathbb{X}})) \prod_{j=1}^k \delta_{y_j}(\mathrm{d}x_j) \Phi^k(\mathrm{d}y_1 \ldots \mathrm{d}y_k)\right], \end{aligned} \tag{6.16}$$

where $\Phi^k$ is the $k$-th power of $\Phi$. Note that the term $\prod_{j=1}^k \delta_{y_j}(\mathrm{d}x_j)$ entails that the integrand is zero on sets of the type

$$\{\boldsymbol{y} \in \mathbb{X}^k : \; y_j = y_\ell \text{ for } j \neq \ell\}.$$

Therefore we can replace $\Phi^k$ with the $k$-th factorial power $\Phi^{(k)}$. Moreover, observe that for all $\boldsymbol{y}$ such that the integrand is non zero, $\Phi(\tilde{\mathbb{X}}) = (\Phi - \sum_{j=1}^k \delta_{y_j})(\tilde{\mathbb{X}})$. Continuing from (6.16), we get

$$\begin{aligned} J_k(\mathrm{d}x_1 \ldots \mathrm{d}x_k) &= \mathsf{E}\left[\int_{\mathbb{X}^k} \mathbb{1}_{\{0\}}((\Phi - \sum_{j=1}^k \delta_{y_j})(\tilde{\mathbb{X}})) \prod_{j=1}^k \delta_{y_j}(\mathrm{d}x_j) \Phi^{(k)}(\mathrm{d}y_1 \ldots \mathrm{d}y_k)\right] \\ &= \int_{\mathbb{X}^k} \mathsf{E}\left[\mathbb{1}_{\{0\}}(\Phi^!_{\boldsymbol{y}}(\tilde{\mathbb{X}}))\right] \prod_{j=1}^k \delta_{y_j}(\mathrm{d}x_j) M^{(k)}_\Phi(\mathrm{d}y_1 \ldots \mathrm{d}y_k) \\ &= \mathsf{E}\left[\mathbb{1}_{\{0\}}(\Phi^!_{\boldsymbol{x}}(\tilde{\mathbb{X}}))\right] M^{(k)}_\Phi(\mathrm{d}x_1 \ldots \mathrm{d}x_k), \end{aligned} \tag{6.17}$$

where the second equality follows from the CLM formula. □

## 6.D Results and proofs for shot noise Cox process of Section 6.4

### 6.D.1 Auxiliary results for the shot noise Cox process

The first lemma describes the Laplace functional of a SNCP.

**Lemma 6.1.** *Let $\Phi \sim \text{SNCP}(\kappa, \nu)$ and any measurable function $f : \mathbb{X} \to \mathbb{R}$. Then,*

$$\mathcal{L}_\Phi(f) = \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-\exp\left\{-\gamma\int_{\mathbb{X}}(1-\exp\{-f(x)\})\kappa(x;\theta)\mathrm{d}x\right\}\right)\nu(\mathrm{d}\theta\mathrm{d}\gamma)\right\}.$$

*Proof.* See Appendix 6.D.6. □

The second lemma describes the Laplace functional, the mean measure and the reduced Palm version of the processes $\Phi_{\zeta_{\boldsymbol{x}_h}}$ appearing in the characterization of the reduced Palm distributions of the SNCP in Theorem 6.2.

**Lemma 6.2.** *Let $\boldsymbol{x} = (x_1, \ldots, x_k)$ and $\Phi_{\zeta_{\boldsymbol{x}}}$ be such that*

$$\Phi_{\zeta_{\boldsymbol{x}}} \,|\, \zeta_{\boldsymbol{x}} = (\theta_{\boldsymbol{x}}, \gamma_{\boldsymbol{x}}) \sim \mathrm{PP}(\gamma_{\boldsymbol{x}}\kappa(x;\theta_{\boldsymbol{x}})\,\mathrm{d}x), \quad \zeta_{\boldsymbol{x}} = (\theta_{\boldsymbol{x}}, \gamma_{\boldsymbol{x}}) \sim f_{\boldsymbol{x}}(\mathrm{d}\theta\,\mathrm{d}\gamma) \propto \gamma^k \prod_{j=1}^k \kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma).$$

*The following holds true.*

*(i) For any measurable $f : \mathbb{X} \to \mathbb{R}$, the Laplace functional of $\Phi_{\zeta_{\boldsymbol{x}}}$ is equal to*

$$\mathcal{L}_{\Phi_{\zeta_{\boldsymbol{x}}}}(f) = \int_{\mathbb{X}\times\mathbb{R}_+}\exp\left\{-\gamma\int_{\mathbb{X}}(1-\exp\{-f(x)\})\kappa(x;\theta)\mathrm{d}x\right\}\gamma^k\prod_{j=1}^k\kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)/\eta(\boldsymbol{x}),$$

*where $\eta(\boldsymbol{x}) = \int_{\mathbb{X}\times\mathbb{R}_+}\gamma^k\prod_{j=1}^k\kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)$.*

*(ii) The mean measure of $\Phi_{\zeta_{\boldsymbol{x}}}$ equals*

$$M_{\Phi_{\zeta_{\boldsymbol{x}}}}(\mathrm{d}y) = \eta(\boldsymbol{x}, \mathrm{d}y)/\eta(\boldsymbol{x}). \tag{6.18}$$

*(iii) The reduced Palm version of $\Phi_{\zeta_{\boldsymbol{x}}}$ at $y \in \mathbb{X}$ corresponds to*

$$(\Phi_{\zeta_{\boldsymbol{x}}})^!_y \overset{d}{=} \Phi_{\zeta_{(\boldsymbol{x},y)}} \tag{6.19}$$

*Proof.* See Appendix 6.D.6. □

The third lemma describes the $k$-th order factorial moment measure of a SNCP. To state this result in a rigorous way, let us introduce the space $\mathcal{T}_k$ as follows. Let $\boldsymbol{t} = (t_1, \ldots, t_k)$ be a vector of natural numbers such that $\max_j t_j = |\boldsymbol{t}|$, where $|\boldsymbol{t}|$ denotes the number of distinct values in $\boldsymbol{t}$. Consider the partition $\mathcal{P}_k = \{C_1, \ldots, C_{|\boldsymbol{t}|}\}$ of $\{1, \ldots, k\}$ induced by the ties in $\boldsymbol{t}$, such that $j \in C_h$ iff $t_j = h$. Then, the space $\mathcal{T}_k$ contains the equivalence classes of $\boldsymbol{t}$ inducing the same partition $\mathcal{P}_k$. This is the correct space where the vector of latent indicators $\boldsymbol{T}$ takes value in Theorem 6.2 and Theorem 6.3.

**Lemma 6.3.** *Let $\Phi \sim \text{SNCP}(\kappa, \nu)$. The $k$-th order factorial moment measure, denoted with $M_\Phi^{(k)}$, is equal to*

$$M_\Phi^{(k)}(\mathrm{d}\boldsymbol{x}) = \sum_{\boldsymbol{t} \in \mathcal{T}_k} \prod_{h=1}^{|\boldsymbol{t}|} \eta(\boldsymbol{x}_h)\mathrm{d}\boldsymbol{x},$$

*where $\boldsymbol{x}_h = (x_\ell : t_\ell = h)$, for $h = 1, \dots, |\boldsymbol{t}|$.*

*Proof.* See Appendix 6.D.6. □

The last theorem characterizes the joint distribution of the number of points $M$ of $\Phi$, when $\Phi$ is a SNCP, together with the latent partition among its points, determined by the $\Phi_i$'s (see Section 6.4.1). This result turns out to be useful for proving Theorem 6.2.

**Theorem 6.5.** *Consider $\Phi \sim \text{SNCP}(\kappa, \nu)$, for some $\kappa$ and $\int_{\mathbb{X}\times\mathbb{R}_+}(1 - e^{-\gamma})\nu(\mathrm{d}\theta\,\mathrm{d}\gamma) < \infty$. The joint distribution on the number of points and the latent partition induced by $\Phi$ is*

$$\mathsf{P}(M, \{C_1, \dots, C_d\}) = \frac{1}{M!} e^{-\int_{\mathbb{X}\times\mathbb{R}_+}(1-e^{-\gamma})\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)} \prod_{h=1}^{d} \int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma}\gamma^{n_h}\nu(\mathrm{d}\theta\,\mathrm{d}\gamma),$$

*where $n_h = |C_h|$, $M = \sum_{h=1}^d n_h$. Clearly, this law does not depend on the kernel $\kappa$.*

*Proof.* See Appendix 6.D.6. □

### 6.D.2 Proof of Theorem 6.2: the Palm distributions of shot noise Cox processes

To prove Theorem 6.2 we proceed by induction. For $k = 1$ the result is given in Møller (2003) and for $k = 2$ the result can be shown by direct calculation thanks to the Palm algebra. Suppose now the statements hold for a general $k$-tuple $\boldsymbol{x}_k = (x_1, \dots, x_k)$ of distinct points, and let $x_{k+1} \neq x_j$, $j = 1, \dots, k$. By the Palm algebra

$$\Phi^!_{\boldsymbol{x}_{k+1}} = (\Phi^!_{\boldsymbol{x}_k})_{x_{k+1}}.$$

Observe that $\Phi^!_{\boldsymbol{x}_k}$ has a mixture representation (by the induction hypothesis). Let $\boldsymbol{t}_k = (t_1, \dots, t_k) \in \mathcal{T}_k$ be indicators describing a partition of $\boldsymbol{x}_k$ into $|\boldsymbol{t}|$ clusters. For ease of notation, define $\Psi_{\boldsymbol{t}_k} = \Phi + \sum_{h=1}^{|\boldsymbol{t}_k|} \Phi_{\zeta_{\boldsymbol{x}_h}}$ where the $\Phi_{\zeta_{\boldsymbol{x}_h}}$'s are as in the statement of the theorem. Then

$$\mathbf{P}_{\Phi^!_{\boldsymbol{x}_k}} = \sum_{\boldsymbol{t}_k \in \mathcal{T}_k} \mathsf{P}(\boldsymbol{T}_k = \boldsymbol{t}_k)\mathbf{P}_{\Psi_{\boldsymbol{t}_k}}.$$

An application of Proposition 6.3 yields

$$\mathbf{P}_{\Phi^!_{\boldsymbol{x}_{k+1}}} = \sum_{\boldsymbol{t}_k \in \mathcal{T}_k} w_{\boldsymbol{t}_k}(x_{k+1})\mathbf{P}^{!x_{k+1}}_{\Psi_{\boldsymbol{t}_k}},$$

where $\mathbf{P}^{!x_{k+1}}_{\Psi_{\boldsymbol{t}_k}}$ is the law of the process $\left(\Phi + \sum_{h=1}^{|\boldsymbol{t}_k|} \Phi_{\zeta_{\boldsymbol{x}_h}}\right)^!_{x_{k+1}}$ and, by Lemma 6.2,

$$w_{\boldsymbol{t}_k}(x_{k+1}) \propto \mathsf{P}(\boldsymbol{T}_k = \boldsymbol{t}_k)\left(\eta(x_{k+1}) + \sum_{h=1}^{\boldsymbol{t}_k} \frac{\eta(\boldsymbol{x}_h, x_{k+1})}{\eta(\boldsymbol{x}_k)}\right) =: \mathsf{P}(\boldsymbol{T}_k = \boldsymbol{t}_k)\lambda_{\boldsymbol{t}}(x_{k+1}).$$

Moreover

$$\left(\Phi+\sum_{h=1}^{|\boldsymbol{t}_k|}\Phi_{\zeta_{\boldsymbol{x}_h}}\right)^{!}_{x_{k+1}} \stackrel{\mathrm{d}}{=} \begin{cases}\Phi+\Phi_{\zeta_{x_{k+1}}}+\sum_{h=1}^{|\boldsymbol{t}_k|}\Phi_{\zeta_{\boldsymbol{x}_h}} & \text{with prob. } \frac{\eta(x_{k+1})}{\lambda_{\boldsymbol{t}}(x_{k+1})}\\ \Phi+\Phi_{\zeta_{(\boldsymbol{x}_j,x_{k+1})}}+\sum_{h\neq j}|\boldsymbol{t}_k|\Phi_{\zeta_{\boldsymbol{x}_h}} & \text{with prob. } \frac{\eta(\boldsymbol{x}_h,x_{k+1})/\eta(\boldsymbol{x}_h)}{\lambda_{\boldsymbol{t}}(x_{k+1})}\end{cases}$$

Putting things together, we obtain

$$\Phi^{!}_{\boldsymbol{x}_{k+1}} \stackrel{\mathrm{d}}{=} \begin{cases}\Phi+\Phi_{\zeta_{x_{k+1}}}+\sum_{h=1}^{|\boldsymbol{t}_k|}\Phi_{\zeta_{\boldsymbol{x}_h}} & \text{with prob. } w_{\boldsymbol{t}_k}(x_{k+1})\frac{\eta(x_{k+1})}{\lambda_{\boldsymbol{t}}(x_{k+1})}\\ \Phi+\Phi_{\zeta_{(\boldsymbol{x}_j,x_{k+1})}}+\sum_{h\neq j}|\boldsymbol{t}_k|\Phi_{\zeta_{\boldsymbol{x}_h}} & \text{with prob. } w_{\boldsymbol{t}_k}(x_{k+1})\frac{\eta(\boldsymbol{x}_h,x_{k+1})/\eta(\boldsymbol{x}_h)}{\lambda_{\boldsymbol{t}}(x_{k+1})}.\end{cases}$$

The proof follows by observing that $w_{\boldsymbol{t}_k}(x_{k+1})\frac{\eta(x_{k+1})}{\lambda_{\boldsymbol{t}}(x_{k+1})} \propto \prod_{h=1}^{|\boldsymbol{t}_k|}\eta(\boldsymbol{x}_h)\eta(x_{k+1})$ and $w_{\boldsymbol{t}_k}(x_{k+1})$ $\frac{\eta(\boldsymbol{x}_h,x_{k+1})/\eta(\boldsymbol{x}_h)}{\lambda_{\boldsymbol{t}}(x_{k+1})} \propto \prod_{j\neq h}^{|\boldsymbol{t}|}\eta(\boldsymbol{x}_j)\eta(\boldsymbol{x}_h,x_{k+1})$.

### 6.D.3 General statement and proof of Theorem 6.3: the Janossy measures of shot noise Cox processes

We provide the Janossy measures for finite SNCPs, for any arbitrary intensity measure $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)$. Theorem 6.3 corresponds to the special case with the additional assumption $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)=\rho(\mathrm{d}\gamma)G_0(\mathrm{d}\theta)$. This special case is discussed in the following statement as well.

**Theorem 6.6.** *Let $\Phi\sim$ SNCP$(\kappa,\nu)$ such that $\Phi(\mathbb{X})<\infty$ almost surely. Then, its Janossy density equals*

$$j_k(x_1,\ldots,x_k)=\exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}(1-e^{-\gamma})\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\}\sum_{\boldsymbol{t}\in\mathcal{T}_k}\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}\times\mathbb{R}_+}e^{-\gamma}\gamma^{n_h}\prod_{j:t_j=h}\kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma).$$

*Moreover, if $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)=\rho(\mathrm{d}\gamma)G_0(\mathrm{d}\theta)$, where $G_0$ is a probability measure on $\mathbb{X}$, then*

$$j_k(x_1,\ldots,x_k)=k!\,\mathsf{P}(\Phi(\mathbb{X})=k)\,\mathsf{E}\left[\prod_{h=1}^{|\boldsymbol{T}|}\int_{\mathbb{X}}\prod_{j:T_j=h}\kappa(x_j;\theta)G_0(\mathrm{d}\theta)\right],$$

*where the expectation is taken with respect to the indicators $\boldsymbol{T}=(T_1,\ldots,T_k)\in\mathcal{T}_k$ with distribution*

$$\mathsf{P}(\boldsymbol{T}=\boldsymbol{t})\propto\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{R}_+}e^{-\gamma}\gamma^{n_h}\rho(\mathrm{d}\gamma).$$

*Proof.* The proof follows from specializing Theorem 6.4 for a SNCP. In particular, Theorem 6.4 states that

$$J_k(\mathrm{d}x_1\,\ldots\,\mathrm{d}x_k)=\mathsf{P}(\Phi^{!}_{(x_1,\ldots,x_k)}(\mathbb{X})=0)M^{(k)}_{\Phi}(\mathrm{d}x_1\,\ldots\,\mathrm{d}x_k). \tag{6.20}$$

Focusing on the first term of (6.20), and denoting with $\boldsymbol{x}=(x_1,\ldots,x_k)$,

$$\mathsf{P}\left(\Phi^{!}_{\boldsymbol{x}}(\mathbb{X})=0\right)=\mathsf{E}\left[\mathsf{P}\left(\Phi^{!}_{\boldsymbol{x}}(\mathbb{X})=0\,|\,\boldsymbol{T}\right)\right]=\mathsf{E}\left[\mathsf{P}\left(\Phi(\mathbb{X})=0\right)\prod_{h=1}^{|\boldsymbol{T}|}\mathsf{P}\left(\Phi_{\zeta_{\boldsymbol{x}_h}}(\mathbb{X})=0\right)\right], \tag{6.21}$$

where $\boldsymbol{T}$ and the $\Phi_{\zeta_{\boldsymbol{x}_h}}$'s are introduced in Theorem 6.2. It follows that (6.21) writes as

$$\begin{aligned}
\mathsf{P}\left(\Phi^!_{\boldsymbol{x}}(\mathbb{X}) = 0\right) &= \mathsf{E}\left[\mathsf{P}\left(\Phi(\mathbb{X}) = 0 \,|\, \Lambda\right)\right] \mathsf{E}\left[\prod_{h=1}^{|\boldsymbol{T}|} \mathsf{E}\left[\mathsf{P}\left(\Phi_{\zeta_{\boldsymbol{x}_h}}(\mathbb{X}) = 0 \,|\, \zeta_{\boldsymbol{x}_h}\right)\right]\right] \\
&= \mathsf{E}\left[e^{-\int_{\mathbb{X}}\int_{\mathbb{X}\times\mathbb{R}_+}\gamma\kappa(x;\theta)\Lambda(\mathrm{d}\theta\,\mathrm{d}\gamma)\mathrm{d}x}\right] \mathsf{E}\left[\prod_{h=1}^{|\boldsymbol{T}|} \mathsf{E}\left[e^{-\int_{\mathbb{X}}\gamma_{\boldsymbol{x}_h}\kappa(x;\theta_{\boldsymbol{x}_h})\mathrm{d}x}\right]\right] \\
&= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-e^{-\gamma}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\} \mathsf{E}\left[\prod_{h=1}^{|\boldsymbol{T}|}\int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma} f_{\boldsymbol{x}_h}(\mathrm{d}\theta\,\mathrm{d}\gamma)\right] \\
&= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-e^{-\gamma}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\} \sum_{\boldsymbol{t}\in\mathcal{T}_k}\left[\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma} f_{\boldsymbol{x}_h}(\mathrm{d}\theta\,\mathrm{d}\gamma)\right] \mathsf{P}(\boldsymbol{T}=\boldsymbol{t}),
\end{aligned}$$

where $f_{\boldsymbol{x}_h}$ is defined in Theorem 6.2. Finally, plugging this last expression in (6.20) and using Lemma 6.3 for $M^{(k)}_\Phi$, we obtain

$$\begin{aligned}
J_k(\mathrm{d}x_1\,\ldots\,\mathrm{d}x_k) &= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-e^{-\gamma}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\} \\
&\qquad\times \sum_{\boldsymbol{t}\in\mathcal{T}_k}\prod_{h=1}^{|\boldsymbol{t}|}\eta(\boldsymbol{x}_h)\int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma} f_{\boldsymbol{x}_h}(\mathrm{d}\theta\,\mathrm{d}\gamma)\,\mathrm{d}x_1\,\ldots\,\mathrm{d}x_k \\
&= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-e^{-\gamma}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\} \\
&\qquad\times \sum_{\boldsymbol{t}\in\mathcal{T}_k}\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma}\gamma^{n_h}\prod_{j:t_j=h}\kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\,\mathrm{d}x_1\,\ldots\,\mathrm{d}x_k,
\end{aligned}$$

and the thesis is proved.

If $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma) = \rho(\mathrm{d}\gamma)G_0(\mathrm{d}\theta)$, where $G_0$ is a probability measure, then the Janossy density boils down to

$$\begin{aligned}
&j_k(x_1,\ldots,x_k) \\
&= \exp\left\{-\int_{\mathbb{R}_+}(1-e^{-\gamma})\rho(\mathrm{d}\gamma)\right\}\sum_{\boldsymbol{t}\in\mathcal{T}_k}\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}}\prod_{j:t_j=h}\kappa(x_j;\theta)G_0(\mathrm{d}\theta)\int_{\mathbb{R}_+} e^{-\gamma}\gamma^{n_h}\rho(\mathrm{d}\gamma).
\end{aligned} \tag{6.22}$$

Specializing the result in Theorem 6.5 for $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma) = \rho(\mathrm{d}\gamma)G_0(\mathrm{d}\theta)$, we obtain

$$\mathsf{P}(M,\{C_1,\ldots,C_d\}) = \frac{1}{M!}e^{-\int_{\mathbb{R}_+}(1-e^{-\gamma})\rho(\mathrm{d}\gamma)}\prod_{h=1}^{d}\int_{\mathbb{R}_+} e^{-\gamma}\gamma^{n_h}\rho(\mathrm{d}\gamma),$$

and the following holds true:

$$\begin{aligned}\mathsf{P}(M=k) &= \sum_{\{C_1,\dots,C_d\}} \mathsf{P}(M=k,\{C_1,\dots,C_d\}) \\ &= \frac{1}{k!}e^{-\int_{\mathbb{R}_+}(1-e^{-\gamma})\rho(\mathrm{d}\gamma)} \sum_{\{C_1,\dots,C_d\}} \prod_{h=1}^{d}\int_{\mathbb{R}_+} e^{-\gamma}\gamma^{n_h}\rho(\mathrm{d}\gamma) \\ &= \frac{1}{k!}e^{-\int_{\mathbb{R}_+}(1-e^{-\gamma})\rho(\mathrm{d}\gamma)} \sum_{\boldsymbol{t}\in\mathcal{T}_k} \prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{R}_+} e^{-\gamma}\gamma^{n_h}\rho(\mathrm{d}\gamma),\end{aligned}$$

where the last equality is just due to a change in notation. Therefore, (6.22) can be written as

$$\begin{aligned}&j_k(x_1,\dots,x_k) = k!\mathsf{P}(M=k) \\ &\times \sum_{\boldsymbol{t}\in\mathcal{T}_k}\left[\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}}\prod_{j:t_j=h}\kappa(x_i;\theta)G_0(\mathrm{d}\theta)\right]\frac{\exp\left\{-\int_{\mathbb{R}_+}(1-e^{-\gamma})\rho(\mathrm{d}\gamma)\right\}\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{R}_+}e^{-\gamma}\gamma^{n_h}\rho(\mathrm{d}\gamma)}{k!\mathsf{P}(M=k)},\end{aligned}$$

where the fraction term corresponds to $\mathsf{P}(\boldsymbol{T}=\boldsymbol{t})$, where the law of $\boldsymbol{T}$ is defined in the statement. In conclusion, the thesis follows by expressing the last equation as an expected value with respect to $\boldsymbol{T}$. □

### 6.D.4 Proofs of Propositions 6.1 and 6.2

The proofs of these results are based on Theorems 4.1 and 4.2 in Chapter 4 and Theorem 6.2 above.

*Proof of Proposition 6.1.* The marginal distribution of $\boldsymbol{Z}$ is obtained by specializing Theorem 4.1 in Chapter 4, which writes in generality as

$$\int_{(0,1]^k}\mathsf{E}\left\{e^{\int_{\mathbb{X}\times(0,1]}n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(\mathrm{d}z\,\mathrm{d}t)}\right\}\prod_{\ell=1}^{k}s_\ell^{m_\ell}(1-s_\ell)^{n-m_\ell}M_\Psi^{(k)}(\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{s}), \tag{6.23}$$

where $M_\Psi^{(k)}(\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{s})$ is the $k$-th order factorial moment measure of $\Psi$ and, for $\boldsymbol{x}\in\mathbb{X}^k$ and $\boldsymbol{s}\in(0,1]^k$, $\Psi^!_{\boldsymbol{x},\boldsymbol{s}}$ is the reduced Palm version of $\Psi$ at $((x_1,s_1),\dots,(x_k,s_k))$. Since $\Psi$ is an independently marked process with ground process $\Phi$ and mark kernel $H$, the $k$-th order factorial moment measure equals $M_\Psi^{(k)}(\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{s}) = M_\Phi^{(k)}(\mathrm{d}\boldsymbol{x})\prod_{\ell=1}^{k}H(\mathrm{d}s_\ell\,|\,x_\ell)$, where $M_\Phi^{(k)}$ is given by Lemma 6.3 since $\Phi$ is a shot noise Cox process. With respect to the decomposition $M_\Psi^{(k)}(\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{s}) = \rho^{(k)}(\mathrm{d}\boldsymbol{s}\,|\,\boldsymbol{x})\tilde{m}_\xi^{(k)}(\mathrm{d}\boldsymbol{x})$ exploited in Chapter 4, it holds that $\rho^{(k)}(\mathrm{d}\boldsymbol{s}\,|\,\boldsymbol{x}) = \prod_{\ell=1}^{k}H(\mathrm{d}s_\ell\,|\,x_\ell)$. Moreover, from (Baccelli et al., 2020, Proposition 3.2.14), the reduced Palm version $\Psi^!_{x,s}$ is still an independently marked process, with ground process $\Phi^!_x$ and mark kernel $H$, thus it does not depend on $s$. By Palm algebra, we can extend this property by claiming that $\Psi^!_{\boldsymbol{x},\boldsymbol{s}}$ is an independently marked process, with ground process $\Phi^!_{\boldsymbol{x}}$ and mark kernel $H$. Then, letting $\boldsymbol{x}=(X_1,\dots,X_k)$,

$$\begin{aligned}\mathsf{E}\left\{e^{\int_{\mathbb{X}\times(0,1]}n\log(1-t)\Psi^!_{\boldsymbol{x},\boldsymbol{s}}(\mathrm{d}z\,\mathrm{d}t)}\right\} &= \mathcal{L}_{\Psi^!_{\boldsymbol{x},\boldsymbol{s}}}(-n\log(1-t)) \\ &= \mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[-\log\int_{(0,1]}(1-t)^n H(\mathrm{d}t\,|\,x)\right],\end{aligned} \tag{6.24}$$

where the second equality follows from (Baccelli et al., 2020, Proposition 2.2.20). Continuing from (6.24), and denoting by $\beta_n(x) = \int_{(0,1]}(1-s)^n H(\mathrm{d}s \,|\, x)$, we get

$$\begin{aligned}
\mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[-\log\beta_n(x)\right] &= \mathsf{E}\left[e^{\int_{\mathbb{X}}\log\beta_n(x)\Phi^!_{\boldsymbol{x}}(\mathrm{d}x)}\right] = \mathsf{E}\left[\mathsf{E}\left[e^{\int_{\mathbb{X}}\log\beta_n(x)\Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T}\right]\right] \\
&= \mathsf{E}\left[\mathsf{E}\left[e^{\int_{\mathbb{X}}\log\beta_n(x)\Phi(\mathrm{d}x)}\prod_{h=1}^{|\boldsymbol{T}|} e^{\int_{\mathbb{X}}\log\beta_n(x)\Phi_{\zeta_{\boldsymbol{x}_h}}(\mathrm{d}x)} \,|\, \boldsymbol{T}\right]\right] \\
&= \mathsf{E}\left[e^{\int_{\mathbb{X}}\log\beta_n(x)\Phi(\mathrm{d}x)}\right]\mathsf{E}\left[\prod_{h=1}^{|\boldsymbol{T}|}\mathsf{E}\left[e^{\int_{\mathbb{X}}\log\beta_n(x)\Phi_{\zeta^*_{\boldsymbol{x}_h}}(\mathrm{d}x)} \,|\, \boldsymbol{T}\right]\right],
\end{aligned}$$

where $\boldsymbol{T}$ and the $\Phi_{\zeta_{\boldsymbol{x}_h}}$'s are introduced in Theorem 6.2. The third equality follows from the decomposition in Theorem 6.2, and the last equality is due to the independence of the processes, conditionally to $\boldsymbol{T}$. By using Lemma 6.1 and Lemma 6.2, the previous expression becomes

$$\begin{aligned}
&\mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[-\log\beta_n(x)\right] \\
&\quad= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-\exp\left\{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x\right\}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\} \\
&\qquad\times\mathsf{E}\left[\prod_{h=1}^{|\boldsymbol{T}|}\int_{\mathbb{X}\times\mathbb{R}_+}\exp\left\{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x\right\}\gamma^{n_h}\right. \\
&\qquad\qquad\left.\times\prod_{\ell:T_\ell=h}\kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)/\eta(\boldsymbol{x}_h)\right] \\
&\quad= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-\exp\left\{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x\right\}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\} \\
&\qquad\times\frac{\sum_{\boldsymbol{t}\in\mathcal{T}_k}\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}\times\mathbb{R}_+}\exp\left\{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x\right\}\gamma^{n_h}\prod_{\ell:t_\ell=h}\kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)}{\sum_{\boldsymbol{t}\in\mathcal{T}_k}\prod_{h=1}^{|\boldsymbol{t}|}\eta(\boldsymbol{x}_h)},
\end{aligned}$$

where $\boldsymbol{x}_h = (X_\ell : T_\ell = h)$ and $n_h$ is the cardinality of $\boldsymbol{x}_h$. Finally, plugging all terms in (6.23), the marginal distribution specializes as

$$\begin{aligned}
&\exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-\exp\left\{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x\right\}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\} \\
&\times\prod_{\ell=1}^{k}\int_{(0,1]}s^{m_\ell}(1-s)^{n-m_\ell}H(\mathrm{d}s\,|\,X_\ell) \\
&\times\sum_{\boldsymbol{t}\in\mathcal{T}_k}\prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}\times\mathbb{R}_+}\exp\left\{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x\right\}\gamma^{n_h}\prod_{\ell:t_\ell=h}\kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma),
\end{aligned}$$

which corresponds to the thesis.

*Proof of Proposition 6.2.* The posterior distribution of $\tilde{\mu}$, given a sample $\boldsymbol{Z}$ displaying $k$ features with labels $\boldsymbol{x} = (X_1, \ldots, X_k)$ and corresponding vector of frequency counts $\boldsymbol{m} := (m_1, \ldots, m_k)$, is obtained by specializing Theorem 4.2 in Chapter 4, which provides

the following distributional equality

$$\tilde{\mu} \,|\, \boldsymbol{Z} \stackrel{\mathrm{d}}{=} \sum_{\ell=1}^{k} q_\ell \delta_{X_\ell} + \mu',$$

where $\boldsymbol{q} := (q_1, \ldots, q_k)$ is a vector of positive random variables with joint distribution

$$f_{\boldsymbol{q}}(\mathrm{d}\boldsymbol{s}) \propto \mathsf{E}\left\{ e^{\int_{\mathbb{X}\times(0,1]} n \log(1-t) \Psi^!_{\boldsymbol{x},\boldsymbol{s}}(\mathrm{d}z\,\mathrm{d}t)} \right\} \prod_{\ell=1}^{k} s_\ell^{m_\ell} (1-s_\ell)^{n-m_\ell} \rho^{(k)}(\mathrm{d}\boldsymbol{s} \,|\, \boldsymbol{x}).$$

As previously observed, the reduced Palm version $\Psi^!_{\boldsymbol{x},\boldsymbol{s}}$ does not depend on $\boldsymbol{s}$, thus the $q_\ell$'s turn out to be independent random variables with marginal law $f_{q_\ell}(\mathrm{d}s) \propto s^{m_\ell}(1-s)^{n-m_\ell} H(\mathrm{d}s \,|\, X_\ell)$, as $\ell = 1, \ldots, k$. Conditionally to $\boldsymbol{q}$, from Chapter 4, $\mu'$ can be represented as $\mu' = \sum_{j=1}^{M'} q'_j \delta_{\tilde{X}'_j}$, where $\Psi' = \sum_{j=1}^{M'} \delta_{(\tilde{X}'_j, q'_j)}$ is characterized by the Laplace functional

$$\mathcal{L}_{\Psi' \,|\, \boldsymbol{q}}(g) = \frac{\mathcal{L}_{\Psi^!_{\boldsymbol{x},\boldsymbol{q}}}(g(x,s) - n\log(1-s))}{\mathcal{L}_{\Psi^!_{\boldsymbol{x},\boldsymbol{q}}}(-n\log(1-s))}, \tag{6.25}$$

for any measurable function $g : \mathbb{X} \times (0,1] \to \mathbb{R}$. Since $\Psi^!_{\boldsymbol{x},\boldsymbol{q}}$ is an independently marked process with ground process $\Phi^!_{\boldsymbol{x}}$ and mark kernel $H$, it does not depend on $\boldsymbol{q}$ and (6.25) writes as

$$\mathcal{L}_{\Psi'}(g) = \frac{\mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[ -\log \int_{(0,1]} e^{-g(x,s)} (1-s)^n H(\mathrm{d}s \,|\, x) \right]}{\mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[ -\log \int_{(0,1]} (1-s)^n H(\mathrm{d}s \,|\, x) \right]}.$$

By (Baccelli et al., 2020, Proposition 2.2.20), $\Psi'$ is an independently marked point process with ground process $\Phi'$ (characterized next) and mark kernel $H'(\mathrm{d}s \,|\, x) \propto (1-s)^n H(\mathrm{d}s \,|\, x)$. Specifically, for any measurable function $f : \mathbb{X} \to \mathbb{R}$, $\Phi'$ has Laplace functional given by

$$\mathcal{L}_{\Phi'}(f) = \frac{\mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[ f(x) - \log \int_{(0,1]} (1-s)^n H(\mathrm{d}s \,|\, x) \right]}{\mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[ -\log \int_{(0,1]} (1-s)^n H(\mathrm{d}s \,|\, x) \right]} = \frac{\mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[ f(x) - \log \beta_n(x) \right]}{\mathcal{L}_{\Phi^!_{\boldsymbol{x}}}\left[ -\log \beta_n(x) \right]}. \tag{6.26}$$

From (6.26), simple algebra leads to

$$\begin{aligned}
\mathcal{L}_{\Phi'}(f) &= \frac{\mathsf{E}\left[ e^{-\int_{\mathbb{X}} f(x) - \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \right]}{\mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \right]} = \frac{\mathsf{E}\left[ \mathsf{E}\left[ e^{-\int_{\mathbb{X}} f(x) - \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right] \right]}{\mathsf{E}\left[ \mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right] \right]} \\
&= \mathsf{E}\left\{ \frac{\mathsf{E}\left[ e^{-\int_{\mathbb{X}} f(x) - \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right]}{\mathsf{E}\left[ \mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right] \right]} \right\} \\
&= \mathsf{E}\left\{ \frac{\mathsf{E}\left[ e^{-\int_{\mathbb{X}} f(x) - \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right]}{\mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right]} \frac{\mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right]}{\mathsf{E}\left[ \mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right] \right]} \right\} \\
&= \sum_{\boldsymbol{t} \in \mathcal{T}_k} \frac{\mathsf{E}\left[ e^{-\int_{\mathbb{X}} f(x) - \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t} \right]}{\mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t} \right]} \frac{\mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t} \right]}{\mathsf{E}\left[ \mathsf{E}\left[ e^{\int_{\mathbb{X}} \log \beta_n(x) \Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} \right] \right]} \cdot \mathsf{P}(\boldsymbol{T} = \boldsymbol{t}).
\end{aligned}$$

Let $\boldsymbol{T}$ be the allocation variables defined in Theorem 6.2, then introduce the allocation variables $\boldsymbol{T}^*$ with distribution

$$\begin{aligned}\mathsf{P}(\boldsymbol{T}^* = \boldsymbol{t}) &= \frac{\mathsf{E}\left[e^{\int_{\mathbb{X}} \log \beta_n(x)\Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t}\right]}{\mathsf{E}\left[\mathsf{E}\left[e^{\int_{\mathbb{X}} \log \beta_n(x)\Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T}\right]\right]} \cdot \mathsf{P}(\boldsymbol{T} = \boldsymbol{t}) \\ &\propto \prod_{h=1}^{|\boldsymbol{t}|} \int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma \int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x} \gamma^{n_h} \prod_{\ell:t_\ell=h} \kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma).\end{aligned}$$

The previous computation follows from the same steps performed in (6.D.4).

Then, by (Kallenberg, 2021, Theorem 3.4), the following holds true:

$$\mathsf{E}\left[e^{-\int_{\mathbb{X}} f(x)\Phi'(\mathrm{d}x)} \,|\, \boldsymbol{T}^* = \boldsymbol{t}\right] = \frac{\mathsf{E}\left[e^{-\int_{\mathbb{X}} f(x) - \log\beta_n(x)\Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t}\right]}{\mathsf{E}\left[e^{\int_{\mathbb{X}} \log \beta_n(x)\Phi^!_{\boldsymbol{x}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t}\right]}.$$

Because of the distributional decomposition and the independence between all the involved point processes, provided in Theorem 6.2, the previous Laplace functional writes as

$$\begin{aligned}\mathsf{E}\left[e^{-\int_{\mathbb{X}} f(x)\Phi'(\mathrm{d}x)} \,|\, \boldsymbol{T}^* = \boldsymbol{t}\right] &= \frac{\mathsf{E}\left[e^{-\int_{\mathbb{X}} f(x) - \log\beta_n(x)\Phi(\mathrm{d}x)}\right]}{\mathsf{E}\left[e^{\int_{\mathbb{X}} \log \beta_n(x)\Phi(\mathrm{d}x)}\right]} \\ &\quad \times \prod_{h=1}^{|\boldsymbol{t}|} \frac{\mathsf{E}\left[e^{-\int_{\mathbb{X}} f(x) - \log\beta_n(x)\Phi_{\zeta_{\boldsymbol{x}_h}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t}\right]}{\mathsf{E}\left[e^{\int_{\mathbb{X}} \log \beta_n(x)\Phi_{\zeta_{\boldsymbol{x}_h}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t}\right]}.\end{aligned} \tag{6.27}$$

Lemma 6.1 allows to handle both numerator and denominator of the first ratio of the right-hand side of (6.27), getting

$$\begin{aligned}\frac{\mathsf{E}\left[e^{-\int_{\mathbb{X}} f(x) - \log\beta_n(x)\Phi(\mathrm{d}x)}\right]}{\mathsf{E}\left[e^{\int_{\mathbb{X}} \log \beta_n(x)\Phi(\mathrm{d}x)}\right]} &= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+} \left(1 - e^{-\gamma\int_{\mathbb{X}}(1-e^{-f(x)}\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right. \\ &\quad \left. + \int_{\mathbb{X}\times\mathbb{R}_+} \left(1 - e^{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\} \\ &= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\right. \\ &\quad \left. \times \left[1 - e^{-\gamma\int_{\mathbb{X}}(1-e^{-f(x)})\beta_n(x)\kappa(x;\theta)\mathrm{d}x}\right]\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\},\end{aligned}$$

which is the Laplace functional of a shot noise Cox process, denoted with $\Phi^{(0)}$.

Similarly, using Lemma 6.2 for computing both numerator and denominator of the generic ratio involving $\Phi_{\zeta_{\boldsymbol{x}_h}}$ in (6.27), it is easy to show that

$$\begin{aligned}&\frac{\mathsf{E}\left[e^{-\int_{\mathbb{X}} f(x) - \log\beta_n(x)\Phi_{\zeta_{\boldsymbol{x}_h}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t}\right]}{\mathsf{E}\left[e^{\int_{\mathbb{X}} \log \beta_n(x)\Phi_{\zeta_{\boldsymbol{x}_h}}(\mathrm{d}x)} \,|\, \boldsymbol{T} = \boldsymbol{t}\right]} \\ &= \frac{\int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma\int_{\mathbb{X}}(1-e^{-f(x)}\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\gamma^{n_h}\prod_{\ell:t_\ell=h}\kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)}{\int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\gamma^{n_h}\prod_{\ell:t_\ell=h}\kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)} \\ &= \int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma\int_{\mathbb{X}}(1-e^{-f(x)})\beta_n(x)\kappa(x;\theta)\mathrm{d}x} e^{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\gamma^{n_h}\prod_{\ell:t_\ell=h}\kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)/\tilde{\eta}(\boldsymbol{x}_h),\end{aligned}$$

where $\tilde{\eta}(\boldsymbol{x}_h) = \int_{\mathbb{X}\times\mathbb{R}_+} e^{-\gamma\int_{\mathbb{X}}(1-\beta_n(x))\kappa(x;\theta)\mathrm{d}x}\gamma^{n_h}\prod_{\ell:t_\ell=h}\kappa(X_\ell;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)$. We observe that the last expression corresponds to the Laplace functional of the point process $\Phi^{(\boldsymbol{x}_h)}$ defined in the statement. Conditionally to $\boldsymbol{T}^*$, independence among $\Phi^{(0)}$ and the $\Phi^{(\boldsymbol{x}^*_h)}$'s stems from the factorization of the Laplace functional of $\Phi'$ in (6.27). In the statement of the result, the allocation variables $\boldsymbol{T}^*$ are relabeled as $\boldsymbol{T}$, but note that they follow a different distribution from $\boldsymbol{T}$ in Theorem 6.2.

### 6.D.5 Expectation-maximization algorithm for maximum likelihood estimation

Let $\Phi \sim \text{SNCP}(\kappa, \nu)$ and denote by $\psi$ the set of parameters of $\Phi$, which enters the definition of $\kappa$ and $\nu$. Define the likelihood $L(\psi;\boldsymbol{x}) = j_k(x_1,\dots,x_k)$ as in Theorem 6.3, for $\boldsymbol{x} = (x_1,\dots,x_k)$. The peculiar structure of the likelihood naturally motivates the use of an EM algorithm for its maximization, which we develop in the following. Let us introduce the following shorthand notation

$$p(\boldsymbol{x}\,|\,\boldsymbol{t},\psi) = \prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{X}}\prod_{j:t_j=h}\kappa(x_j;\theta)G_0(\mathrm{d}\theta), \quad p(\boldsymbol{t}\,|\,\psi) \propto \prod_{h=1}^{|\boldsymbol{t}|}\int_{\mathbb{R}_+} e^{-\gamma}\gamma^{n_h}\rho(\mathrm{d}\gamma), \qquad c(\psi) = k!\,\mathsf{P}(\Phi(\mathbb{X})=k), \tag{6.28}$$

where $\boldsymbol{t} = (t_1,\dots,t_k) \in \mathcal{T}_k$. Then

$$L(\psi;\boldsymbol{x}) = c(\psi)p(\boldsymbol{x}\,|\,\psi) = c(\psi)\sum_{\boldsymbol{t}\in\mathcal{T}_k} p(\boldsymbol{x},\boldsymbol{t}\,|\,\psi) = c(\psi)\sum_{\boldsymbol{t}\in\mathcal{T}_k} p(\boldsymbol{x}\,|\,\boldsymbol{t},\psi)p(\boldsymbol{t}\,|\,\psi). \tag{6.29}$$

Following Haugh (2015), let

$$Q(\psi;\psi_{old}) := \mathsf{E}\left[\ln\left\{c(\psi)p(\boldsymbol{x},\boldsymbol{T}\,|\,\psi)\right\}\,|\,\boldsymbol{x},\psi_{old}\right]. \tag{6.30}$$

The EM algorithm alternates computing $Q(\psi;\psi_{old})$ for the current values of parameters $\psi_{old}$ (*E-step*) and finding a new value $\psi_{new}$ maximizing $Q(\psi;\psi_{old})$ (*M-step*). See Haugh (2015) for further details and justifications. Since the expectation in (6.30) is not available analytically, we employ a stochastic approximation of the *E-step* as in Wei and Tanner (1990) and approximate

$$Q(\psi;\psi_{old}) \approx \frac{1}{U}\sum_{u=1}^{U}\ln\left\{c(\psi)p(\boldsymbol{x},\boldsymbol{T}^{(u)}\,|\,\psi)\right\}, \qquad \boldsymbol{T}^{(1)},\dots,\boldsymbol{T}^{(U)} \overset{\text{iid}}{\sim} p(\cdot\,|\,\boldsymbol{x},\psi_{old}),$$

In our examples, we find that $U = 1$ gives satisfactory results. To sample $\boldsymbol{T}$ from $p(\cdot\,|\,\boldsymbol{x},\psi_{old})$, we employ a standard Gibbs sampler developed for Bayesian nonparametric mixture models, see Algorithm 3 in Neal (2000). The Gibbs sampler starts from an initial configuration for $\boldsymbol{T}$ and updates it through a number of iterations; at each iteration, $\boldsymbol{T}$ is updated by sequentially updating each of its elements $T_j$, as $j = 1,\dots,k$. In particular, for $\ell = 1,\dots,k$, the update of $T_\ell$ given the current values of the remaining latent indicators $\boldsymbol{T}_{-\ell} = (T_j : j \neq \ell)$ follows from its full-conditional distribution. Let $d$ be the number of clusters in the partition induced by $\boldsymbol{T}_{-\ell}$, with cluster cardinalities $n_h$ and associated distinct label $t^*_h$, as $h = 1,\dots,d$. The update for $T_\ell$, denoted with $T^{new}_\ell$, follows

$$p(T^{new}_\ell\,|\,\boldsymbol{T}_{-\ell},\boldsymbol{x}) \propto \sum_{h=1}^{d} p(\boldsymbol{x}\,|\,\boldsymbol{t}'_h)p(\boldsymbol{t}'_h)\delta_{t^*_h} + p(\boldsymbol{x}\,|\,\boldsymbol{t}'_{new})p(\boldsymbol{t}'_{new})\delta_{\min\{i:i\neq t^*_h, h=1,\dots,d\}},$$

where $\boldsymbol{t}'_h$ is obtained by inserting the value $t^*_h$ in the $\ell$-th position of the vector $\boldsymbol{T}_{-\ell}$, shifting subsequent elements one place to the right. Similarly, $\boldsymbol{t}'_{new}$ is obtained by inserting the value $\min\{i : i \neq t^*_h, h = 1, \dots, d\}$ in the $\ell$-th position of the vector $\boldsymbol{T}_{-\ell}$. Note that the number of clusters in the partition induced by $\boldsymbol{t}'_h$ is $d$, for any $h$, while the number of clusters in the partition induced by $\boldsymbol{t}'_{new}$ is $d+1$. The probability density functions $p(\boldsymbol{x} \,|\, \boldsymbol{t})$ and $p(\boldsymbol{t})$ are described in (6.28). In our experiments, we run the Gibbs sampler for 20 iterations and keep the last partition visited as a draw from $p(\cdot \,|\, \boldsymbol{x}, \psi_{old})$. The *M-step* is carried out by a simple gradient descent method, obtaining the gradients of $Q$ via automatic differentiation.

### 6.D.6 Proofs of the auxiliary results in Appendix 6.D.1

#### Proof of Lemma 6.1

The proof is straightforwardly obtained by exploiting the decomposition in (6.5):

$$\begin{aligned}
\mathcal{L}_\Phi(f) &= \mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\Phi(\mathrm{d}x)\right\}\right] = \mathsf{E}\left[\mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\Phi(\mathrm{d}x)\right\} \,|\, \Lambda\right]\right] \\
&= \mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}}(1-e^{-f(x)})\sum_{j\geq 1}\gamma_j\kappa(x;\theta_j)\mathrm{d}x\right\}\right] \\
&= \mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\gamma\int_{\mathbb{X}}(1-e^{-f(x)})\kappa(x;\theta)\mathrm{d}x\,\Lambda(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\}\right] \\
&= \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}_+}\left(1-\exp\left\{-\gamma\int_{\mathbb{X}}(1-\exp\{-f(x)\})\kappa(x;\theta)\mathrm{d}x\right\}\right)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\right\}.
\end{aligned}$$

The third equality follows from the fact that $\Phi \,|\, \Lambda$ is a Poisson process, while the last step holds because $\Lambda$ is Poisson process.

#### Proof of Lemma 6.2

*Point (i).* For the Laplace functional of $\Phi_{\zeta_{\boldsymbol{x}}}$, with $\boldsymbol{x} = (x_1, \dots, x_k)$, the proof follows from direct computation. Let $f : \mathbb{X} \to \mathbb{R}_+$ be any measurable function, then

$$\begin{aligned}
\mathcal{L}_{\Phi_{\zeta_{\boldsymbol{x}}}}(f) &= \mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\Phi_{\zeta_{\boldsymbol{x}}}(\mathrm{d}x)\right\}\right] = \mathsf{E}\left[\mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\Phi_{\zeta_{\boldsymbol{x}}}(\mathrm{d}x)\right\} \,|\, \zeta_{\boldsymbol{x}}\right]\right] \\
&= \mathsf{E}\left[\exp\left\{-\int_{\mathbb{X}}(1-e^{-f(x)})\gamma_{\boldsymbol{x}}\kappa(x;\theta_{\boldsymbol{x}})\mathrm{d}x\right\}\right] \\
&= \int_{\mathbb{X}\times\mathbb{R}_+}\exp\left\{-\gamma\int_{\mathbb{X}}(1-\exp\{-f(x)\})\kappa(x;\theta)\mathrm{d}x\right\}\gamma^k\prod_{j=1}^{k}\kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)/\eta(\boldsymbol{x}).
\end{aligned}$$

The third equality follows from the fact that $\Phi_{\zeta_{\boldsymbol{x}}} \,|\, \zeta_{\boldsymbol{x}}$ is distributed as Poisson process.

*Point (ii).* Still by direct computation, the mean measure of $\Phi_{\zeta_{\boldsymbol{x}}}$ writes as

$$\begin{aligned} M_{\Phi_{\zeta_{\boldsymbol{x}}}}(\mathrm{d}y) &= \mathsf{E}\left[\mathsf{E}\left[\Phi_{\zeta_{\boldsymbol{x}}}(\mathrm{d}y) \,|\, \zeta_{\boldsymbol{x}}\right]\right] = \mathsf{E}\left[\gamma_{\boldsymbol{x}}\kappa(y;\theta_{\boldsymbol{x}})\mathrm{d}y\right] = \\ &= \int_{\mathbb{X}\times\mathbb{R}_+} \gamma\kappa(y;\theta)\mathrm{d}y\, \gamma^k \prod_{j=1}^{k} \kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)/\eta(\boldsymbol{x}) \\ &= \int_{\mathbb{X}\times\mathbb{R}_+} \gamma^{k+1}\kappa(y;\theta) \prod_{j=1}^{k} \kappa(x_j;\theta)\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)\mathrm{d}y/\eta(\boldsymbol{x}) \\ &= \eta(\boldsymbol{x},y)\mathrm{d}y/\eta(\boldsymbol{x}) \end{aligned}$$

and the thesis is proved.

*Point (iii).* By (Baccelli et al., 2020, Proposition 3.2.5), if $\Psi$ is a Cox process directed by the random measure $\mu$, then the reduced Palm version $\Psi_y^!$ is a Cox process directed by $\mu_y$ (the Palm version of $\mu$ at $y$). Now, observe that $\Phi_{\zeta_{\boldsymbol{x}}}$ is a Cox process directed by the random measure $\mu(\mathrm{d}x) = \gamma_{\boldsymbol{x}}\kappa(x;\theta_{\boldsymbol{x}})\,\mathrm{d}x$, with $(\theta_{\boldsymbol{x}},\gamma_{\boldsymbol{x}}) \sim f_{\boldsymbol{x}}(\mathrm{d}\theta\,\mathrm{d}\gamma)$ and $f_{\boldsymbol{x}}$ is defined in the statement of the present lemma. Consequently, the reduced Palm version $(\Phi_{\zeta_{\boldsymbol{x}}})_y^!$ is a Cox process directed by $\mu_y$. We are then left with determining the distribution of $\mu_y$, which we derive by computing its Laplace functional $\mathcal{L}_{\mu_y}(f)$, for any measurable function $f : \mathbb{X} \to \mathbb{R}_+$. To this end, rely on (Baccelli et al., 2020, Proposition 3.2.1), which states that, for any measurable functions $f, g : \mathbb{X} \to \mathbb{R}_+$, the Palm distribution of a random measure $\mu$ satisfies

$$\frac{\partial}{\partial t}\mathcal{L}_{\mu}(f+tg)\big|_{t=0} = -\mathsf{E}\left[\mu(g)e^{-\mu(f)}\right] = -\int_{\mathbb{X}} g(y)\mathcal{L}_{\mu_y}(f)M_{\mu}(\mathrm{d}y), \tag{6.31}$$

where $\mu(f) := \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x)$. Therefore, since $\mu(\mathrm{d}x) = \gamma_{\boldsymbol{x}}\kappa(x;\theta_{\boldsymbol{x}})\,\mathrm{d}x$, the term $\mathsf{E}\left[\mu(g)e^{-\mu(f)}\right]$ writes as

$$\begin{aligned} \mathsf{E}\left[\mu(g)e^{-\mu(f)}\right] &= \mathsf{E}\left[\int_{\mathbb{X}} g(y)\gamma_{\boldsymbol{x}}\kappa(y;\theta_{\boldsymbol{x}})\mathrm{d}y \cdot \exp\left\{-\int_{\mathbb{X}} f(z)\gamma_{\boldsymbol{x}}\kappa(z;\theta_{\boldsymbol{x}})\mathrm{d}z\right\}\right] \\ &= \int_{\mathbb{X}\times\mathbb{R}_+}\int_{\mathbb{X}} g(y)\gamma\kappa(y;\theta)\mathrm{d}y \cdot \exp\left\{-\int_{\mathbb{X}} f(z)\gamma\kappa(z;\theta)\mathrm{d}z\right\} \cdot f_{\boldsymbol{x}}(\mathrm{d}\theta\,\mathrm{d}\gamma) \\ &= \int_{\mathbb{X}} g(y)\int_{\mathbb{X}\times\mathbb{R}_+} \exp\left\{-\int_{\mathbb{X}} f(z)\gamma\kappa(z;\theta)\mathrm{d}z\right\}\gamma\kappa(y;\theta)f_{\boldsymbol{x}}(\mathrm{d}\theta\,\mathrm{d}\gamma)\cdot\mathrm{d}y. \end{aligned} \tag{6.32}$$

Observing that $M_{\mu}(\mathrm{d}y) = \mathsf{E}[\gamma_{\boldsymbol{x}}\kappa(y;\theta_{\boldsymbol{x}})\mathrm{d}y] = \int_{\mathbb{X}\times\mathbb{R}_+}\gamma\kappa(y;\theta)f_{\boldsymbol{x}}(\mathrm{d}\theta\,\mathrm{d}\gamma)\,\mathrm{d}y$, (6.32) can be expressed as

$$\mathsf{E}\left[\mu(g)e^{-\mu(f)}\right] = \int_{\mathbb{X}} g(y)\int_{\mathbb{X}\times\mathbb{R}_+} \exp\left\{-\int_{\mathbb{X}} f(z)\gamma\kappa(z;\theta)\mathrm{d}z\right\}\frac{\gamma\kappa(y;\theta)f_{\boldsymbol{x}}(\mathrm{d}\theta\,\mathrm{d}\gamma)}{\int_{\mathbb{X}\times\mathbb{R}_+}\gamma'\kappa(y;\theta')f_{\boldsymbol{x}}(\mathrm{d}\theta'\,\mathrm{d}\gamma')}\cdot M_{\mu}(\mathrm{d}y)$$

By identification in (6.31), we see that

$$\mathcal{L}_{\mu_y}(f) = \int_{\mathbb{X}\times\mathbb{R}_+} \exp\left\{-\int_{\mathbb{X}} f(z)\gamma\kappa(z;\theta)\mathrm{d}z\right\}\frac{\gamma\kappa(y;\theta)f_{\boldsymbol{x}}(\mathrm{d}\theta\,\mathrm{d}\gamma)}{\int_{\mathbb{X}\times\mathbb{R}_+}\gamma'\kappa(y;\theta')f_{\boldsymbol{x}}(\mathrm{d}\theta'\,\mathrm{d}\gamma')},$$

which corresponds to say that $\mu_y(\mathrm{d}x) = \gamma^*\kappa(x;\theta^*)\mathrm{d}x$, with

$$(\theta^*,\gamma^*) \sim f_*(\mathrm{d}\theta\,\mathrm{d}\gamma) \propto \gamma\kappa(y;\theta)f_{\boldsymbol{x}}(\mathrm{d}\theta\,\mathrm{d}\gamma) \propto f_{(\boldsymbol{x},y)}(\mathrm{d}\theta\,\mathrm{d}\gamma).$$

Summing up, we have found that

$$(\Phi_{\zeta_{\boldsymbol{x}}})^!_y \mid (\theta^*, \gamma^*) \sim \mathrm{PP}(\gamma^* \kappa(x; \theta^*)\, \mathrm{d}x), \qquad (\theta^*, \gamma^*) \sim f_{(\boldsymbol{x},y)}(\mathrm{d}\theta\, \mathrm{d}\gamma),$$

which is equivalent to say $(\Phi_{\zeta_{\boldsymbol{x}}})^!_y \stackrel{\mathrm{d}}{=} \Phi_{\zeta_{(\boldsymbol{x},y)}}$, and the thesis is proved.

Proof of Lemma 6.3

For any $B_1, \dots, B_k \in \mathscr{X}$, it holds that

$$\begin{aligned}
M^{(k)}_{\Phi}(B_1, \dots, B_k) &= \mathsf{E}\left[\sum_{X_1,\dots,X_k \in \Phi}^{\neq} \prod_{j=1}^{k} \mathbb{1}_{B_j}(X_j)\right] = \mathsf{E}\left[\mathsf{E}\left[\sum_{X_1,\dots,X_k \in \Phi}^{\neq} \prod_{j=1}^{k} \mathbb{1}_{B_j}(X_j) \mid \Lambda\right]\right] \\
&= \mathsf{E}\left[M^{(k)}_{\Phi \mid \Lambda}(B_1, \dots, B_k)\right] = \mathsf{E}\left[\prod_{j=1}^{k} M_{\Phi \mid \Lambda}(B_j)\right] \\
&= \mathsf{E}\left[\prod_{j=1}^{k} \int_{B_j} \int_{\mathbb{X}\times\mathbb{R}_+} \gamma \kappa(x_j; \theta) \Lambda(\mathrm{d}\theta\, \mathrm{d}\gamma) \mathrm{d}x_j\right] \\
&= \mathsf{E}\left[\int_{(\mathbb{X}\times\mathbb{R}_+)^k} \prod_{j=1}^{k} \gamma_j \int_{B_j} \kappa(x_j; \theta_j) \mathrm{d}x_j \Lambda^k(\mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\gamma})\right] \\
&= \int_{(\mathbb{X}\times\mathbb{R}_+)^k} \prod_{j=1}^{k} \gamma_j \int_{B_j} \kappa(x_j; \theta_j) \mathrm{d}x_j M^k_{\Lambda}(\mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\gamma}),
\end{aligned} \tag{6.33}$$

where $\Lambda$ is the Poisson process appearing in (6.5) and the symbol $\neq$ over the summation in the first line means that the sum is extended over all pairwise distinct points of $\Phi$.

Moreover, from (Baccelli et al., 2020, Lemma 14.E.4), the following holds true for any point process $\xi$ on a Polish space $\mathbb{Y}$ (embedded with the Borel $\sigma$-algebra $\mathscr{Y}$), for any $A_1, \dots, A_k \in \mathscr{Y}$,

$$M^k_{\xi}(A_1, \dots, A_k) = \sum_{q=1}^{k} \sum_{\{J_1,\dots,J_q\}} M^{(q)}_{\xi}\left(\prod_{h=1}^{q} (\cap_{m \in J_h} A_m)\right),$$

where the summation is over all partitions $\{J_1, \dots, J_q\}$ of $\{1, \dots, k\}$. In the special case of $\xi$ being a Poisson process, using a limit argument, we obtain

$$M^k_{\xi}(\mathrm{d}y_1 \dots \mathrm{d}y_k) = \sum_{q=1}^{k} \sum_{\{J_1,\dots,J_q\}} \prod_{h=1}^{q} M_{\xi}(\mathrm{d}y_{(J_h)_1}) \prod_{m \in J_h} \delta_{y_{(J_h)_1}}(\mathrm{d}y_m), \tag{6.34}$$

where $(J_h)_1$ indicates the first index of $J_h$, according to any arbitrary ordering.

Then, since $\Lambda$ in (6.33) is a Poisson process, we can apply the expression in (6.34),

leading to

$$
\begin{aligned}
M_{\Phi}^{(k)}(B_1,\dots,B_k) &= \int_{(\mathbb{X}\times\mathbb{R}_+)^k} \left[\prod_{j=1}^{k}\gamma_j \int_{B_j}\kappa(x_j;\theta_j)\mathrm{d}x_j\right] \\
&\quad \times \sum_{q=1}^{k}\sum_{\{J_1,\dots,J_q\}}\prod_{h=1}^{q} M_{\Lambda}(\mathrm{d}\theta_{(J_h)_1}\mathrm{d}\gamma_{(J_h)_1}) \prod_{m\in J_h}\delta_{(\theta_{(J_h)_1},\gamma_{(J_h)_1})}(\mathrm{d}\theta_m\,\mathrm{d}\gamma_m) \\
&= \sum_{q=1}^{k}\sum_{\{J_1,\dots,J_q\}}\prod_{h=1}^{q}\int_{(\mathbb{X}\times\mathbb{R}_+)^{k_h}}\left[\prod_{m\in J_h}\gamma_m\int_{B_m}\kappa(x_m;\theta_m)\mathrm{d}x_m\right] \\
&\quad \times M_{\Lambda}(\mathrm{d}\theta_{(J_h)_1}\,\mathrm{d}\gamma_{(J_h)_1}) \prod_{m\in J_h}\delta_{(\theta_{(J_h)_1},\gamma_{(J_h)_1})}(\mathrm{d}\theta_m\,\mathrm{d}\gamma_m) \\
&= \sum_{q=1}^{k}\sum_{\{J_1,\dots,J_q\}}\prod_{h=1}^{q}\int_{\mathbb{X}\times\mathbb{R}_+}\gamma^{k_h}\prod_{m\in J_h}\int_{B_m}\kappa(x_m;\theta)\mathrm{d}x_m\; M_{\Lambda}(\mathrm{d}\theta\,\mathrm{d}\gamma) \\
&= \sum_{q=1}^{k}\sum_{\{J_1,\dots,J_q\}}\prod_{h=1}^{q}\int_{\times_{m\in J_h}B_m}\int_{\mathbb{X}\times\mathbb{R}_+}\gamma^{k_h}\prod_{m\in J_h}\kappa(x_m;\theta)M_{\Lambda}(\mathrm{d}\theta\,\mathrm{d}\gamma)\prod_{m\in J_h}\mathrm{d}x_m,
\end{aligned}
$$

where $k_h$ is the cardinality of set $J_h$. For $\boldsymbol{x}_{J_h} = (x_m : m \in J_h)$, define

$$
\eta(\boldsymbol{x}_{J_h}) = \int_{\mathbb{X}\times\mathbb{R}_+}\gamma^{k_h}\prod_{m\in J_h}\kappa(x_m;\theta)M_{\Lambda}(\mathrm{d}\theta\,\mathrm{d}\gamma).
$$

Therefore, the previous expression writes as

$$
\begin{aligned}
M_{\Phi}^{(k)}(B_1,\dots,B_k) &= \sum_{q=1}^{k}\sum_{\{J_1,\dots,J_q\}}\prod_{h=1}^{q}\int_{\times_{m\in J_h}B_m}\eta(\boldsymbol{x}_{J_h})\prod_{m\in J_h}\mathrm{d}x_m \\
&= \sum_{q=1}^{k}\sum_{\{J_1,\dots,J_q\}}\int_{B_1\times\cdots\times B_k}\prod_{h=1}^{q}\eta(\boldsymbol{x}_{J_h})\,\mathrm{d}x_1\,\dots\,\mathrm{d}x_k \\
&= \int_{B_1\times\cdots\times B_k}\sum_{q=1}^{k}\sum_{\{J_1,\dots,J_q\}}\prod_{h=1}^{q}\eta(\boldsymbol{x}_{J_h})\,\mathrm{d}x_1\,\dots\,\mathrm{d}x_k,
\end{aligned}
$$

and the thesis follows, with the introduction of the allocation variables $\boldsymbol{t} \in \mathcal{T}_k$.

Proof of Theorem 6.5

The formulation of the SNCP as cluster process in (6.5) yields a clustering structure at the latent level. Based on this, the generating mechanism of the points can be described in terms of the clustering structure and the cluster-specific parameters of the kernel $\kappa$. Specifically, consider the event $((\theta_1^*, N_1^*),\dots,(\theta_d^*, N_d^*))$, where $d$ denotes the number of clusters among the $M$ points of $\Phi$, $\theta_h^*$ represents the parameter of kernel $\kappa$ in cluster $h$, and $N_h^*$ is the number of points of $\Phi$ in cluster $h$, given a uniform random ordering of the clusters. It holds that $M = \sum_{h=1}^{d} N_h^*$ and the cluster-specific parameters $\theta_h^*$'s are distinct. Denote with $\boldsymbol{\theta}^* = (\theta_1^*,\dots,\theta_d^*)$. The main computation concerns the probability of the

event just described, that is

$$\begin{aligned}\mathsf{P}((\theta_1^*,N_1^*),\dots,(\theta_d^*,N_d^*)) &= \mathsf{E}\left[\mathsf{P}((\theta_1^*,N_1^*),\dots,(\theta_d^*,N_d^*)\,|\,\Lambda)\right]\\ &= \frac{1}{d!}\mathsf{E}\left[\int_{(\mathbb{X}\times\mathbb{R}_+)^d}\prod_{h=1}^{d}\frac{1}{N_h^*!}e^{-\gamma_h}\gamma_h^{N_h^*}\delta_{\theta_h^*}(\theta_h)\cdot\prod_{j>d}e^{-\gamma_j}\Lambda^d(\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\gamma})\right]\\ &= \frac{1}{d!}\mathsf{E}\left[\int_{(\mathbb{X}\times\mathbb{R}_+)^d}e^{-\int_{\mathbb{X}\times\mathbb{R}_+}s\Lambda(\mathrm{d}x\,\mathrm{d}s)}\cdot\prod_{h=1}^{d}\frac{1}{N_h^*!}\gamma_h^{N_h^*}\delta_{\theta_h^*}(\theta_h)\Lambda^d(\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\gamma})\right],\end{aligned}$$

where $\Lambda^d$ is the $d$-th power of $\Lambda$. Since the term $\prod_{h=1}^{d}\delta_{\theta_h^*}(\theta_h)$ entails that the integrand is zero on sets of the type

$$\{(\boldsymbol{\theta},\boldsymbol{\gamma})\in(\mathbb{X}\times\mathbb{R}_+)^d:\ \theta_i=\theta_j \text{ for } i\neq j\},$$

then we can replace $\Lambda^d$ with the $d$-th factorial power $\Lambda^{(d)}$. By applying the CLM formula, we obtain

$$\begin{aligned}&\mathsf{P}((\theta_1^*,N_1^*),\dots,(\theta_d^*,N_d^*))\\ &\quad= \frac{1}{d!}\mathsf{E}\left[\int_{(\mathbb{X}\times\mathbb{R}_+)^d}e^{-\int_{\mathbb{X}\times\mathbb{R}_+}s\Lambda(\mathrm{d}x\,\mathrm{d}s)}\cdot\prod_{h=1}^{d}\frac{1}{N_h^*!}\gamma_h^{N_h^*}\delta_{\theta_h^*}(\theta_h)\Lambda^{(d)}(\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\gamma})\right]\\ &\quad= \frac{1}{d!}\int_{(\mathbb{X}\times\mathbb{R}_+)^d}\mathsf{E}\left[e^{-\int_{\mathbb{X}\times\mathbb{R}_+}s\Lambda^!_{\boldsymbol{\theta},\boldsymbol{\gamma}}(\mathrm{d}x\,\mathrm{d}s)}\right]e^{-\sum_{h=1}^{d}\gamma_h}\prod_{h=1}^{d}\frac{1}{N_h^*!}\gamma_h^{N_h^*}\delta_{\theta_h^*}(\theta_h)\ M_\Lambda^{(d)}(\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\gamma})\\ &\quad= \frac{1}{d!}\int_{\mathbb{R}_+^d}\mathsf{E}\left[e^{-\int_{\mathbb{X}\times\mathbb{R}_+}s\Lambda^!_{\boldsymbol{\theta}^*,\boldsymbol{\gamma}}(\mathrm{d}x\,\mathrm{d}s)}\right]\prod_{h=1}^{d}\frac{1}{N_h^*!}e^{-\gamma_h}\gamma_h^{N_h^*}\ M_\Lambda^{(d)}(\mathrm{d}\boldsymbol{\theta}^*\,\mathrm{d}\boldsymbol{\gamma}).\end{aligned}\tag{6.35}$$

Since $\Lambda$ is a Poisson process with intensity measure $\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)$, then (6.35) yields

$$\begin{aligned}\mathsf{P}((\theta_1^*,N_1^*),\dots,(\theta_d^*,N_d^*)) &= \frac{1}{d!}\mathsf{E}\left[e^{-\int_{\mathbb{X}\times\mathbb{R}_+}s\Lambda(\mathrm{d}x\,\mathrm{d}s)}\right]\int_{\mathbb{R}_+^d}\prod_{h=1}^{d}\frac{1}{N_h^*!}e^{-\gamma_h}\gamma_h^{N_h^*}\ \nu^d(\mathrm{d}\boldsymbol{\theta}^*\,\mathrm{d}\boldsymbol{\gamma})\\ &= \frac{1}{d!}e^{-\int_{\mathbb{X}\times\mathbb{R}_+}(1-e^{-\gamma})\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)}\prod_{h=1}^{d}\int_{\mathbb{R}_+}\frac{1}{N_h^*!}e^{-\gamma_h}\gamma_h^{N_h^*}\ \nu(\mathrm{d}\theta_h^*\,\mathrm{d}\gamma_h).\end{aligned}$$

For any clustering $(C_1,\dots,C_d)$ of the $M$ points of $\Phi$, with cluster cardinalities $N_h^*=|C_h|$, $h=1,\dots,d$, note that the following relationship holds

$$\mathsf{P}((\theta_1^*,C_1),\dots,(\theta_d^*,C_d)) = \binom{M}{N_1^*,\dots,N_k^*}^{-1}\mathsf{P}((\theta_1^*,N_1^*),\dots,(\theta_d^*,N_d^*)).$$

The marginal distribution of $(C_1,\dots,C_d)$ is simply obtained by integrating out $(\theta_1^*,\dots,\theta_d^*)$ from the previous expression, leading to

$$\mathsf{P}(C_1,\dots,C_d) = \frac{1}{M!}\frac{1}{d!}e^{-\int_{\mathbb{X}\times\mathbb{R}_+}(1-e^{-\gamma})\nu(\mathrm{d}\theta\,\mathrm{d}\gamma)}\prod_{h=1}^{d}\int_{\mathbb{X}\times\mathbb{R}_+}e^{-\gamma_h}\gamma_h^{N_h^*}\ \nu(\mathrm{d}\theta_h\,\mathrm{d}\gamma_h).$$

Observe that $\mathsf{P}(\{C_1,\dots,C_d\})$ is obtained by multiplying by $d!$ the previous expression. In the statement of the theorem, we stress that this probability distribution describes the number of points $M$ as well.

# Bibliography


Aldous, D. J. (1985). Exchangeability and related topics. *Ecole d'Eté de Probabilités de Saint-Flour XIII1983*, 1–198.

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

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

Ascolani, F., B. Franzolini, A. Lijoi, and I. Prünster (2024). Nonparametric priors with full-range borrowing of information. *Biometrika 111*(3), 945–969.

Ayed, F. and F. Caron (2021). Nonnegative Bayesian nonparametric factor models with completely random measures. *Statistics and Computing 31*(5), 31–63.

Bacallado, S., M. Battiston, S. Favaro, and L. Trippa (2017). Sufficientness Postulates for Gibbs-Type Priors and Hierarchical Generalizations. *Statistical Science 32*(4), 487 – 500.

Baccelli, F., B. Błaszczyszyn, and M. Karray (2020). Random measures, point processes, and stochastic geometry. *HAL preprint available at https://hal.inria.fr/hal-02460214/*.

Baddeley, A., E. Rubak, and R. Turner (2015). *Spatial Point Patterns: Methodology and Applications with R*. Boca Raton: Chapman & Hall/CRC.

Baddeley, A. and R. Turner (2005). Spatstat: an R package for analysing spatial point patterns. *Journal of Statistical Software 12*(6), 1–42.

Balocchi, C., S. Favaro, and Z. Naulet (2024). Bayesian Nonparametric Inference for "Species-sampling" Problems. *arXiv preprint arXiv:2203.06076*.

Bassetti, F., R. Casarin, and L. Rossini (2020). Hierarchical Species Sampling Models. *Bayesian Analysis 15*(3), 809 – 838.

Battiston, M., S. Favaro, D. M. Roy, and Y. W. Teh (2018). A characterization of product-form exchangeable feature probability functions. *The Annals of Applied Probability 28*(3), 1423–1448.

Benedetto, G. D., F. Caron, and Y. W. Teh (2020). Non-exchangeable feature allocation models with sublinear growth of the feature sizes. In S. Chiappa and R. Calandra (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and*

*Statistics*, Volume 108 of *Proceedings of Machine Learning Research*, pp. 3208–3218. PMLR.

Beraha, M., R. Argiento, F. Camerlenghi, and A. Guglielmi (2025). Bayesian mixture models with repulsive and attractive atoms. *Journal of the Royal Statistical Society Series B: Statistical Methodology (In press)*.

Beraha, M., R. Argiento, J. Møller, and A. Guglielmi (2022). MCMC Computations for Bayesian Mixture Models Using Repulsive Point Processes. *Journal of Computational and Graphical Statistics 31*(2), 422–435.

Beraha, M. and S. Favaro (2025). Transform-scaled process priors for trait allocations in Bayesian nonparametrics. *Electronic Journal of Statistics 19*(2), 3532–3563.

Beraha, M. and J. E. Griffin (2023). Normalised latent measure factor models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(4), 1247–1270.

Beraha, M., A. Guglielmi, and F. A. Quintana (2021). The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions. *Bayesian Analysis 16*(4), 1187–1219.

Beraha, M., L. Masoero, S. Favaro, and T. Richardson (2024). A nonparametric bayes approach to online activity prediction. *Manuscript available upon request*.

Berti, P., E. Dreassi, F. Leisen, L. Pratelli, and P. Rigo (2025). A Probabilistic View on Predictive Constructions for Bayesian Learning. *Statistical Science 40*(1), 25–39.

Bianchini, I., A. Guglielmi, and F. A. Quintana (2020). Determinantal point process mixtures via spectral density approach. *Bayesian Analysis 15*, 187–214.

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

Borgoni, R., C. Galimberti, and D. Zappa (2021). Identification of spatial defects in semiconductor manufacturing. *Applied Stochastic Models in Business and Industry 37*(5), 878–893.

Broderick, T., M. I. Jordan, and J. Pitman (2012). Beta processes, stick-breaking and power laws. *Bayesian Analysis 7*(2), 439–476.

Broderick, T., L. Mackey, J. Paisley, and M. I. Jordan (2015). Combinatorial clustering and the beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 290–306.

Broderick, T., J. Pitman, and M. I. Jordan (2013). Feature Allocations, Probability Functions, and Paintboxes. *Bayesian Analysis 8*(4), 801 – 836.

Broderick, T., A. C. Wilson, and M. I. Jordan (2018). Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli 24*(4B), 3181 – 3221.

Calderoni, F., D. Brunetto, and C. Piccardi (2017). Communities in criminal networks: A case study. *Social Networks 48*, 116–125.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodríguez (2019). Latent Nested Nonparametric Priors (with Discussion). *Bayesian Analysis 14*(4), 1303 – 1356.

Camerlenghi, F. and S. Favaro (2021). On Johnson's "Sufficientness" postulates for feature-sampling models. *Mathematics 9*(22).

Camerlenghi, F., S. Favaro, L. Masoero, and T. Broderick (2024). Scaled process priors for Bayesian nonparametric estimation of the unseen genetic variation. *Journal of the American Statistical Association 119*(545), 320–331.

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

Campbell, T., D. Cai, and T. Broderick (2018). Exchangeable trait allocations. *Electronic Journal of Statistics 12*(2), 2290–2322.

Catalano, M., H. Lavenant, A. Lijoi, and I. Prünster (2024). A Wasserstein Index of Dependence for Random Measures. *Journal of the American Statistical Association 119*(547), 2396–2406.

Catalano, M., A. Lijoi, and I. Prünster (2021). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *The Annals of Statistics 49*(5), 2916–2947.

Chakraborty, S., A. Arora, C. B. Begg, and R. Shen (2019). Using somatic variant richness to mine signals from rare variants in the cancer genome. *Nature Communications 10*, 5506.

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

Charalambides, C. and J. Singh (1988). A review of the stirling numbers, their generalizations and statistical applications. *Communication in Statistics - Theory and Methods 17*, 2533–2595.

Charalambides, C. A. (2005). *Combinatorial Methods in Discrete Distributions*. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons.

Chiu, C.-H. (2022). Incidence-data-based species richness estimation via a beta-binomial model. *Methods in Ecology and Evolution 13*(11), 2546–2558.

Chiu, C.-H. (2023). Sample coverage estimation, rarefaction, and extrapolation based on sample-based abundance data. *Ecology 104*(8), 1–15.

Chiu, C.-H., Y.-T. Wang, B. A. Walther, and A. Chao (2014). An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. *Biometrics 70*(3), 671–682.

Choi, C.-S., J. O. Woo, and J. G. Andrews (2017). An analytical framework for modeling a spatially repulsive cellular network. *arXiv preprint arXiv:1701.02261*.

Coeurjolly, J.-F., J. Møller, and R. Waagepetersen (2017). A tutorial on palm distributions for spatial point processes. *International Statistical Review 85*(3), 404–420.

Colombi, A., R. Argiento, F. Camerlenghi, and L. Paci (2025). Hierarchical Mixture of Finite Mixtures. *Bayesian Analysis*, 1 – 29.

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

Colwell, R. K., A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino (2012, 03). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology 5*(1), 3–21.

Cremaschi, A., T. M. Wertz, and M. De Iorio (2024). Repulsion, chaos, and equilibrium in mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 87*(2), 389–432.

Daley, D. J. and D. Vere-Jones (2003). *An introduction to the theory of point processes. Vol. I* (Second ed.). Probability and its Applications (New York). New York: Springer-Verlag. Elementary theory and methods.

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

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

De Blasi, P., A. Lijoi, and I. Prünster (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica 23*(3), 1299–1321.

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

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles 739*, 5–18. Translated In: Studies in Inductive and Probability, II. Jeffrey, R. (ed.) University of California Press: Berkeley 1980.

Deng, C., T. Daley, G. De Sena Brandine, and A. D. Smith (2019). Molecular heterogeneity in large-scale biological data: Techniques and applications. *Annual Review of Biomedical Data Science 2*(Volume 2, 2019), 39–67.

Denti, F., F. Camerlenghi, M. Guindani, and A. Mira (2023). A common atoms model for the bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association 118*(541), 405–416.

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.

Dunson, D. B. and C. Xing (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association 104*(487), 1042–1051.

Erdélyi, A. and F. G. Tricomi (1951). The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics 1*(1), 133 – 142.

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

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology 3*(1), 87–112.

Favaro, S., A. Lijoi, R. H. Mena, and I. Prünster (2009). Bayesian non-parametric inference for species variety with a two-parameter poisson–dirichlet process prior. *Journal of the Royal Statistical Society Series B: Statistical Methodology 71*(5), 993–1008.

Favaro, S., A. Lijoi, and I. Prünster (2012). Asymptotics for a Bayesian nonparametric estimator of species variety. *Bernoulli 18*(4), 1267 – 1283.

Favaro, S., B. Nipoti, and Y. W. Teh (2016). Rediscovery of good–turing estimators via bayesian nonparametrics. *Biometrics 72*(1), 136–145.

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

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi, J. S. Rustagi, and D. Siegmund (Eds.), *Recent Advances in Statistics*, pp. 287–302. Academic Press.

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

Fong, E., C. Holmes, and S. G. Walker (2023). Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(5), 1357–1391.

Fortini, S. and S. Petrone (2012). Predictive construction of priors in Bayesian nonparametrics. *Brazilian Journal of Probability and Statistics 26*(4), 423 – 449.

Franzolini, B., M. De Iorio, and J. Eriksson (2023). Conditional partial exchangeability: a probabilistic framework for multi-view clustering. *arXiv preprint arXiv:2307.01152*.

Franzolini, B., A. Lijoi, I. Prünster, and G. Rebaudo (2025). Multivariate species sampling models. *arXiv preprint arXiv:2503.24004*.

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

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

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

Gnedin, A. and J. Pitman (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov, POMI 325*, 83–102.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40*(3-4), 237–264.

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

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*(2), 215–231.

Gotelli, N. J. and R. K. Colwell (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters 4*(4), 379–391.

Gravel, S. (2014). Predicting discovery rates of genomic features. *Genetics 197*(2), 601–610.

Gregoire, G. (1984). Negative binomial distributions for point processes. *Stochastic Processes and their Applications 16*(2), 179–188.

Griffin, J. E., M. Kolossiatis, and M. F. Steel (2013). Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society Series B: Statistical Methodology 75*(3), 499–529.

Griffin, J. E. and F. Leisen (2017). Compound random measures and their use in bayesian non-parametrics. *Journal of the Royal Statistical Society Series B: Statistical Methodology 79*(2), 525–545.

Griffiths, T. L. and Z. Ghahramani (2005). Infinite latent feature models and the indian buffet process. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'05, Cambridge, MA, USA, pp. 475–482. MIT Press.

Griffiths, T. L. and Z. Ghahramani (2011). The Indian buffet process: an introduction and review. *Journal of Machine Learning Research 12*(32), 1185–1224.

Hagenaars, J. A. and A. L. McCutcheon (2002). *Applied Latent Class Analysis*. Cambridge University Press.

Haugh, M. (2015). The EM Algorithm. Lecture notes for IEOR E4570: Machine Learning for OR&FE Department at Columbia University.

Heaukulani, C. and D. M. Roy (2016). The combinatorial structure of beta negative binomial processes. *Bernoulli 22*(4), 2301 – 2324.

Hewitt, E. and L. J. Savage (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society 80*(2), 470–501.

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

Hough, J. B., M. Krishnapur, Y. Peres, and B. Viràg (2006). Determinantal processes and independence. *Probability Surveys 3*, 206–229.

Hu, Y., K. Zhai, S. Williamson, and J. Boyd-Graber (2012). Modeling images using transformed Indian buffet processes. In *International Conference of Machine Learning*.

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan (2008). *Statistical analysis and modelling of spatial point patterns*. Statistics in Practice. John Wiley & Sons, Ltd., Chichester.

Ionita-Laza, I., C. Lange, and N. M. Laird (2009). Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences 106*(13), 5008–5013.

James, L. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *arXiv preprint arXiv:math/0205093*.

James, L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics 33*(4), 1771 – 1799.

James, L. F. (2006). Poisson calculus for spatial neutral to the right processes. *The Annals of Statistics 34*(1), 416 – 440.

James, L. F. (2017). Bayesian Poisson calculus for latent feature modeling via generalized Indian Buffet Process priors. *The Annals of Statistics 45*(5), 2016–2045.

James, L. F., J. Lee, and A. Pandey (2024). Bayesian analysis of generalized hierarchical Indian buffet processes for within and across group sharing of latent features. *arXiv preprint arXiv:2304.05244*.

James, L. F., J. Lee, and A. Pandey (2025). Poisson hierarchical indian buffet processes for within and across group sharing of latent features-with indications for microbiome species sampling models. *arXiv preprint arXiv:2502.01919*.

James, L. F., P. Orbanz, and Y. W. Teh (2015). Scaled subordinators and generalizations of the indian buffet process. *arXiv preprint arXiv:1510.07309*.

Kallenberg, O. (1973). Characterization and convergence of random measures and point processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 27*(1), 9–21.

Kallenberg, O. (1984). An informal guide to the theory of conditioning in point processes. *International Statistical Review / Revue Internationale de Statistique 52*(2), 151–164.

Kallenberg, O. (2017). *Random measures, theory and applications*, Volume 1. Springer.

Kallenberg, O. (2021). *Foundations of Modern Probability*. Springer.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *The Annals of Statistics 27*(2), 562–588.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics 21*, 59–78.

Knowles, D. and Z. Ghahramani (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics 5*(2B), 1534–1552.

Kulesza, A. and B. Taskar (2012). *Determinantal Point Processes for Machine Learning*, Volume 5 of *Foundations and Trends in Machine Learning*. Now Publishers.

Last, G. and M. Penrose (2017). *Lectures on the Poisson process*, Volume 7. Cambridge University Press.

Lavancier, F., J. Møller, and E. Rubak (2015). Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(4), 853–877.

Lavancier, F. and E. Rubak (2023). On simulation of continuous determinantal point processes. *Statistics and Computing 33*(45). Publisher Copyright: © 2023, The Author(s).

Lazarsfeld, P. F. and N. W. Henry (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lee, J., X. Miscouridou, and F. Caron (2023). A unified construction for series representations and finite approximations of completely random measures. *Bernoulli 29*(3), 2142–2166.

Legramanti, S., T. Rigon, D. Durante, and D. B. Dunson (2022). Extended stochastic block models with application to criminal networks. *The Annals of Applied Statistics 16*(4), 2369–2395.

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

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

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

Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. In *Bayesian nonparametrics*, Volume 28 of *Cambridge Series in Statistical and Probabilistic Mathematics*, pp. 80–136. Cambridge Univ. Press, Cambridge.

Lijoi, A., I. Prünster, and S. G. Walker (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *The Annals of Applied Probability 18*(4), 1519–1547.

Lijoi, A., I. Prünster, and G. Rebaudo (2023). Flexible clustering via hidden hierarchical dirichlet priors. *Scandinavian Journal of Statistics 50*(1), 213–234.

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

Lijoi, A., I. Prünster, and T. Rigon (2024). Finite-dimensional discrete random structures and Bayesian clustering. *Journal of the American Statistical Association 119*(546), 929–941.

Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics 12*(1), 351 – 357.

Lo, A. Y. (1991). A characterization of the dirichlet process. *Statistics & Probability Letters 12*(3), 185–187.

Lu, C., D. Durante, and N. Friel (2025). Zero-inflated stochastic block modeling of efficiency-security trade-offs in weighted criminal networks. *Journal of the Royal Statistical Society Series A: Statistics in Society (In Press)*.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pp. 50–55.

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

Magurran, A. E. and B. J. McGill (2011). *Biological Diversity: frontiers in measurement and assessment*. Oxford Biology.

Masoero, L., M. Beraha, S. Favaro, and T. Richardson (2024). Improved prediction of future user activity in online a/b testing. *Manuscript available upon request*.

Masoero, L., F. Camerlenghi, S. Favaro, and T. Broderick (2018). Posterior representations of hierarchical completely random measures in trait allocation models. In *NeurIPS Workshop on All of Bayesian Nonparametrics*.

Masoero, L., F. Camerlenghi, S. Favaro, and T. Broderick (2022). More for less: predicting and maximizing genomic variant discovery via Bayesian nonparametrics. *Biometrika 109*(1), 17–32.

Masoero, L., J. Schraiber, and T. Broderick (2021). Bayesian nonparametric strategies for power maximization in rare variants association studies. *arXiv preprint arXiv:2112.02032*.

Mazziotta, A., J. Heilmann-Clausen, H. H. Bruun, O. Fritz, E. Aude, and A. P. Tøttrup (2016a). Dataset on species incidence, species richness and forest characteristics in a danish protected area. *Data in Brief 9*, 895–897.

Mazziotta, A., J. Heilmann-Clausen, H. H. Bruun, O. Fritz, E. Aude, and A. P. Tøttrup (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.

Meilă, M. (2003). Comparing clusterings by the variation of information. In B. Schölkopf and M. K. Warmuth (Eds.), *Learning Theory and Kernel Machines*, Berlin, Heidelberg, pp. 173–187. Springer Berlin Heidelberg.

Meyer, S., L. Held, and M. Höhle (2017). Spatio-temporal analysis of epidemic phenomena using the `R` package `surveillance`. *Journal of Statistical Software 77*(11), 1–55.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356.

Miller, K. T., T. L. Griffiths, and M. I. Jordan (2009). Nonparametric latent feature models for link prediction. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'09, Red Hook, NY, USA, pp. 1276–1284. Curran Associates Inc.

Møller, J. (2003). Shot noise Cox processes. *Advances in Applied Probability 35*(3), 614–640.

Møller, J. and R. P. Waagepetersen (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.

Momozawa, Y. and K. Mizukami (2021). Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of Human Genetics 66*, 11–23.

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

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Ngo Bieng, M. A., C. Ginisty, and F. Goreaud (2011). Point process models for mixed sessile forest stands. *Annals of Forest Science 68*, 267–274.

Nguyen, T. D., J. Huggins, L. Masoero, L. Mackey, and T. Broderick (2024). Independent finite approximations for Bayesian nonparametric inference. *Bayesian Analysis 19*(4), 1187–1224.

Ogata, Y. (1998). Space–time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics 50*(2), 379–402.

Ord, K. (1978). How many trees in a forest? *The Mathematical Scientist 3*, 23–33.

Orlitsky, A., A. T. Suresh, and Y. Wu (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences 113*(47), 13283–13288.

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

Perman, M., J. Pitman, and M. Yor (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields 92*(1), 21–39.

Petralia, F., V. Rao, and D. Dunson (2012). Repulsive mixtures. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 25. Curran Associates, Inc.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields 102*(2), 145–158.

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.

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

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

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

Regazzini, E. (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *Giornale dell'Istituto italiano degli attuari 41*, 77–89.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability 13*(2), 255–266.

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

Sanders, J. G., S. Nurk, R. A. Salido, J. Minich, Z. Z. Xu, Q. Zhu, C. Martino, M. Fedarko, T. D. Arthur, and F. Chen (2019). Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biology 20*(1), 1–14.

Shen, Y., L. Masoero, J. G. Schraiber, and T. Broderick (2024). Double trouble: Predicting new variant counts across two heterogeneous populations. *arXiv preprint arXiv:2403.02154* .

Steele, J. M. (1994). Le cam's inequality and poisson approximations. *The American Mathematical Monthly 101*(1), 48–54.

Stolf, F. and D. B. Dunson (2025). Infinite joint species distribution models. *Biometrika (In press)*.

Tanaka, U. and Y. Ogata (2014). Identification and estimation of superposed neyman–scott spatial cluster processes. *Annals of the Institute of Statistical Mathematics 66*, 687–702.

Teh, Y. and D. Gorur (2009). Indian buffet processes with power-law behavior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, Volume 22.

Teh, Y. W. and M. I. Jordan (2010). Hierarchical bayesian nonparametric models with applications. In *Bayesian nonparametrics*, Volume 28 of *Cambridge Series in Statistical and Probabilistic Mathematics*, pp. 158–207. Cambridge Univ. Press, Cambridge.

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

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

Titsias, M. (2007). The infinite gamma-poisson feature model. In J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), *Advances in Neural Information Processing Systems*, Volume 20. Curran Associates, Inc.

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

Wang, Y., A. Degleris, A. Williams, and S. W. Linderman (2023). Spatiotemporal clustering with neyman-scott processes via connections to bayesian nonparametric mixture models. *Journal of the American Statistical Association*, 1–14.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research 14*(27), 867–897.

Wei, G. C. and M. A. Tanner (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association 85*(411), 699–704.

Williamson, S., C. Wang, K. A. Heller, and D. M. Blei (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.

Xie, F. and Y. Xu (2019). Bayesian repulsive gaussian mixture model. *Journal of the American Statistical Association*, 187–203.

Xu, H., D. Luo, X. Chen, and L. Carin (2018). Benefits from superposed hawkes processes. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, Volume 84 of *PMLR*, pp. 623–631.

Xu, Y., P. Müller, and D. Telesca (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics 72*(3), 955–964.

Zabell, S. L. (1982). W. E. Johnson's "Sufficientness" Postulate. *The Annals of Statistics 10*(4), 1090 – 1099.

Zabell, S. L. (2005). *Symmetry and its discontents*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge University Press, New York. Essays on the history of inductive probability, With a preface by Brian Skyrms.

Zhang, M. J., V. Ntranos, and D. Tse (2020). Determining sequencing depth in a single-cell rna-seq experiment. *Nature Communications 11*(774).

Zhou, M., S. Favaro, and S. G. Walker (2017). Frequency of frequencies distributions and size-dependent exchangeable random partitions. *Journal of the American Statistical Association 112*(520), 1623–1635.

Zhou, M., L. Hannah, D. Dunson, and L. Carin (2012). Beta-negative binomial process and poisson factor analysis. In N. D. Lawrence and M. Girolami (Eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, Volume 22 of *Proceedings of Machine Learning Research*, La Palma, Canary Islands, pp. 1462–1471. PMLR.

Zhou, M., O. H. M. Padilla, and J. G. Scott (2016). Priors for random count matrices derived from a family of negative binomial processes. *Journal of the American Statistical Association 111*(515), 1144–1156.

Zito, A., T. Rigon, and D. B. Dunson (2023). Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa. *Methods in Ecology and Evolution 14*(2), 529–542.

Zito, A., T. Rigon, and D. B. Dunson (2024). Bayesian Nonparametric Modeling of Latent Partitions via Stirling-Gamma Priors. *Bayesian Analysis*, 1 – 28.

Zito, A., T. Rigon, O. Ovaskainen, and D. B. Dunson (2023). Bayesian modeling of sequential discoveries. *Journal of the American Statistical Association 188*(544), 2521–2532.

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