On New Advances In Nonparametric Bayesian Priors and Their Applications

Sadegh Chegini

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Mathematics and Statistics*

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

*The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Bayesian nonparametric inference requires the construction of priors on infinite-dimensional spaces, such as the space of cumulative distribution functions. Well-known priors on this space include the Dirichlet process and the two-parameter Poisson–Dirichlet process. In this thesis, we explore a distinctive functional of the Poisson point process, known as the negative binomial process. While the increments of the negative binomial process are not independent, they become conditionally independent given an underlying gamma variable. We propose a novel point process representation for the negative binomial process, which extends the Poisson-Kingman distribution and its associated random discrete probability measure. The new proposed family of the discrete random probability measures which is defined by normalizing the points of the negative binomial process provides a new set of useful priors for Bayesian nonparametric models with more flexibility compared to the random discrete probability measure which are obtained by normalizing the points of a Poisson point process. We illustrate how this family encompasses several well-known priors, such as the Dirichlet process, the normalized positive α-stable process, and the Poisson–Dirichlet process. Using the same gamma Lévy measure, we derive an extension of the Dirichlet process along with an almost sure approximation. Additionally, leveraging our negative binomial process representation, we develop a new series representation for the Poisson–Dirichlet process. Through simulations, we demonstrate how adopting priors from this family can enhance the performance of Bayesian nonparametric hierarchical models.

In the literature, the term *negative binomial process* has been used to describe several distinct stochastic processes, each playing a significant role in probability theory and statistics, particularly in Bayesian nonparametric analysis. However, the presence of multiple, and at times conflicting, definitions has led to considerable ambiguity. This thesis addresses this issue by systematically reviewing the various definitions and clarifying their distinctions. The aim is to provide a comprehensive overview that helps practitioners recognize the differences between these processes and avoid potential misunderstandings. Furthermore, for one of the definitions of the negative binomial process, we present an extension from the univariate case to a bivariate form.

We also examine the Liouville distribution, a well-known conjugate prior for the multinomial distribution, which addresses certain limitations of the Dirichlet distribution, particularly its tendency to induce negative correlations. We construct a discrete random probability measure based on a random vector following a Liouville distribution and establish its weak limit to define the proposed *Liouville process*. This process takes the form of a spike-and-slab model, where the slab is represented by a Dirichlet process and the spike corresponds to a single point drawn from its mean. These components are combined through a random convex mixture, with weights governed by the Liouville distribution. By placing the Liouville process as a prior over the space of probability measures, we derive both its posterior and predictive distributions.

Dedications

To the loving memory of my father. You will always be in my heart

Acknowledgements

I would like to express my deepest respect and sincere gratitude to my supervisor, Professor Mahmoud Zarepour, for his invaluable supervision, expert guidance, unwavering support, and constant encouragement throughout my PhD journey. His mentorship has been a cornerstone of this work, and I am truly grateful for his presence every step of the way. This thesis would not have been possible without him.

I am also grateful to Professor Mayer Alvo, Professor Rafal Kulik, Professor Andrei Volodin, and Professor Jason Nielson for generously serving on my committee and for their thoughtful feedback and contributions.

I extend my appreciation to the University of Ottawa and the Department of Mathematics and Statistics for providing me with financial support and an intellectually stimulating academic environment throughout my studies.

My heartfelt thanks also go to everyone who supported me in any way during the process of writing this thesis.

Finally, I am profoundly grateful to my mother and siblings for their unconditional love, patience, sacrifices, and unwavering support. Words cannot fully express how much their strength and understanding have meant to me. Thank you from the bottom of my heart.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In statistical data analysis, we often encounter datasets where the goal is to fit a model to infer key characteristics of the underlying distribution. These characteristics can range from simple parameters, such as the mean, to more complex descriptions, such as the entire probability distribution. Regardless of the complexity, constructing an appropriate statistical model that accurately reflects the data is essential. This process involves estimating a set of parameters that capture the underlying structural aspects of the observed data. There are two primary approaches to address this problem: the Frequentist and Bayesian frameworks.

The **Frequentist approach** interprets probability as the long-run frequency of events. In this paradigm, data are treated as repeatable random samples, and the underlying parameters are assumed to remain fixed across repetitions of the experiment.

On the other hand, **Bayesian parametric inference** assumes that the data are generated from a family of probability measures $\{F_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ represents the parameter space. Typically, the parameter θ is finite dimensional and is treated as a random variable, with a prior distribution assigned to incorporate any prior knowledge or beliefs about its possible values. Bayes' theorem then combines the likelihood of the observed data with the prior distribution to form the posterior distribution of θ. This posterior allows us to make inferences about the parameters, such as calculating posterior means, standard deviations, or credible intervals.

However, parametric models can be limiting if the chosen form of the model does not adequately capture the underlying data-generating process, potentially leading to unsatisfactory inferences. To address this limitation, **Bayesian nonparametric inference** provides a more flexible framework by constructing priors over infinite-dimensional spaces, such as the space of all possible distribution functions. In the literature, this nonparametric extension of Bayesian inference is referred to as *Bayesian nonparametric inference* [22].

There are two key requirements when constructing priors in Bayesian nonparametric settings: (1) the posterior distribution must remain mathematically tractable given the observations, and (2) there must be an efficient simulation algorithm to draw samples from both the prior and posterior distributions. A desirable property in this context is *conjugacy*, where the posterior distribution belongs to the same family as the prior distribution. Conjugate priors are particularly advantageous from a computational perspective, as they enable the use of similar algorithms for sampling from both the prior and the posterior.

A prominent prior that plays a central role in Bayesian nonparametric inference is the Dirichlet process, introduced by [18]. However, the Dirichlet process is not universally suitable for all types of data. As a result, considerable research has focused on developing generalizations and alternative processes to extend its applicability. Notable examples include the two-parameter Poisson-Dirichlet process [56], normalized random measures with independent increments [58, 37, 38], the normalized inverse-Gaussian process [47], the normalized generalized gamma process [48], the beta process [26], and the beta-Stacy process [66].

The outline of this thesis is as follows: In Chapter 2, we introduce foundational definitions and concepts, along with a review of key properties of Bayesian nonparametric priors, with a particular focus on the Dirichlet process and the two-parameter Poisson-Dirichlet process. The notation and results presented in this chapter will serve as essential tools for the subsequent chapters.

In Chapter 3, we derive a new representation for the negative binomial process directly as a functional of the Poisson random measure. Then using this representation of the negative binomial process, we provide a family of generalized Poisson-Kingman distribution and its associated random discrete probability measure which contains many well-known priors in nonparametric Bayesian analysis such as the Dirichlet process, the Poisson-Dirichlet process, the normalized generalized gamma process, etc. A natural extension of the Dirichlet process is formulated as a functional of the proposed series representation for the negative binomial process. We also provide an almost sure convergent approximation for this extended Dirichlet process. Another by-product of our proposed series representation for the negative binomial process is a new series representation for the two-parameter Poisson-Dirichlet process. It is shown that an approximation based on this new representation for the Poisson-Dirichlet process is very efficient, as illustrated in a simulation study.

In Chapter 4, we apply the new family of priors introduced in Chapter 3 within a Bayesian nonparametric hierarchical model. We then compare the results with those obtained using a model based on the Dirichlet prior through a simulation study.

The term "negative binomial process" is used to describe various processes across different fields, particularly in Bayesian nonparametrics, leading to confusion due to conflicting definitions in the literature. In Chapter 5, we address this ambiguity by systematically reviewing the various definitions of the negative binomial process and emphasizing their key distinctions. The objective is to provide a clear and comprehensive overview that enables practitioners to distinguish between these processes and avoid potential misinterpretations. In addition, we extend the univariate negative binomial process introduced in Chapter 3 to a bivariate form.

In Chapter 6, we address some limitations of the Dirichlet distribution, such as its tendency to impose negative correlations, and explore the Liouville distribution as an effective alternative. We introduce a discrete random probability measure based on a random vector following a Liouville distribution and derive its weak limit to define the Liouville process. The resulting process takes the form of a spike-and-slab model, where the Dirichlet process serves as the slab, and a single point drawn from its mean represents the spike. These two components are linearly combined with a random weight governed by a Liouville distribution. By placing the Liouville process as a prior on the space of probability measures, we derive its corresponding posterior and predictive distributions.

Finally, in Chapter 7, we provide a summary of the thesis and discuss potential directions for future research.

Chapter 2

Basic tools and background

In this chapter, we lay the groundwork for the upcoming chapters by first introducing essential tools and definitions. We then review key properties of some well-known Bayesian nonparametric priors, with particular focus on the Dirichlet process, the two-parameter Poisson-Dirichlet process, and pure jump processes. The notations and results presented here will be applied throughout the subsequent chapters. For conciseness, the proofs of these results are omitted, as they are available in the original literature with precise references provided.

2.1 Definitions and concepts

Definition 2.1.1 (**Random measure** [59])**.** Let $\mathbb{E}$ be a Polish space and $\mathscr{E}$ be the Borel σ-algebra generated by the open sets in $\mathbb{E}$. A measure μ is called Radon if $\mu(K) < \infty$ for any compact set K in $\mathbb{E}$. Let $M(\mathbb{E})$ be the space of Radon measures on $\mathbb{E}$. Let $\mathcal{M}(\mathbb{E})$ be the smallest σ-algebra of subsets of $M(\mathbb{E})$ making the maps $\mu \to \mu(f) = \int f \, d\mu$ from $M(\mathbb{E})$ to $\mathbb{R}$ measurable for all functions $f \in C_K^+(\mathbb{E})$, where $C_K^+(\mathbb{E})$ denotes the set of continuous functions $f : \mathbb{E} \to [0, \infty)$ with compact support. Note that $\mathcal{M}(\mathbb{E})$ is the Borel σ-algebra generated by the topology of vague convergence. If $\mu_n, \mu \in M(\mathbb{E})$, we say that $(\mu_n)_{n\geq 1}$ converges vaguely to μ (and we write $\mu_n \xrightarrow{v} \mu$) if $\mu_n(f) \to \mu(f)$ for any $f \in C_K^+(\mathbb{E})$. A *random measure* on $\mathbb{E}$ is any measurable map ξ defined on a probability space $(\Omega, \mathcal{F}, P)$ with values in $(M(\mathbb{E}), \mathcal{M}(\mathbb{E}))$.

Definition 2.1.2 (**Completely random measure** [42]). A random measure ξ is a completely random measure if, for any finite collection $A_1, \ldots, A_n$ of disjoint sets, the random variables $\xi(A_1), \ldots, \xi(A_n)$ are independent.

Definition 2.1.3 (**Point process**). Let $\mathbb{E}$ be a locally compact space with a countable basis. Let $\mathscr{E}$ be a Borel σ-algebra of subsets of $\mathbb{E}$. Let $\{y_i\}_{i\geq 1}$ be a countable collection of not necessarily distinct points of $\mathbb{E}$. A point measure on $\mathbb{E}$ is a measure m of the following form:

$$m = \sum_{i=1}^{\infty} \delta_{y_i},$$

where δ_{y_i} denotes the Dirac measure at y_i, i.e., $\delta_{y_i}(A) = 1$ if $y_i \in A$ and 0 otherwise for a set $A \in \mathscr{E}$. Note that m is a Radon measure. Take $M_p(\mathbb{E})$ as the space of all point measures defined on $\mathbb{E}$ and $\mathcal{M}_p(\mathbb{E})$ be the smallest σ-algebra containing all sets of the form $\{m \in M_p(\mathbb{E}) : m(A) \in B\}$ for $A \in \mathscr{E}, B \in \mathscr{B}$ where $\mathscr{B}$ is the Borel σ-algebra generated by the open sets in $\mathbb{R}_+$. Alternatively, $\mathcal{M}_p(\mathbb{E})$ is the smallest σ-algebra making all evaluation maps $m \to m(A)$ measurable for all $A \in \mathscr{E}$. A point process ξ on $\mathbb{E}$ is a measurable map from the probability space $(\Omega, \mathcal{F}, P) \to (M_p(\mathbb{E}), \mathcal{M}_p(\mathbb{E}))$. Therefore, a point process is a random element of $M_p(\mathbb{E})$.

The Laplace functional is a useful tool for determining the distribution of point processes. Notably, the Laplace functional of a random measure uniquely characterizes the distribution of that measure.

Definition 2.1.4 (**Laplace functional of a point process**). Let Q be a probability measure on $(M_p(\mathbb{E}), \mathcal{M}_p(\mathbb{E}))$. If $\xi : (\Omega, \mathcal{F}) \to (M_p(\mathbb{E}), \mathcal{M}_p(\mathbb{E}))$ is a point process, then the Laplace functional of ξ for a non-negative measurable function f on $\mathbb{E}$ is defined by

$$\begin{aligned}\Psi_\xi(f) &= E\left(\exp\{-\xi(f)\}\right) = \int_\Omega \exp\{-\xi(\omega, f)\} P(d\omega) \\ &= \int_{M_p(\mathbb{E})} \exp\left\{-\int_{\mathbb{E}} f(x) m(dx)\right\} Q(dm).\end{aligned}$$

Definition 2.1.5 (Poisson random measure). Let Λ be a Radon measure on $\mathscr{E}$. A point process ξ is called a Poisson point process or a Poisson random measure with mean measure Λ, denoted by PRM(Λ), if it satisfies:

1. For $A \in \mathscr{E}$,

$$P(\xi(A) = k) = \begin{cases} \frac{(\Lambda(A))^k e^{-\Lambda(A)}}{k!} & \text{if } \Lambda(A) < \infty, \\ 0 & \text{if } \Lambda(A) = \infty. \end{cases}$$

2. For any $n \geq 1$, if $A_1, \ldots, A_n$ are mutually disjoint sets in $\mathscr{E}$, then $\xi(A_1), \ldots, \xi(A_n)$ are independent random variables.

Therefore, ξ is a Poisson random measure if the random number of points in a set A has a Poisson distribution with parameter $\Lambda(A)$ and the number of points in disjoint sets are independent random variables.

Proposition 2.1.6. Let $\xi \sim$ PRM(Λ). The Laplace functional of PRM(Λ) uniquely determines the law of ξ. It is given for any nonnegative measurable function f by

$$\Psi_\xi(f) = \exp\left\{ -\int_{\mathbb{E}} \left(1 - e^{-f(x)}\right) \Lambda(dx) \right\}. \tag{2.1.1}$$

For proof, see [59, Proposition 3.6].

The following straightforward proposition derives a representation for the Poisson random measure with Lebesgue mean measure. There are various approaches to demonstrate this result. However, given that the recursive technique introduced in [5] proves useful in other similar scenarios, we opt to present it here.

Proposition 2.1.7. Let $\xi \sim$ PRM(λ) where λ is the Lebesgue measure on $[0, \infty)$. Then ξ can be written as follows

$$\xi = \sum_{i=1}^{\infty} \delta_{\Gamma_i}, \tag{2.1.2}$$

where

$$\Gamma_i = E_1 + \cdots + E_i, \tag{2.1.3}$$

and $(E_i)_{i \geq 1}$ is a sequence of independent and identically distributed (i.i.d.) random variables with an exponential distribution of mean 1.

Proof. For any $t \geq 0$, define

$$\xi_t = \sum_{i=1}^{\infty} \delta_{\Gamma_i + t}$$

such that $\xi_0 = \xi$. Now, for any nonnegative function f on $[0, \infty)$,

$$\begin{aligned}
\Psi_{\xi_t}(f) &= E(e^{-\xi_t(f)}) = E\left(e^{-\sum_{i=1}^{\infty} f(\Gamma_i + t)}\right) \\
&= E\left(E\left(e^{-\sum_{i=1}^{\infty} f(\Gamma_i + t)} \mid \Gamma_1 = s\right)\right) \\
&= \int_0^{\infty} e^{-f(s+t)} \Psi_{\xi_{s+t}}(f) e^{-s} \mathrm{d}s.
\end{aligned}$$

Using a change of variable we get

$$e^{-t}\Psi_{\xi_t}(f) = \int_t^{\infty} e^{-f(v)} \Psi_{\xi_v}(f) e^{-v} \mathrm{d}v.$$

Differentiating both sides with respect to t, we have

$$\Psi_{\xi_t}(f) = \exp\left(-\int_t^{\infty} (1 - e^{-f(s)}) \mathrm{d}s\right).$$

Now, take $t = 0$ to get

$$\Psi_{\xi_0}(f) = \exp\left(-\int_0^{\infty} (1 - e^{-f(s)}) \mathrm{d}s\right)$$

which equals (2.1.1) with $\Lambda(\mathrm{d}s) = \lambda(\mathrm{d}s) = \mathrm{d}s$. ∎

Two important transformations of the PRM which we will use frequently in the next sections are presented in the following two propositions. The first proposition shows that mapping the points of a PRM yields a new PRM with a certain representation for its mean measure while the other proposition shows that starting from a PRM on one dimension, we may construct a new PRM whose points live in a higher dimension.

Proposition 2.1.8 (Proposition 3.7 in [59]). Let $\mathbb{E}_i$, $i = 1, 2$, be two locally compact spaces with countable bases and $\mathscr{E}_i$, $i = 1, 2$, be the associated σ-algebras. Suppose $T : (\mathbb{E}_1, \mathscr{E}_1) \to (\mathbb{E}_2, \mathscr{E}_2)$ be measurable. If ξ is a PRM(μ) on $\mathbb{E}_1$, then $\tilde{\xi} = \xi \circ T^{-1}$ is a PRM($\tilde{\mu}$) on $\mathbb{E}_2$ such that $\tilde{\mu} = \mu \circ T^{-1}$. If ξ has the representation $\sum_{i=1}^{\infty} \delta_{Y_i}$ where $(Y_i)_{i \geq 1}$ are random elements on $\mathbb{E}_1$ then $\tilde{\xi} = \sum_{i=1}^{\infty} \delta_{T(Y_i)}$.

Proposition 2.1.9 (Proposition 3.8 in [59]). Let $\mathbb{E}_i$, $i = 1, 2$, be two locally compact spaces with countable bases and $\sum_{i=1}^{\infty} \delta_{Y_i}$ be a PRM(μ) on $\mathbb{E}_1$ and $(U_i)_{i\geq 1}$ are i.i.d. random elements on $\mathbb{E}_2$ with common probability distribution F. Assume $(U_i)_{i\geq 1}$ are independent from $(Y_i)_{i\geq 1}$. Then the point process $\sum_{i=1}^{\infty} \delta_{(Y_i,U_i)}$ on $\mathbb{E}_1 \times \mathbb{E}_2$ is a PRM with mean measure $\mu \times F$.

Applying Proposition 2.1.8 and 2.1.9 on PRM(λ) defined in (2.1.2), we can derive useful PRMs which in turn lead to other processes with applications in nonparametric Bayesian inference. Take $T(x) = L^{-1}(x)$ where L (the Lévy measure) is a Borel measure defined on $(0, \infty)$ by $L(x) = L((x, \infty)) = \int_x^{\infty} \mathrm{d}L(u)$ and for each $\epsilon > 0$,

$$\int_{\epsilon}^{\infty} L^{-1}(u)\mathrm{d}u < \infty, \tag{2.1.4}$$

in which

$$L^{-1}(y) = \inf\{x > 0 : L(x) \geq y\}.$$

Also, let $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random elements in a Polish space $\mathbb{E}$, drawn from a probability measure H, independent of $(\Gamma_i)_{i\geq 1}$. Obviously,

$$\sum_{i=1}^{\infty} \delta_{L^{-1}(\Gamma_i)} \sim \text{PRM}(L), \tag{2.1.5}$$

$$\sum_{i=1}^{\infty} \delta_{(Z_i,L^{-1}(\Gamma_i))} \sim \text{PRM}(H \times L). \tag{2.1.6}$$

For example, $\sum_{i=1}^{\infty} \delta_{\Gamma_i^{-1/\alpha}}$ follows a PRM(L) with

$$L(x) = x^{-\alpha} = \int_x^{\infty} \alpha u^{-\alpha-1}\mathrm{d}u, \; x > 0, \; \alpha \in (0, 1). \tag{2.1.7}$$

In fact, (2.1.7) denotes the Lévy measure of the positive α-stable random variable $S_\alpha = \sum_{i=1}^{\infty} \Gamma_i^{-1/\alpha}$. Notice that since $\Gamma_i/i \xrightarrow{a.s.} 1$, S_α converges for $\alpha \in (0, 1)$, almost surely. In this thesis, $\xrightarrow{d}$, $\xrightarrow{p}$, and $\xrightarrow{a.s.}$ denote convergence in distribution, convergence in probability, and almost sure convergence, respectively.

As another example, we can set L as the gamma Lévy measure to construct both the gamma process and the Dirichlet process, which we will explore in the next section.

2.2 Bayesian nonparametric priors

In this section, we review some properties of Bayesian nonparametric priors, with special emphasis on the Dirichlet process and the two-parameter Poisson-Dirichlet process.

2.2.1 The Dirichlet process

The Dirichlet process, formally introduced in [18], is considered the cornerstone of Bayesian nonparametric inference. It is a prior law over the space of probability distribution functions, whose finite-dimensional marginals have a Dirichlet distribution. We begin by recalling the definition of the Dirichlet distribution.

Definition 2.2.1 (**Dirichlet Distribution**). The vector of proportions $(p_1, \ldots, p_{n+1})$ follows a Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_{n+1})$, denoted by $(p_1, \ldots, p_{n+1}) \sim \mathrm{D}(\alpha_1, \ldots, \alpha_{n+1})$ if the n-dimensional distribution of $(p_1, \ldots, p_n)$ is absolutely continuous with respect to the Lebesgue measure, with density

$$f(p_1, \ldots, p_n) = \frac{\Gamma\left(\sum_{i=1}^{n+1} \alpha_i\right)}{\prod_{i=1}^{n+1} \Gamma(\alpha_i)} \prod_{i=1}^{n} p_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n} p_i\right)^{\alpha_{n+1}-1} I_{\mathbb{S}_n}(p_1, \ldots, p_n), \tag{2.2.1}$$

where $\alpha_i > 0$ for all $i = 1, \ldots, n+1$, and $\mathbb{S}_n$ is the n-simplex

$$\mathbb{S}_n = \left\{ (p_1, \ldots, p_n) : p_i \geq 0, \sum_{i=1}^{n} p_i \leq 1 \right\}.$$

In the above definition, $\sum_{i=1}^{n+1} p_i = 1$, and

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t, \quad x > 0$$

is the gamma function.

Following [18], the Dirichlet process is defined as follows:

Definition 2.2.2 (**Dirichlet Process**). Let $(\mathbb{E}, \mathscr{E})$ be an arbitrary measurable space and H be a probability measure on $(\mathbb{E}, \mathscr{E})$. Let $\theta > 0$ be arbitrary. A random probability measure $P = \{P(A)\}_{A \in \mathscr{E}}$ is called a Dirichlet process on $(\mathbb{E}, \mathscr{E})$ with parameters θ and H, if for any finite measurable partition $\{A_1, \dots, A_k\}$ of $\mathbb{E}$, the joint distribution of the vector $(P(A_1), \dots, P(A_k))$ has the Dirichlet distribution with parameters $(\theta H(A_1), \dots, \theta H(A_k))$, where $k \geq 2$. We assume that if $H(A) = 0$, then $P(A) = 0$ with probability one. If P is a Dirichlet process with parameters a and H, we write $P \sim DP(\theta, H)$.

Throughout this thesis, we use the same notation for the probability measure and its corresponding cumulative distribution function, i.e., $P(t) = P((-\infty, t])$ and $H(t) = H((-\infty, t])$ when $\mathbb{E} = \mathbb{R}$.

For any $A \in \mathscr{E}$, $P(A)$ has a beta distribution with parameters $\theta H(A)$ and $\theta(1 - H(A))$. Thus,

$$E(P(A)) = H(A) \quad \text{and} \quad \mathrm{Var}(P(A)) = \frac{H(A)(1 - H(A))}{1 + \theta}. \tag{2.2.2}$$

The probability measure H is called the base measure of P. Clearly, from (2.2.2), H plays the role of the center of the process, while θ can be viewed as the concentration parameter. The larger θ is, the more likely it is that the realization of P is close to H. Specifically, for any fixed set $A \in \mathscr{E}$ and $\epsilon > 0$, we have $P(A) \to H(A)$ as $\theta \to \infty$ since

$$\Pr\{|P(A) - H(A)| > \epsilon\} \leq \frac{H(A)(1 - H(A))}{\epsilon^2(1 + \theta)}.$$

The use of the Dirichlet process requires finding its posterior distribution, i.e., the conditional distribution of the Dirichlet process given the sample. At first, we introduce a notion of a sample from a random probability distribution.

Definition 2.2.3 (**Sample from a random probability measure**). Let P be a random probability measure on $(\mathbb{E}, \mathscr{E})$. We say that $X_1, \dots, X_m$ is a sample of size m from P if for any $n = 1, 2, \dots$ and sets $B_1, \dots, B_n, C_1, \dots, C_m \in \mathscr{E}$;

$$\Pr\{X_1 \in C_1, \dots, X_m \in C_m \mid P(B_1), \dots, P(B_n), P(C_1), \dots, P(C_m)\} = \prod_{i=1}^{m} P(C_i)$$

almost surely. See [18].

That is, $X_1, \dots, X_m$ is a sample of size m from P, if, given $P(C_1), \dots, P(C_m)$, the events $\{X_1 \in C_1\}, \dots, \{X_m \in C_m\}$ are independent from the rest of the process and independent among themselves, with $\Pr\{X_i \in C_i \mid P(C_1), \dots, P(C_m)\} = P(C_i)$ a.s. for $i = 1, \dots, m$.

The following theorem outlines the posterior distribution of the Dirichlet process given the observed data. It shows that the Dirichlet process has the conjugacy property. That is, the posterior distribution given the data is again a Dirichlet process. For the proof, see [18, Theorem 1]. In the theorem and throughout the thesis, we use a "*" as a superscript to denote posterior quantities.

Theorem 2.2.4. If $X_1, \ldots, X_n$ is a sample from $P \sim DP(\theta, H)$, then the posterior distribution of P given $X_1, \ldots, X_n$ coincides with the distribution of the Dirichlet process with parameters θ^* and H^*, where

$$\theta^* = \theta + n, \quad H^* = \frac{\theta}{\theta + n} H + \frac{n}{\theta + n} \frac{\sum_{i=1}^n \delta_{X_i}}{n}. \tag{2.2.3}$$

Notice that the posterior base distribution H^* is a convex combination of the base distribution and the empirical distribution. The weight associated with the prior base distribution H is proportional to θ, giving another reason to call θ the concentration parameter. The weight associated with the empirical distribution is proportional to the number of observations n. The posterior base distribution H^* approaches the prior base measure H for large values of θ. On the other hand, for small values of θ, H^* is close to the empirical distribution.

A very useful approximation of the Dirichlet process can be obtained from certain finite mixture models. In particular, [34] proved the next result.

Theorem 2.2.5. For any $N \geq 1$, consider the finite-dimensional Dirichlet prior

$$P_N^{F.D.}(\cdot) = \sum_{i=1}^N p_{i,N} \delta_{Z_i}(\cdot), \tag{2.2.4}$$

where $(Z_i)_{i \geq 1}$ is a sequence of i.i.d. random variables with common distribution H, and $(p_{1,N}, \ldots, p_{N,N})$ has the Dirichlet distribution with parameter $(\theta/N, \ldots, \theta/N)$, independent of $(Z_i)_{i \geq 1}$. Let P be a Dirichlet process with parameters θ and H. Then

$$P_N^{F.D.}(g) := \int g(x) P_N^{F.D.}(\mathrm{d}x) \xrightarrow{d} P(g) := \int g(x) P(\mathrm{d}x),$$

as $N \to \infty$, for any measurable function $g : \mathbb{R} \to \mathbb{R}$ with $\int_{\mathbb{R}} |g(x)| H(\mathrm{d}x) < \infty$. In particular, $(P_N^{F.D.})_{N \geq 1}$ converges in distribution to P, where $P_N^{F.D.}$ and P are random variables with values in the space $M_1(\mathbb{R})$ of probability measures on $\mathbb{R}$, endowed with the topology of weak convergence.

To generate the sequence $(p_{1,N}, \ldots, p_{N,N})$, one can define $p_{i,N}$ as

$$p_{i,N} = \frac{G_{i,N}}{\sum_{i=1}^{N} G_{i,N}},$$

where $(G_{i,N})_{i\geq 1}$ is a sequence of i.i.d. random variables with a Gamma$(\theta/N, 1)$ distribution, independent of $(Z_i)_{i\geq 1}$.

Ferguson (1973) proposed a series representation in [18] as an alternative definition of the Dirichlet process. This representation is based on the earlier work of Ferguson and Klass (1972) [20], in which they provided a sum representation for processes with independent increments (and no Gaussian part) based on the arrival times of a homogeneous Poisson process. The representation is described in the following theorem. For the proof, see [20, 18, 5].

Theorem 2.2.6. Let $(\Gamma_i)_{i\geq 1}$ be defined as in (2.1.3), and let $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with a common distribution H, independent of $(\Gamma_i)_{i\geq 1}$. Let

$$P^{Ferg.}(\cdot) = \sum_{i=1}^{\infty} \frac{L^{-1}(\Gamma_i)}{\sum_{j=1}^{\infty} L^{-1}(\Gamma_j)}\, \delta_{Z_i}(\cdot), \tag{2.2.5}$$

where

$$L(x) = \theta \int_x^{\infty} t^{-1} e^{-t}\, \mathrm{d}t, \quad x > 0. \tag{2.2.6}$$

Then $P^{Ferg.}$ is a Dirichlet process with parameters θ and H.

Note that $L(x) = L([x, \infty))$, where

$$L(\mathrm{d}x) = \theta x^{-1} e^{-x}\, \mathrm{d}x$$

is the (tail) Lévy measure of a random variable Y with a Gamma$(\theta, 1)$ distribution, i.e.,

$$E(e^{iuY}) = (1 - iu)^{-\theta} = \exp\left\{\theta \int_0^{\infty} (e^{iux} - 1)\, L(\mathrm{d}x)\right\}, \quad u \in \mathbb{R}.$$

From Theorem 2.2.6, it follows clearly that a realization of the Dirichlet process must necessarily be a discrete probability measure, even when the base measure is absolutely continuous. This fact was noted by [18], and [6]. It is worth mentioning that this discreteness property is no more troublesome than the discreteness of the empirical process. In spite of the discreteness property, the support of the Dirichlet process is very large. For more details, see Proposition 3 of [18].

Remark 2.2.7. The series $\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)\, \delta_{Z_i}(\cdot)$ is the Ferguson and Klass representation of the gamma process. Thus, the Dirichlet process can be viewed as a normalized gamma process.

Sampling from the Dirichlet process based in (2.2.5) is difficult in practice since there is no closed form for the inverse of the gamma Lévy measure (2.2.6). Moreover, to determine the random weights in (2.2.5), an infinite sum must be computed.

The Bondesson (1982) sum representation of the Dirichlet process with parameters θ and H is defined in the next theorem.

Theorem 2.2.8. Let $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with common distribution H and let $(E_i')_{i\geq 1}$ be a sequence of i.i.d. exponential random variables with mean 1, independent of $(\Gamma_i)_{i\geq 1}$ and $(Z_i)_{i\geq 1}$. Then,

$$P^{\text{Bond.}}(\cdot) = \sum_{i=1}^{\infty} \frac{e^{-\Gamma_i/\theta} E_i'}{\sum_{i=1}^{\infty} e^{-\Gamma_i/\theta} E_i'} \delta_{Z_i}(\cdot). \tag{2.2.7}$$

is a Dirichlet process with parameters θ and H.

Note that the Bondesson's representation overcomes the problem of inverting the gamma Lévy measure. However, the infinite number of terms to compute in (2.2.7) make it difficult to sample from the Dirichlet process. As an alternative, one can approximate the Dirichlet process by using a truncation

$$P_N^{\text{Bond.}}(\cdot) = \sum_{i=1}^{N} \frac{e^{-\Gamma_i/\theta} E_i'}{\sum_{i=1}^{N} e^{-\Gamma_i/\theta} E_i'} \delta_{Z_i}(\cdot).$$

For a given tolerance value $\epsilon \in (0,1)$, a truncation value $N = N(\epsilon)$ can be selected by

$$N = \inf\left\{ j : \frac{e^{-\Gamma_j/\theta} E_j'}{\sum_{i=1}^{j} e^{-\Gamma_i/\theta} E_i'} < \epsilon \right\}.$$

The Ferguson's and Bondesson's sum representation for the Dirichlet process are based on the normalized Gamma process. A radically different (constructive) definition of the Dirichlet process is given by Sethuraman (1994) in [63] using a "stick-breaking" approach. This representation is stated in the next theorem.

Theorem 2.2.9. Let $(\beta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with a Beta$(1,\theta)$ distribution. Define

$$p_1 = \beta_1, \quad p_i = \beta_i \prod_{k=1}^{i-1}(1-\beta_k), \quad i \geq 2.$$

Moreover, let $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with common distribution H, independent of $(\beta_i)_{i\geq 1}$. Define

$$P^{Seth.}(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{Z_i}(\cdot). \tag{2.2.8}$$

Then $P^{Seth.}$ is a Dirichlet process with parameters θ and H.

Sethuraman's stick-breaking representation can be used to approximately simulate the Dirichlet process using a truncation argument. By truncating the higher order terms in the sum we can approximate the Sethuraman stick-breaking representation by

$$P_N^{Seth.}(\cdot) = \sum_{i=1}^{N} p_i \delta_{Z_i}(\cdot),$$

where $(\beta_i)_{i\geq 1}, (p_i)_{i\geq 1}$, and $(Z_i)_{i\geq 1}$ are as defined in (2.2.8) with $\beta_N = 1$ (hence β_N does not have a beta distribution). The assumption that $\beta_N = 1$ is necessary to make the weights add to 1, almost surely. A random stopping rule for choosing $N = N(\epsilon)$ was proposed in [51] where, for $\epsilon \in (0,1)$,

$$N = \inf\left\{ i : p_i = \beta_i \prod_{k=1}^{i-1}(1-\beta_k) < \epsilon \right\}.$$

It is demonstrated in [67] that the weights in both Bondesson's representation and the Sethurman sum representation are not strictly decreasing almost surely. This makes the truncated stick-breaking representation and Bondesson's representation inefficient for simulation purposes. More precisely, it can be proved that, for $i \geq 1$,

$$\Pr\left\{ \frac{e^{-\Gamma_{i+1}/\theta}E'_{i+1}}{\sum_{i=1}^{\infty} e^{-\Gamma_i/\theta}E'_i} < \frac{e^{-\Gamma_i/\theta}E'_i}{\sum_{i=1}^{\infty} e^{-\Gamma_i/\theta}E'_i} \right\} = \sum_{k=0}^{\infty}(-1)^k \frac{\theta}{k+\theta},$$

and

$$\Pr\{p_{i+1} < p_i\} = \sum_{k=0}^{\infty}(-1)^k \frac{\theta}{k+\theta}.$$

In [67], Zarepour and Al Labadi derive a finite sum representation of the Dirichlet process, which almost surely converges to Ferguson's representation (2.2.5). Due to the monotonically decreasing nature of the weights in this new representation, it provides a highly accurate approximation of the Dirichlet process. The details of this representation are presented in the following theorem.

Theorem 2.2.10. Let Y_n be a random variable with distribution Gamma$(\theta/n, 1)$. Define

$$G_n(x) = \Pr(Y_n > x) = \int_x^\infty \frac{1}{\Gamma(\theta/n)} e^{-t} t^{\theta/n-1} \mathrm{d}t$$

and

$$G_n^{-1}(y) = \inf\{x : G_n(x) \geq y\}, \quad 0 < y < 1.$$

Let L be the gamma Lévy measure (2.2.6) and $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with values in $\mathbb{E}$ and common distribution H, independent of $(\Gamma_i)_{i\geq 1}$. Then, as $n \to \infty$

$$P_n^{Zar\&Al-Lab} = \sum_{i=1}^{n} \frac{G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)}{\sum_{i=1}^{n} G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)} \delta_{Z_i} \xrightarrow{a.s.} P^{Ferg.} \tag{2.2.9}$$

on $\mathbb{E}$ with respect to the weak topology.

2.2.2 The two-parameter Poisson-Dirichlet process

An important extension of the Dirichlet process is the two-parameter Poisson-Dirichlet process developed by Pitman and Yor (1997) [56]. We start by recalling the Pitman and Yor (1997) stick-breaking definition of the two-parameter Poisson-Dirichlet process on an arbitrary measurable space $(\mathbb{E}, \mathscr{E})$.

Definition 2.2.11. Let $0 \leq \alpha < 1$, $\theta > -\alpha$, and $(\beta_i)_{i\geq 1}$ be a sequence of independent random variables with a Beta$(1-\alpha, \theta + i\alpha)$ distribution. Define

$$p_1' = \beta_1, \quad p_i' = \beta_i \prod_{j=1}^{i-1}(1-\beta_j), \quad i \geq 2.$$

Let $p_1 \geq p_2 \geq \ldots$ be the ranked values of $(p_i')_{i\geq 1}$. Moreover, let $(Z_i)_{i\geq 1}$ be i.i.d. random variables with values in $\mathbb{E}$ and common distribution H and independent of p_i''s. Then the random probability measure

$$P_{\alpha,\theta,H}(\cdot) \stackrel{d}{=} \sum_{i=1}^{\infty} p_i \delta_{Z_i}(\cdot) \tag{2.2.10}$$

is called two-parameter Poisson-Dirichlet process with parameters α, θ, H, and denoted by $\mathrm{PDP}(H; \alpha, \theta)$.

It is worth mentioning that [31] referred to the process

$$\tilde{P}_{\alpha,\theta,H}(\cdot) = \sum_{i=1}^{\infty} p'_i \delta_{Z_i}(\cdot)$$

as the Pitman-Yor process. Note that $(p'_1, p'_2, \ldots) \sim \mathrm{GEM}(\alpha, \theta)$ with the notation used in [9]. The law of the weights $(p_1, p_2, \ldots)$ is called the *two-parameter Poisson-Dirichlet distribution*, denoted by $\mathrm{PD}(\alpha, \theta)$. The two-parameter Poisson-Dirichlet distribution has many applications in different fields such as population genetics, ecology, statistical physics, and number theory. See [17] for more details. On the other hand, the two-parameter Poisson-Dirichlet process has been recently used in applications in Bayesian nonparametric statistics, such as computer science [65], species sampling [39], and genomics [16].

Note that for the special case when $\alpha = 0$, $(\beta_i)_{i \geq 1}$ become a sequence of independent random variables with $\mathrm{Beta}(1, \theta)$ distribution. Thus, the two-parameter Poisson-Dirichlet process $P_{0,\theta,H}$ becomes simply the Dirichlet process. On the other hand, when $\theta = 0$, [56] show that the two-parameter Poisson-Dirichlet process $P_{\alpha,0,H}$ becomes the normalized non-negative Stable law process with index $\alpha \in (0, 1)$.

Note that for any measurable subset A of $\mathbb{E}$, the two-parameter Poisson-Dirichlet process has the following properties:

$$E\{P_{\alpha,\theta,H}(A)\} = H(A),$$

$$\mathrm{Var}\{P_{\alpha,\theta,H}(A)\} = H(A)(1 - H(A)) \frac{1-\alpha}{1+\theta}.$$

For more details on the calculation of the moments for the two-parameter Poisson-Dirichlet process, interested readers can refer to [9].

Similar to the Dirichlet process, the base measure H plays the role of the center of the process. However, both α and θ govern the variability of $P_{\alpha,\theta,H}$ around its base measure H.

The next theorem derives the posterior distribution of $P_{\alpha,\theta,H}$ given the data set. Note that unlike the Dirichlet process, the two-parameter Dirichlet process does not have the conjugacy property.

Theorem 2.2.12. Let $X_1, \ldots, X_n$ be a sample from $P_{\alpha,\theta,H}$. Let k be the number of unique values within n observations, and $X'_1, \ldots, X'_k$ denote the distinct values. Also, let n_j be the number of X_i equal to X'_j. Then

$$(P_{\alpha,\theta,H} \mid X_1, \ldots, X_n) \stackrel{d}{=} \sum_{j=1}^{k} W_j \delta_{X'_j} + W_{k+1} P_{\alpha,\theta+k\alpha,H},$$

where $P_{\alpha,\theta+k\alpha,H} \sim \mathrm{PDP}(H, \alpha, \theta + k\alpha)$, $(W_1, \ldots, W_{k+1}) \sim \mathrm{D}(n_1 - \alpha, \ldots, n_k - \alpha, \theta + k\alpha)$, and $P_{\alpha,\theta+k\alpha,H}$ and $(W_1, \ldots, W_{k+1})$ are conditionally independent given $X_1, \ldots, X_n$.

Remark 2.2.13. In the above theorem, if we take $\alpha = 0$, then

$$(P_{0,\theta,H} \mid X_1, \ldots, X_n) \sim DP(\theta^*, H^*)$$

where θ^* and H^* are defined in (2.2.3).

Remark 2.2.14. In [53], it is shown that the finite dimensional Dirichlet process $P_N^{F.D.}$, defined in (2.2.4), is a two-parameter Poisson-Dirichlet process with parameters $-\theta/N$ and θ, i.e. $P_N^{F.D.} \sim \mathrm{PDP}(H; -\theta/N, \theta)$. Consequently, if $X_1, \ldots, X_n$ is a sample from $P^{F.D.}$, then

$$(P_N^{F.D.} \mid X_1, \ldots, X_n) \stackrel{d}{=} \sum_{j=1}^{k} W_j \delta_{X'_j} + W_{k+1} P_{-\theta/N,\theta-k\theta/N,H},$$

where $P_{-\theta/N,\theta-k\theta/N,H} \sim \mathrm{PDP}(H, -\theta/N, \theta - k\theta/N)$, $(W_1, \ldots, W_{k+1}) \sim \mathrm{D}(n_1 + \theta/N, \ldots, n_k + \theta/N, \theta - k\theta/N)$, and $P_{-\theta/N,\theta-k\theta/N,H}$ and $(W_1, \ldots, W_{k+1})$ are conditionally independent given $X_1, \ldots, X_n$. If $N \to \infty$, then

$$(P_N^{F.D.} \mid X_1, \ldots, X_n) \sim DP(\theta^*, H^*)$$

where θ^* and H^* are defined in (2.2.3).

The infinite sum in (2.2.10) makes it difficult in practice to draw a sample from the Poisson-Dirichlet process. An approximation of the two-parameter Poisson-Dirichlet process can be made by truncating the higher order terms in the sum (2.2.10). For more details, see [3]. Let $(\beta_i)_{i\geq 1}$ and $(p_i)_{i\geq 1}$ be as defined earlier, with $\beta_N = 1$. The random probability measure

$$P_{\alpha,\theta,H,N}(\cdot) = \sum_{i=1}^{N} p_i \delta_{Z_i}(\cdot) \tag{2.2.11}$$

is the finite approximation of the two-parameter Poisson-Dirichlet process. A random stopping rule for choosing $N = N(\epsilon)$, where $\epsilon \in (0, 1)$, is:

$$N = \inf \left\{ i : p_i' = (1-\beta_1)\cdots(1-\beta_{i-1})\beta_i < \epsilon \right\}.$$

According to [3, Lemma 2.1], the weights $(p_i)_{i\geq 1}$ in (2.2.11), prior to being ordered, are not strictly decreasing with probability one. This suggests that the truncated representation (2.2.11) may not serve as an efficient approximation to the two-parameter Poisson-Dirichlet process.

A different approach to construct the two-parameter Poisson-Dirichlet process is proposed in [56, Proposition 22]. This is described in the next proposition.

Proposition 2.2.15. For $0 < \alpha < 1$ and $\theta > 0$, suppose that $(p_1(0, \theta), p_2(0, \theta), \ldots)$ and $(p_1(\alpha, 0), p_2(\alpha, 0), \ldots)$ have respective distributions $PD(0, \theta)$ and $PD(\alpha, 0)$. Independent of $(p_1(0, \theta), p_2(0, \theta), \ldots)$, let $\{p_1^i(\alpha, 0), p_2^i(\alpha, 0), \ldots\}, i = 1, 2, \ldots,$ be a sequence of independent copies of $(p_1(\alpha, 0), p_2(\alpha, 0), \ldots)$. Let $(p_i)_{i\geq 1}$ be the descending order statistics of $\{p_i(0, \theta)p_j^i(\alpha, 0), i, j = 1, 2, \ldots\}$. Then $(p_1, p_2, \ldots)$ has a $PD(\alpha, \theta)$ distribution.

Note that in Proposition 2.2.15, the weights of the two-parameter Poisson-Dirichlet process are constructed based on two different choices of parameters. One with $\alpha = 0$ which corresponds to the Dirichlet process $P_{0,\theta,H}$ and another when $\theta = 0$ which corresponds to the normalized Stable law process $P_{\alpha,0,H}$. The index of the Stable law process α is in $(0, 1)$. Therefore, approximating the two-parameter Poisson-Dirichlet process necessitates simulating both the Dirichlet process and the normalized stable law process.

In [56, Proposition 10], it is shown that the sum representation of the normalized Stable law process can be written as follows

$$P_{\alpha,0,H}(\cdot) = \sum_{i=1}^{\infty} \frac{\Gamma_i^{-1/\alpha}}{\sum_{i=1}^{\infty} \Gamma_i^{-1/\alpha}} \delta_{Z_i}(\cdot).$$

Therefore, the approximation of the normalized Stable law process is

$$P_{\alpha,0,H,N}(\cdot) = \sum_{i=1}^{N} \frac{\Gamma_i^{-1/\alpha}}{\sum_{i=1}^{N} \Gamma_i^{-1/\alpha}} \delta_{Z_i}(\cdot), \tag{2.2.12}$$

for large enough N. Note that the weights $\left(\frac{\Gamma_i^{-1/\alpha}}{\sum_{k=1}^{N} \Gamma_k^{-1/\alpha}} \right)_{1 \leq i \leq N}$ are strictly decreasing. In Algorithm A of [3], Al Labadi and Zarepour propose an approximation of the two-parameter Poisson-Dirichlet process by first approximating the Dirichlet process using (2.2.9), followed by approximating the normalized Stable law process with (2.2.12). They compare the accuracy of their approximation with the stick-breaking approximation given in (2.2.11). Through simulations, they demonstrate that their approach yields more precise results than the stick-breaking approximation in (2.2.11).

2.2.3 Lévy processes

Definition 2.2.16 ([4])**.** A *cadlag* [1] (i.e., right-continuous with left limits a.s.) stochastic process $\{X(t)\}_{t \geq 0}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a Lévy process if it satisfies the following conditions:

1. $X(t)$ starts at 0, i.e., $X(0) = 0$ a.s.

2. $X(t)$ has independent increments, i.e., the random variables $X(t_0), X(t_1) - X(t_0), \ldots, X(t_n) - X(t_{n-1})$ are independent for all $n \geq 1$ and $0 \leq t_0 < t_1 < \ldots < t_n$.

3. $X(t)$ has stationary increments, i.e., $X(t+s) - X(t) \stackrel{d}{=} X(s)$ for all $s, t \geq 0$.

4. $X(t)$ is stochastically continuous, i.e., for all $\epsilon > 0$ and for all $s \geq 0$,

$$\lim_{t \to s} \Pr\{|X(t) - X(s)| > \epsilon\} = 0,$$

   or equivalently $X(t) \stackrel{p}{\to} X(s)$ as $t \to s$.

[1]Some authors do not impose the cadlag property in the definition of a Lévy process. In fact, every Lévy process (defined without he cadlag property) has a unique modification that is cadlag [62, Theorem 30]. Therefore, the cadlag property can be assumed without loss of generality ([11, Definition 3.1]).

Note that, one may conclude (1) from (3) by setting $s = 0$. Condition (4) does not imply that the sample paths are continuous. The Brownian motion, the Poisson process, and the gamma process are some examples of Lévy processes. A Lévy process $X(t)$ is called Brownian motion if it has continuous sample paths and $X(t) \sim N(0, t)$. It is called a Poisson process of intensity $\lambda > 0$ if $X(t) \sim \text{Poisson}(\lambda t)$. It is called a gamma process with parameters $a, b > 0$ if $X(t) \sim \text{Gamma}(at, b)$.

The Lévy-Khintchine representation theorem states that a Lévy process is the sum of a deterministic function, a Brownian motion, and a pure jump part ([4]). Its characteristic function is given by

$$E(e^{iuX(t)}) = \exp\left\{ i\gamma_t u - \frac{1}{2}\sigma_t^2 u^2 + \int_{\mathbb{R}\setminus\{0\}} \left(e^{ius} - 1 - ius\mathbf{1}_{\{|s|\leq 1\}}\right) L_t(\mathrm{d}s) \right\}, \quad u \in \mathbb{R},$$

where $(\gamma_t, \sigma_t^2, L_t)$ is called the characterizing triplet. The Lévy measure L_t satisfies $L_t(\{0\}) = 0$, $\int_{\mathbb{R}} s^2/(1+s^2) L_t(\mathrm{d}s) < \infty$ or equivalently $\int_0^1 s^2 L_t(\mathrm{d}s) < \infty$ and $\int_1^\infty L_t(\mathrm{d}s) < \infty$.

Lévy processes have been used in Bayesian nonparametric inference by [13, 18, 19, 21, 26] with the following additional properties (which in this case, $X(t)$ is called a subordinator):

5. $X(t)$ is defined on $\mathbb{R}^+$
6. $X(t)$ is nonnegative: $X(t) \geq 0$ a.s.
7. $X(t)$ is nondecreasing a.s.

For a comprehensive review of applications of the Lévy processes in Bayesian nonparametric inference, refer to [2, Chapter 2 and 3].

Chapter 3

Random discrete probability measures based on a negative binomial process

A distinctive functional of Poisson point process is the negative binomial process (NBP) for which the increments are not independent, but are independent conditional on an underlying gamma variable. In this chapter, using a new point process representation for the NBP, we generalize the Poisson-Kingman distribution and its corresponding random discrete probability measure. This new proposed family of the discrete random probability measures which is defined by normalizing the points of the NBP provides a new set of useful priors for Bayesian nonparametric models with more flexibility than the random discrete probability measure which are obtained by normalizing the points of a Poisson point process. We illustrate how this family of random discrete probability measures contains the nonparametric Bayesian priors such as the DP, the normalized positive α-stable process, Poisson-Dirichlet process (PDP), and others. With the same gamma Lévy measure, we derive an extension of the DP and its almost sure approximation. Using our representation for the NBP, we develop a new series representation for the PDP. By applying a stopping rule, we truncate this new series representation to obtain an approximation of the original series. We provide a simulation study to compare the efficiency of this approximation with other approximation of this process that exists in the literature, particularly, [3].

3.1 Introduction

The Dirichlet process defined in (2.2.5) can be viewed as a functional of the Poisson random measure (PRM) given in (2.1.6). Another important distribution which results from a PRM is the Poisson-Kingman distribution. Consult [43] and [54] for properties and applications of this distribution. The vector of the normalized points of the PRM defined in equation (2.1.5) will follow a Poisson-Kingman distribution denoted by PK(L), i.e.,

$$\left(\frac{L^{-1}(\Gamma_1)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)}, \frac{L^{-1}(\Gamma_2)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)}, \dots\right) \sim \mathrm{PK}(L). \tag{3.1.1}$$

Equation (3.1.1) defines a random discrete distribution on the infinite dimensional simplex $\mathbb{S}_\infty := \{(x_1, x_2, \dots) : x_i \geq 0, i = 1, 2, \dots, \sum_{i=1}^{\infty} x_i = 1\}$.

Equivalently, a Poisson-Kingman distribution can be constructed using subordinators. Let $(X_t)_{t\geq 0}$ be a subordinator with Lévy measure L and write $(\Delta X_t := X_t - X_{t-})_{t>0}$ for the jump process of X_t, and $\Delta X_t^{(1)} \geq \Delta X_t^{(2)} \geq \cdots$ for the ordered jumps occurring prior to time $t > 0$. Then

$$\left(\frac{\Delta X_t^{(1)}}{X_t}, \frac{\Delta X_t^{(2)}}{X_t}, \dots\right) \sim \mathrm{PK}(tL). \tag{3.1.2}$$

We see that the weights of the random discrete probability measure identified in expression (2.2.5) are simply the Poisson-Kingman sequence specified in expression (3.1.1). In the rest of this chapter, we will generalize the Poisson-Kingman distribution and the resulting random discrete probability measure by utilizing a point process representation for the negative binomial process. The point process representation for the negative binomial process which we use here is itself constructed directly from a PRM, unlike the representation in [27], where they use a trimmed subordinator.

3.2 The negative binomial process

In this section, we present a novel point process representation for the NBP, derived directly from a PRM. Later, we use this representation to define a more general form of the Poisson-Kingman distribution and its corresponding random discrete probability measure. First we note that for any constant $c > 0$, a simple use of Proposition 2.1.8 shows that the process $\sum_{i=1}^{\infty} \delta_{\Gamma_i + c}$ is a PRM(λ) on $\mathbb{E} = [c, \infty)$, and $\sum_{i=1}^{\infty} \delta_{\Gamma_i / c}$ follows PRM($c\lambda$) on $\mathbb{E} = [0, \infty)$. Now, for any positive integer r consider the random measure

$$\eta = \sum_{i=1}^{\infty} \delta_{\Gamma_{r+i}/\Gamma_r}.$$

First note that conditional on $\{\Gamma_r = u\}$, the process η follows PRM($u\lambda$) on $\mathbb{E} = (1, \infty)$. So the Laplace functional of η is

$$\begin{aligned}
E(e^{-\eta(f)}) &= E\left[E(e^{-\eta(f)} \mid \Gamma_r = u)\right] \\
&= \int_0^{\infty} E(e^{-\eta(f)} \mid \Gamma_r = u)\, P(\Gamma_r \in \mathrm{d}u) \\
&= \int_0^{\infty} \exp\left\{-\int_1^{\infty} (1 - e^{-f(x)})\, u\, \lambda(\mathrm{d}x)\right\} P(\Gamma_r \in \mathrm{d}u) \\
&= \int_0^{\infty} \exp\left\{-u \int_1^{\infty} (1 - e^{-f(x)})\, \lambda(\mathrm{d}x)\right\} \frac{u^{r-1} e^{-u}}{\Gamma(r)}\, \mathrm{d}u \\
&= \left(1 + \int_1^{\infty} (1 - e^{-f(x)})\, \lambda(\mathrm{d}x)\right)^{-r}.
\end{aligned}$$

Indeed, this represents the Laplace functional of the NBP as defined by [23]. We denote this process by NBP(r, λ) and write $\eta \sim$ NBP(r, λ) on the space $\mathbb{E} = (1, \infty)$. The preceding result leads to the following theorem.

Theorem 3.2.1. With a decreasing bijection $L : (0, \infty) \to (0, \infty)$ such that $\sum_{i=1}^{\infty} L^{-1}(\Gamma_i) < \infty$, the point process

$$\kappa = \sum_{i=1}^{\infty} \delta_{L^{-1}(\Gamma_{r+i}/\Gamma_r)} \tag{3.2.1}$$

follows an NBP(r, L) on $\mathbb{E} = (0, L^{-1}(1))$.

The proof of the theorem is similar to that outlined above for the process η.

The NBP was introduced by [23] purely via its Laplace functional, lacking a representation as a point process. As it is seen, the point process representation of NBP(r, L) identified in expression (3.2.1) was derived directly as a functional of a PRM.

Remark 3.2.2. In [27], a point process representation of the NBP, which equals the representation identified in expression (3.2.1) in distribution, was derived using ordered jumps of a trimmed subordinator. If $(X_t)_{t\geq 0}$ is a driftless subordinator with Lévy measure L, by writing $(\Delta X_t := X_t - X_{t-})_{t>0}$ for the jump process of X_t, and $\Delta X_t^{(1)} \geq \Delta X_t^{(2)} \geq \cdots$ for the ordered jumps at $t > 0$, then the point process $\mathbb{B}^{(r)} = \sum_{i=1}^{\infty} \delta_{J_r(i)}$ follows NBP(r, L) where $J_r(i) = \Delta X_1^{(r+i)} / \Delta X_1^{(r)}$, for all $i \in \{1, 2, \ldots\}$.

Remark 3.2.3. In the literature, the terminology "negative binomial process" is used for mathematically distinct concepts. Therefore, to avoid confusion, it seems wise to clarify these concepts. In this chapter, as in works by [27], along with [28, 29, 30], we adhere to the definition of the NBP as outlined by [23]. However, some authors in engineering and computer science present alternative definitions. For example, the definition offered in [68] is different from the one given in [69] and [8], all of which differ from the Gregoire definition which is the focus of our investigations in this chapter.

Following Definition 2.1 of [27], by normalizing the points of the NBP identified in expression (3.2.1), the sequence

$$\left(\frac{L^{-1}(\Gamma_{r+1}/\Gamma_r)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_{r+i}/\Gamma_r)}, \frac{L^{-1}(\Gamma_{r+2}/\Gamma_r)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_{r+i}/\Gamma_r)}, \ldots \right) \tag{3.2.2}$$

defines a 2-parameter random discrete distribution on the infinite-dimensional simplex $\mathbb{S}_\infty$. This distribution is called a Poisson-Kingman distribution generated by NBP(r, L) and is denoted by PK$^{(r)}(L)$. Also, since an NBP(r, L) can be characterized as a PRM with randomized intensity measure $\Gamma_r L$, i.e. PRM$(\Gamma_r L) \stackrel{d}{=}$ NBP(r, L), an equivalent construction of the PK$^{(r)}(L)$ sequence can be constructed from a gamma-subordinated Lévy process.

Theorem 3.2.4. Let $(X_t)_{t\geq 0}$ be a driftless subordinator with Lévy measure L and let $(\sigma_r)_{r>0}$ be an independent gamma subordinator, i.e., a subordinator having Lévy measure specified by expression (2.2.6) with $\theta = 1$. The sequence

$$\left(\frac{\Delta X_{\sigma_r}^{(1)}}{X_{\sigma_r}}, \frac{\Delta X_{\sigma_r}^{(2)}}{X_{\sigma_r}}, \ldots \right) \tag{3.2.3}$$

equal in distribution the representation identified in expression (3.2.2).

Given the form identified in expression (3.1.2), the proof follows a similar logic to that used to prove Theorem 3.2.1.

Definition 3.2.5. Let $(Z_i)_{i\geq 1}$ be i.i.d. random variables with values in $\mathbb{E}$ and common distribution H. Then we may introduce the following random discrete probability measure on $\mathbb{E}$ as a functional of expression (3.2.1) or using the sequence identified in expression (3.2.2) to obtain

$$P_{r,L,H}(\cdot) = \sum_{i=1}^{\infty} \frac{L^{-1}(\Gamma_{r+i}/\Gamma_r)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_{r+i}/\Gamma_r)} \delta_{Z_i}(\cdot). \tag{3.2.4}$$

We employ the notation $\mathrm{PKP}^{(r)}(H;L)$ for the distribution of the random discrete probability measure defined in expression (3.2.4), and we write $P_{r,L,H} \sim \mathrm{PKP}^{(r)}(H;L)$. Note that if $r = 0$ and we set $\Gamma_0 = 1$, then expression (3.2.4) reduces to the representation identified in expression (2.2.5).

The mean and variance of $P_{r,L,H}$ for a $A \in \mathscr{E}$ are

$$E\{P_{r,L,H}(A)\} = H(A), \quad Var\{P_{r,L,H}(A)\} = kH(A)(1 - H(A))$$

where

$$k = k(r) = E\left\{\sum_{i=1}^{\infty}\left(\frac{L^{-1}(\Gamma_{r+i}/\Gamma_r)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_{r+i}/\Gamma_r)}\right)^2\right\}.$$

They are obtained in a similar fashion as those derived in [52, section 3].

Direct calculation of $k(r)$ may appear challenging. However it can be calculated from a simple simulation. A simulation for $k(r)$ is displayed in Figure 3.1 for the case where L represents the gamma Lévy measure specified in expression (2.2.6). For positive integer r, the points are connected by straight lines only for visual convenience. As it is depicted in Figure 3.1, $k(r)$ is a decreasing sequence of r; therefore, it suggests that we need to take a moderately small value of r to avoid having a rigid prior when we use $P_{r,L,H}$ as a prior in a hierarchical nonparametric Bayesian model in Subsection 4.2.

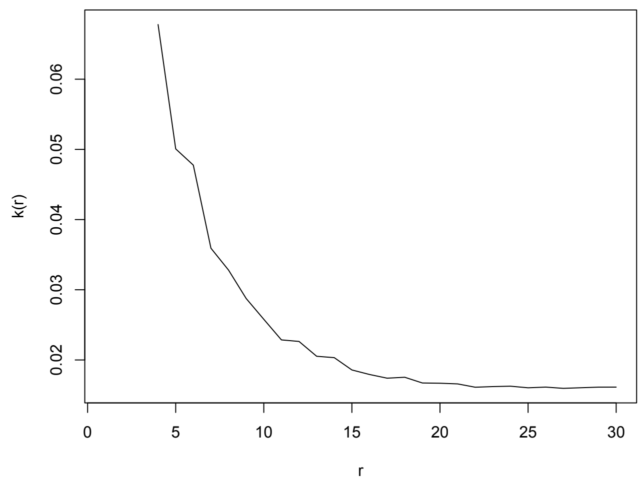

Figure 3.1: Plot of simulated $k(r)$ when L is a gamma Lévy measure with $\theta = 3$.

3.3 The extended Dirichlet process and its approximation

When L is defined as the α-stable Lévy measure identified in expression (2.1.7), the distribution $\mathrm{PK}^{(r)}(L)$ has been extensively studied. [28] derive its connection to other Poisson–Dirichlet models as $r \to \infty$. Moreover, [30] apply this distribution to gene and species sampling data. They also highlight some analytical advantages of using the additional parameter r.

We now take the probability measure identified in expression (3.2.4) when L is the gamma Lévy measure defined in expression (2.2.6) to develop a prior distribution on the space of all probability distributions. This prior would be a natural extension of the DP (the DP appears when $r = 0$ and $\Gamma_0 = 1$). In the following theorem we provide an efficient approximation for this extended DP.

Theorem 3.3.1. Let G_n and G_n^{-1} be defined as in Theorem 2.2.10. Also, let L be the gamma Lévy measure as defined in expression (2.2.6) and $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with values in $\mathbb{E}$ and common distribution H, independent of $(\Gamma_i)_{i\geq 1}$.

Then, as $n \to \infty$

$$P_{n,r,H} = \sum_{i=1}^{n} \frac{G_n^{-1}\left(\frac{\Gamma_{r+i}}{\Gamma_r \Gamma_{n+1}}\right)}{\sum_{i=1}^{n} G_n^{-1}\left(\frac{\Gamma_{r+i}}{\Gamma_r \Gamma_{n+1}}\right)} \delta_{Z_i} \xrightarrow{a.s.} P_{r,L,H} = \sum_{i=1}^{\infty} \frac{L^{-1}\left(\frac{\Gamma_{r+i}}{\Gamma_r}\right)}{\sum_{i=1}^{\infty} L^{-1}\left(\frac{\Gamma_{r+i}}{\Gamma_r}\right)} \delta_{Z_i}$$

on $\mathbb{E}$ with respect to the weak topology.

Proof. The proof is similar to that of Theorem 1 in [67]. Using the fact that $G_n^{-1}(x/(cn)) \xrightarrow{a.s.} L^{-1}(x/c)$, we have $G_n^{-1}(\Gamma_{r+i}/(\Gamma_r\Gamma_{n+1})) \xrightarrow{a.s.} L^{-1}(\Gamma_{r+i}/\Gamma_r)$ by taking the constant $c = \Gamma_r$, $x = \Gamma_{r+i}$, and $n/\Gamma_{n+1} \xrightarrow{a.s.} 1$ as $n \to \infty$. ∎

Our proposed approximation offers several advantages. Notably, our representation circumvents the need for an infinite sum, employing a finite number of weights. These weights are derived from the quantile functions of the Gamma$(\theta/n, 1)$ distribution, evaluated at $1 - \Gamma_{r+i}/\Gamma_r\Gamma_{n+1}$. In the previous representation, it is necessary to calculate L^{-1}, which cannot be written in a closed form. In addition, our introduced weights are stochastically decreasing contrary to the stick-breaking weights in [27]. A similar proposal for the DP can be found in [67].

3.4 A new alternative series representation for the Poisson-Dirichlet process

The definition of the Poisson-Dirichlet process, along with some of its approximations, has been discussed in Subsection 2.2.2. As noted in that subsection, [3, Lemma 2.1] demonstrates that the p_i''s are not strictly decreasing with probability one. Consequently, an approximation of $P_{\alpha,\theta,H}$ using a truncation method has been shown through simulations to be inefficient. Another approach, in particular Algorithm A in [3], was proposed to approximate $P_{\alpha,\theta,H}$ which was based on Proposition 22 of [56]. This approach was faster and more accurate than the method based on truncating the stick-breaking representation, although it requires more complex steps in its algorithm. In this section, we apply our new representation of the NBP (3.2.1) to the Proposition 21 in [56] to give a new representation for the Poisson-Dirichlet process. Then, we will show how simulating the Poisson-Dirichlet process driving from this representation is less complex and more precise compared to the algorithm A in [3].

Now, following the Proposition 21 in [56], for $\theta > 0$ and $0 < \alpha < 1$, let $(X_t)_{t\geq 0}$ be a subordinator with Lévy measure

$$L(x) = \frac{\alpha}{\Gamma(1-\alpha)} \int_x^\infty u^{-\alpha-1} e^{-u} \mathrm{d}u, \;\; x > 0 \tag{3.4.1}$$

and let $(\sigma_r)_{r>0}$ be an independent gamma subordinator. Then

$$\left(\frac{\Delta X_T^{(1)}}{X_T}, \frac{\Delta X_T^{(2)}}{X_T}, \ldots \right) \sim \mathrm{PD}(\alpha, \theta) \;\; \text{if} \;\; T = \sigma_{\theta/\alpha}. \tag{3.4.2}$$

Comparing expression (3.4.2) to expression (3.2.3) we see that $\mathrm{PD}(\alpha,\theta) \overset{d}{=} \mathrm{PK}^{(\theta/\alpha)}(L)$ with L specified in expression (3.4.1). Also, since expression (3.2.3) equals expression (3.2.2) in distribution, we can conclude that $\mathrm{PDP}(H;\alpha,\theta) \overset{d}{=} \mathrm{PKP}^{(\theta/\alpha)}(H;L)$ for $\theta > 0, 0 < \alpha < 1$ and L identified in expression (3.4.1). In other words, the random probability measure

$$P_{r,L,H}(\cdot) = \sum_{i=1}^{\infty} \frac{L^{-1}(\Gamma_{r+i}/\Gamma_r)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_{r+i}/\Gamma_r)} \delta_{Z_i}(\cdot) \tag{3.4.3}$$

with L identified in expression (3.4.1) and $r = \theta/\alpha$ distributed as either $\mathrm{PDP}(H;\alpha,\theta)$ or $\mathrm{PKP}^{(\theta/\alpha)}(H;L)$. Therefore, expression (3.4.3) provides another series representation of the PDP for the case $\theta > 0$ and $0 < \alpha < 1$ through a negative binomial process.

Remark 3.4.1. If we take $r = 0$ and set $\Gamma_0 = 1$, we see that normalized generalized Gamma process introduced in [48] has the same representation, namely expression (3.4.3), of the PDP with $r = 0$.

Remark 3.4.2. Note that $\mathrm{PK}(L) \overset{d}{=} \mathrm{PD}(\alpha, 0)$ for $0 < \alpha < 1$ and L identified in expression (2.1.7) and also, $\mathrm{PK}(L) \overset{d}{=} \mathrm{PD}(0,\theta)$ for $\theta > 0$ and L as specified in expression (2.2.6). See [27] for stick-breaking representations of $\mathrm{PK}^{(r)}(L)$ with L as specified in expressions (2.1.7) and (2.2.6).

3.4.1 A novel approximation method for the Poisson–Dirichlet process

By truncating this new series representation of the Poisson-Dirichlet process PDP$(H;\alpha,\theta)$ specified in expression (3.4.3) we can approximate this process using the finite sum

$$P_{N,r,L,H}(\cdot) = \sum_{i=1}^{N} \frac{L^{-1}(\Gamma_{r+i}/\Gamma_r)}{\sum_{i=1}^{N} L^{-1}(\Gamma_{r+i}/\Gamma_r)} \delta_{Z_i}(\cdot) \tag{3.4.4}$$

for L as specified in expression (3.4.1) and $r = \theta/\alpha$ where $0 < \alpha < 1$ and $\theta > 0$; we suggest using the stopping rule $N = N(\epsilon)$, where

$$N = \inf \left\{ i : \frac{L^{-1}(\Gamma_{r+i}/\Gamma_r)}{\sum_{j=1}^{i} L^{-1}(\Gamma_{r+j}/\Gamma_r)} < \epsilon \right\} \quad \text{for } \epsilon \in (0,1).$$

Since the weights

$$w_i = \frac{L^{-1}(\Gamma_{r+i}/\Gamma_r)}{\sum_{i=1}^{N} L^{-1}(\Gamma_{r+i}/\Gamma_r)}, \quad i \in \{1,\ldots,N\}$$

are strictly decreasing, simulating the PDP via the representation identified in expression (3.4.4) is very efficient. The main complexity in using this approximation lies in applying the Newton-Raphson method to calculate L^{-1} at points Γ_{r+i}/Γ_r in order to determine the probability weights w_i.

Here, we compare our approximation with Algorithm A in [3] which is based on Proposition 22 from [56] and also presented as Proposition 2.2.15 in Subsection 2.2.2.

In Algorithm A of [3], an approximation of the DP(θ, H) is used to generate n probability weights

$$(p_1(0,\theta), p_2(0,\theta), \ldots, p_n(0,\theta))$$

from $PD(0,\theta)$. Additionally, an approximation of the normalized α-stable law process is employed to obtain n i.i.d. copies of the weights

$$(p_1(\alpha,0), p_2(\alpha,0), \ldots, p_m(\alpha,0)),$$

denoted as

$$(p_1^1(\alpha,0),\ldots,p_m^1(\alpha,0)),\ldots,(p_1^n(\alpha,0),\ldots,p_m^n(\alpha,0)),$$

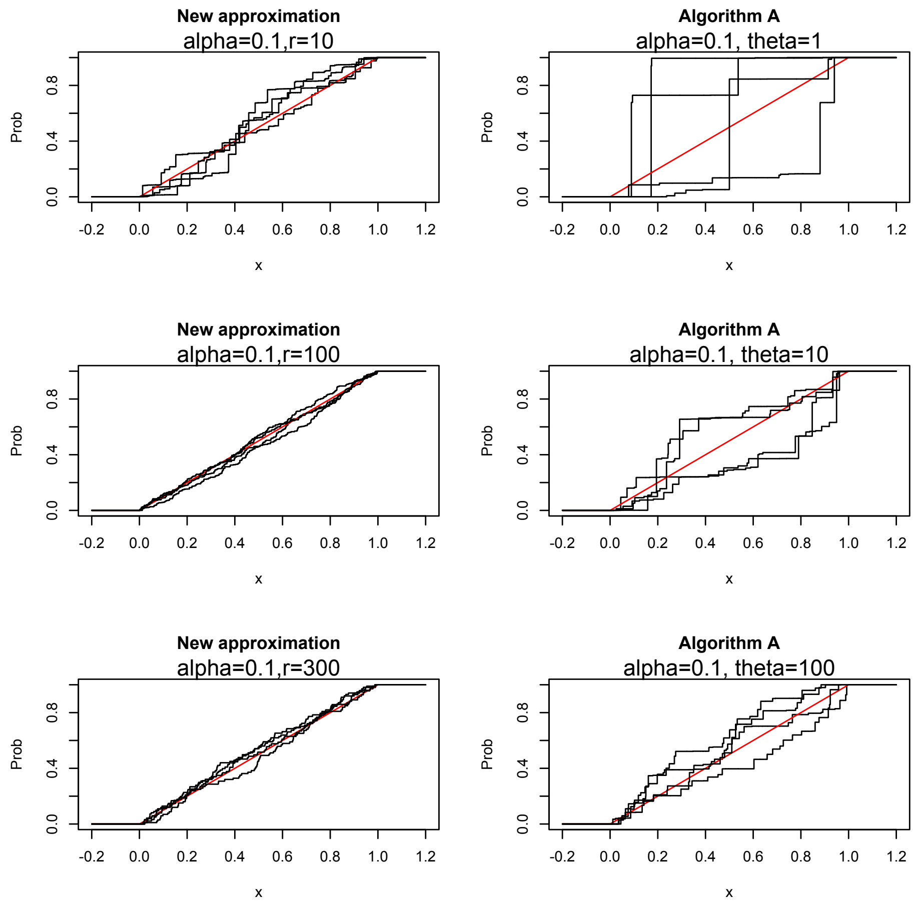

Figure 3.2: Sample paths of the Poisson-Dirichlet process, where H is the uniform distribution on $[0,1]$, $\alpha = 0.1$, and $\theta = 1, 10, 100$. The red line denotes the cumulative distribution function of H.

from $PD(\alpha, 0)$. Finally, the PDP is approximated using a truncated version of expression (3.4.3) with truncation value $N = n \times m$ and probability weights $(p_i)_{1 \le i \le N}$ equal to

$$(p_1(0,\theta)p_1^1(\alpha,0), \ldots, p_1(0,\theta)p_m^1(\alpha,0), \ldots, p_n(0,\theta)p_1^n(\alpha,0), \ldots, p_n(0,\theta)p_m^n(\alpha,0))$$

when arranged in a descending order.

Figures 3.2, 3.3, and 3.4 show sample paths for the approximate PDP for different values of α and θ ($r = \theta/\alpha$). Clearly, the approximation specified in expression (3.4.4) outperforms the approximation given in Algorithm A in [3] in all cases, as the sample paths for this approximation deviate less from the base measure H. This behaviour agrees with Chebyshev's

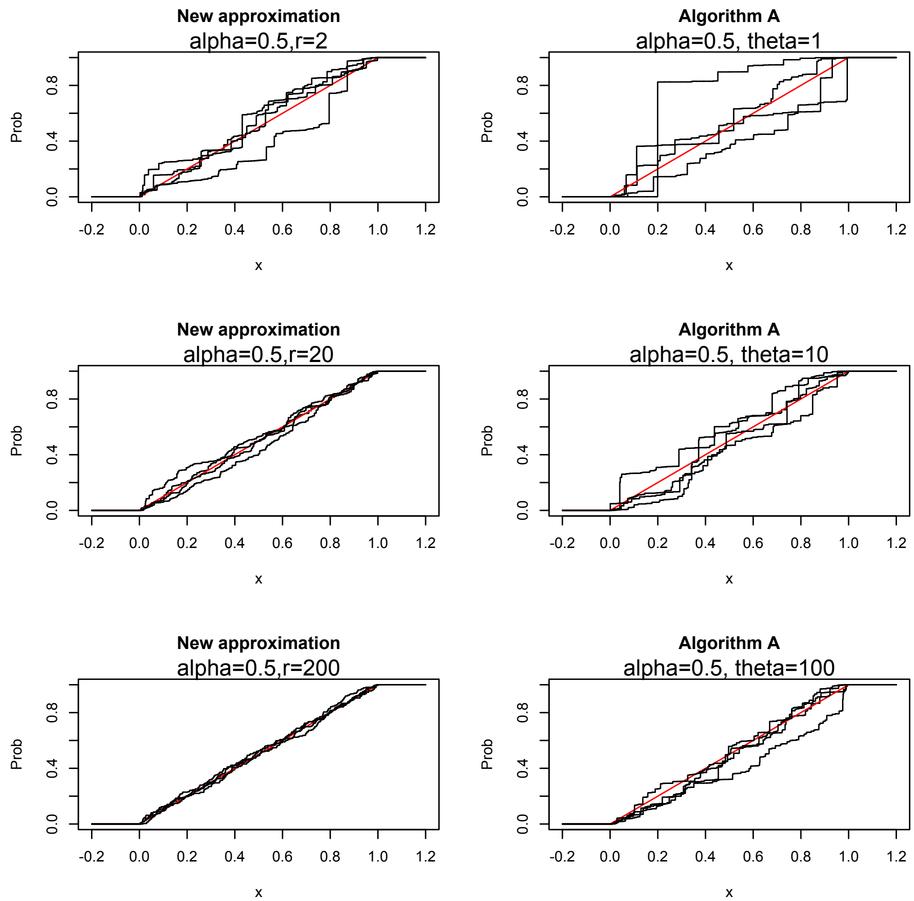

Figure 3.3: Sample paths of the Poisson-Dirichlet process, where H is the uniform distribution on $[0, 1]$, $\alpha = 0.5$, and $\theta = 1, 10, 100$. The red line denotes the cumulative distribution function of H.

inequality. As is expected, a sample path should approach the base measure faster when either α or θ gets larger.

In the simulation, we set $n = 20$, $m = 20$ in Algorithm A of [3], and $N = 20 \times 20 = 400$ in expression (3.4.4). In each simulation scenario we fixed the base measure H to be the uniform distribution on $[0, 1]$. We also computed the Kolmogorov distance between the PDP and H for different values of α and θ ($r = \theta/\alpha$). The Kolmogorov distance between $P_{N,r,L,H}$ and H, denoted by $d(P_{N,r,L,H}, H)$, is defined as

$$d(P_{N,r,L,H}, H) = \sup_{x \in \mathbb{R}} |P_{N,r,L,H}(x) - H(x)|.$$

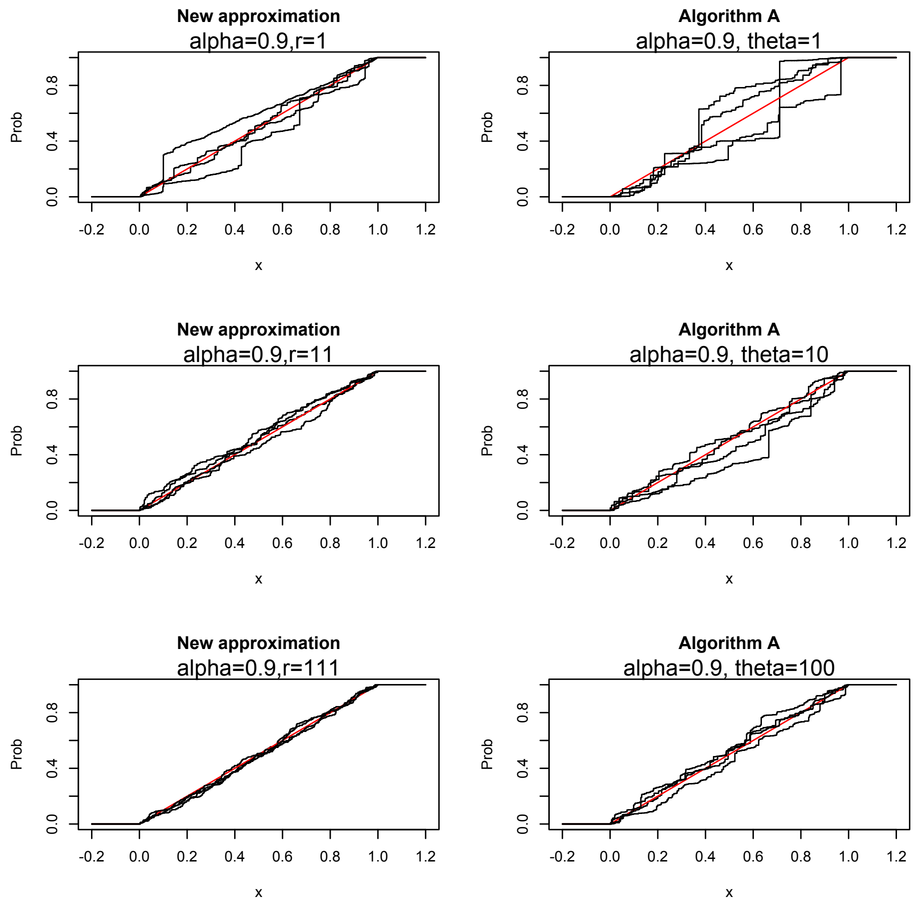

Figure 3.4: Sample paths of the Poisson-Dirichlet process, where H is the uniform distribution on $[0, 1]$, $\alpha = 0.9$, and $\theta = 1, 10, 100$. The red line denotes the cumulative distribution function of H.

For different values of α and θ, we simulated 500 Kolmogorov distances and reported the average of these values in Table 3.1. From the simulation results summarized in Table 3.1, we can conclude that our proposed approximation outperforms in all cases.

Table 3.1: This table reports the Kolmogorov distance $d(P_{N,r,L,H}, H)$, where H is a uniform distribution on $[0, 1]$.

			New	Algorithm A
α	θ	r	d	d
0.1	1	10	0.03133	0.35899
0.1	10	100	0.03819	0.19301
0.1	100	300	0.06443	0.13835
0.5	1	2	0.14497	0.24855
0.5	10	20	0.06549	0.14268
0.5	100	200	0.04606	0.10126
0.9	1	1	0.09294	0.18038
0.9	10	11	0.05178	0.10100
0.9	100	111	0.03998	0.07124

Chapter 4

Applications

In this chapter, we first highlight key properties of samples from the random discrete probability measure introduced in Chapter 3. We then apply the new family of priors within a Bayesian nonparametric hierarchical model and compare the results with those obtained using a model based on the Dirichlet prior. To demonstrate the performance of these models, we conduct a simulation study using both real-world and simulated datasets.

4.1 Regulating the number of clusters

We represent an i.i.d. sample of size n from a random discrete probability measure P, as

$$(X_1, \ldots, X_n)|P \overset{iid}{\sim} P.$$

For a sufficiently large n, it is likely to observe ties within the sample with positive probability. Let $K_n = k$, where $k \in \{1, \ldots, n\}$, indicate the number of unique values within n observations. Denote these distinct values as $X'_1, \ldots, X'_k$. Moreover, take $n_j = \sum_{i=1}^{n} I(X_i = X'_j)$ for $j \in \{1, 2, \ldots, k\}$ where I denotes the indicator function; obviously, $n_1 + \cdots + n_k = n$. This framework can partition the set $\{1, 2, \ldots, n\}$ into k subsets or blocks. Alternatively, a partition can be represented by the blocks count vector $\boldsymbol{M} = (M_1, M_2, \ldots, M_n)$, where M_j denotes the number of blocks with j representatives in a sample of size n. Hence, $\sum_{j=1}^{n} jM_j = n$, and $K_n = M_1 + \cdots + M_n$. The joint distribution of cluster sizes and the distribution of the number of clusters, when probability weights of $P_{r,L,H}$ follow $\mathrm{PK}^{(r)}(L)$ with L as specified in expression (2.1.7), are provided in [28, Theorem 5.1] and Equation (5.13) of that paper, respectively. That theorem can be rewritten for any Lévy measure L satisfying $\sum_{i=1}^{\infty} L^{-1}(\Gamma_i) < \infty$.

Theorem 4.1.1. The distribution of the number of clusters is given by

$$\Pr(K_n = k) = \int_0^\infty \frac{r^{[k]}\lambda^{n-1}}{\Gamma(n)\Psi(\lambda)^{r+k}} B_{n,k}\left(\bar{\Psi}^{(1)}(\lambda), \ldots, \bar{\Psi}^{(n-k+1)}(\lambda)\right) \mathrm{d}\lambda$$

and the joint distribution of the cluster sizes is

$$p(n_1, \ldots, n_k) = \frac{r^{[k]}}{\Gamma(n)} \int_0^\infty \lambda^{n-1}\Psi(\lambda)^{-r-k} \prod_{i=1}^{k} \bar{\Psi}^{(n_j)}(\lambda)\mathrm{d}\lambda$$

where

$$r^{[k]} = r(r+1)\ldots(r+n-1),\ n \in \mathbb{N},$$
$$\Psi(\lambda) = 1 + \int_0^1 \left(1 - e^{-\lambda x}\right) \tilde{L}(\mathrm{d}x),$$
$$\bar{\Psi}^{(j)}(\lambda) = (-1)^{j-1}\frac{d^j}{d\lambda^j}\Psi(\lambda) = \int_0^1 x^j e^{-\lambda x}\tilde{L}(\mathrm{d}x),$$
$$\tilde{L}(\mathrm{d}x) = L(\mathrm{d}x)I(0 < x < L^{-1}(1)),$$

and $B_{n,k}$ is the Bell partition polynomial, satisfying, for $\xi_j > 0$, $j \in \{1, \ldots, n\}$

$$B_{n,k}(\xi_1, \ldots, \xi_{n-k+1}) = \sum_{\boldsymbol{m}\in A_{nk}} \frac{n!}{m_1! \ldots m_n!} \prod_{j=1}^{n} \left(\frac{\xi_j}{j!}\right)^{m_j}$$

where

$$A_{nk} := \{\boldsymbol{m} = (m_1, \ldots, m_n) : m_j \geq 0, \sum_{j=1}^{n} jm_j = n, \sum_{j=1}^{n} m_j = k\}.$$

As demonstrated by [45] and [55] for the Dirichlet process, we have

$$K_n/\log(n) \xrightarrow{a.s.} \theta \ \text{ as } \ n \to \infty.$$

According to this result, the random number of distinct values K_n grows slowly as $n \to \infty$, only in a logarithmic fashion. The reason for this behaviour is because the Dirichlet process prior assigns most of the largest weights to its initial points. This property causes inflexibility

Table 4.1: The first ten probability weights in (3.2.4) when L is the gamma Lévy measure (2.2.6) with $\theta = 3$.

$r = 0$	$r = 3$	$r = 5$	$r = 10$
0.367597022	0.24369485	0.16427045	0.06353087
0.168573239	0.23947841	0.14319002	0.05886303
0.165457071	0.10281577	0.12842541	0.05418117
0.149080111	0.08716432	0.10524699	0.05112053
0.058821776	0.07639968	0.07391248	0.04858739
0.056183134	0.05990201	0.06339345	0.04626107
0.012551887	0.03862790	0.05279059	0.04349017
0.007812625	0.03184211	0.04298598	0.03795044
0.003792634	0.02524151	0.03705701	0.03606621
0.001971704	0.01939928	0.03245281	0.03021665

in the use of the Dirichlet process as a prior mixing measure in nonparametric Bayesian hierarchical mixture models or so called density estimation problem [49, 15, 31, 47, 48]. Adding the new parameter r and working with expression (3.2.4) instead of expression (2.2.5) enables us to remove those initial large probability weights and produce a more robust allocation of the weights instead. For the case when L is the gamma Lévy measure, the values reported in Table 4.1 illustrates how choosing larger values of r leads to more robust versions of the probability weights in expression (3.2.4). To generate this table, we used Theorem 3.3.1. Recall that the Dirichlet process corresponds to the case when $r = 0$. In the following section, we explore the benefits of using the random discrete probability measure $P_{r,L,H}$ instead of the DP in hierarchical nonparametric Bayesian models.

4.2 Hierarchical nonparametric Bayesian models

Here, we compare the hierarchical nonparametric Bayesian model that utilizes the prior $P_{r,L,H}$ (with a gamma Lévy measure) as the mixing measure, to the hierarchical nonparametric

Bayesian model that employs the DP as the mixing measure. The structure of a hierarchical nonparametric Bayesian model is as follows:

$$\begin{aligned} (X_i|Y_i) &\overset{ind}{\sim} \mathscr{L}(X_i|Y_i), \quad i \in \{1,\dots,n\} \\ (Y_i|P) &\overset{iid}{\sim} P, \\ P &\sim \mathscr{P}. \end{aligned} \tag{4.2.1}$$

In this setting, the vector $X = (X_1, \dots, X_n)$ represents the observed data, whereas $Y = (Y_1, \dots, Y_n)$ denotes unobserved random variables. These variables take values within a measurable space $(\mathbb{E}, \mathscr{E})$. According to the model structure outlined at statement (4.2.1), we assume that the observed data, X_i, are conditionally independent given the unobserved variables, Y_i. Furthermore, given a probability measure P, the Y_i variables are independently and identically distributed following P.

The hierarchical model we discuss is nonparametric, aimed at modelling the conditional distribution of X_i given Y_i, denoted as $\mathscr{L}(X_i|Y_i)$. This is achieved by modelling the distribution of Y_i through a random probability measure $\mathscr{P}$. In this context, $\mathscr{P}$ specifically refers to $\mathrm{PKP}^{(r)}(H;L)$ or DP. Our focus is on comparing the use of $\mathrm{PKP}^{(r)}(H;L)$ as an alternative to the DP.

Additionally, the model can be extended to a semiparametric hierarchical model. This extension involves incorporating a finite-dimensional parameter μ. Within this extended framework, $\mathscr{L}(X_i|Y_i, \mu)$ represents the conditional distribution of X_i given Y_i and the parameter μ.

To draw samples from the posterior distributions in the model identified by statement (4.2.1), two algorithms are typically employed: the Pólya urn Gibbs sampler and the blocked Gibbs sampler. Constructing a Pólya urn Gibbs sampler requires knowledge of the predictive distribution of the latent variables, i.e., the Y_i's. When $P \sim \mathrm{PKP}^{(r)}(H;L)$, this predictive distribution, along with the corresponding Pólya urn Gibbs sampler, is detailed in [24]. Specifically, if $P \sim \mathrm{PKP}^{(r)}(H;L)$, then the predictive distribution for Y_{n+1} given Y follows the same law as

$$\Pr(Y_{n+1} \in \mathrm{d}y \mid Y) = \omega_0^{(n)} H(\mathrm{d}y) + \frac{1}{n} \sum_{i=1}^{k} \omega_i^{(n)} \delta_{Y_i'}(\mathrm{d}y)$$

where

$$\omega_0^{(n)} = \frac{r+k}{n} \int_0^\infty v \frac{\bar{\Psi}^{(1)}(v)}{\Psi(v)} \, g_r(v, \boldsymbol{n}) \, \mathrm{d}v,$$

and for $i \in \{1, \ldots, k\}$,

$$\omega_i^{(n)} = \int_0^\infty v \, \frac{\bar{\Psi}^{(n_i+1)}(v)}{\bar{\Psi}^{(n_i)}(v)} \, g_r(v, \boldsymbol{n}) \, \mathrm{d}v.$$

In above, $\boldsymbol{n} = \{n_1, \ldots, n_k\}$ and

$$g_r(v, \boldsymbol{n}) := \frac{r^{[k]}}{(\Psi(v))^{r+k}} \frac{v^{n-1}}{\Gamma(n)} \prod_{i=1}^{k} \bar{\Psi}^{(n_i)}(v), \quad v > 0.$$

Remark 4.2.1. According to Proposition 1 in [7], $\mathrm{PKP}^{(r)}(H; L)$ is not a Gibbs–type prior.

In contrast, to construct the blocked Gibbs sampler introduced in [31], we need to have a stick-breaking construction for the mixing prior in the model identified in statement (4.2.1). A stick-breaking random discrete probability measure which we denote it by P^{stick} is as follows

$$P^{stick} = \sum_{k=1}^{\infty} p_k \delta_{Z_k}$$

where

$$p_1 = V_1, \quad p_k = V_k \prod_{j=1}^{k-1} (1 - V_j), \quad k = 2, \ldots$$

and the V_k are independent $Beta(a_k, b_k)$ random variables and Z_k's are iid from a nonatomic distribution H independent of V_k's. Since P^{stick} is represented as an infinite sum, it is approximated by truncating the sum, i.e.,

$$P^{stick} \approx P_N^{stick} = \sum_{k=1}^{N} p_k \delta_{Z_k}$$

where

$$p_1 = V_1, \quad p_k = V_k \prod_{j=1}^{k-1} (1 - V_j), \quad k \in \{2, \ldots, N\}$$

and the V_k are independent $Beta(a_k, b_k)$ random variables for $k \leq N - 1$ and $V_N = 1$. Note that, in this case, $p = (p_1, \ldots, p_N)$ follows the generalized Dirichlet distribution denoted by

$$p \sim \mathscr{G}(a_1, b_1, \ldots, a_{N-1}, b_{N-1}).$$

It is well-known that the generalized Dirichlet distribution is conjugate for multinomial sampling. By [10], the density function of the generalized Dirichlet distribution equals

$$\left\{\prod_{k=1}^{N-1} \frac{\Gamma(a_k+b_k)}{\Gamma(a_k)\Gamma(b_k)}\right\} p_1^{a_1-1} \dots p_{N-1}^{a_{N-1}-1} p_N^{b_{N-1}-1} \times (1-P_1)^{b_1-(a_2+b_2)} \dots (1-P_{N-2})^{b_{N-2}-(a_{N-1}+b_{N-1})}, \tag{4.2.2}$$

where $P_k = p_1 + \ldots + p_k$. For DP, we take $a_k = 1$ and $b_k = \theta$. A stick-breaking representation for $\text{PKP}^{(r)}(H; L)$ when L is the gamma Lévy measure specified in expression (2.2.6) is given in [27, Theorem 4.1]. More precisely, it follows that $P_{r,L,H}$ with L as the gamma Lévy measure specified in expression (2.2.6) is equal to P^{stick} in distribution if

$$V_k|\Gamma_r = v \overset{iid}{\sim} Beta(1, v\theta), \quad k \in \{1, 2, \ldots\}$$

where Γ_r is a Gamma$(r, 1)$ random variable. In this case, the density functions of V_k and p, respectively, are

$$f_{V_k}(x) = \frac{r\theta}{1-x} \frac{1}{(1-\theta\ln(1-x))^{r+1}}, \quad x \in (0,1)$$

and

$$f_{p_1,\ldots,p_{N-1}}(p_1, \ldots, p_{N-1}) = \prod_{i=1}^{N-1} \frac{r\theta}{1-\sum_{i=1}^{j} p_i} \left\{1 - \theta\ln\left(\frac{1-\sum_{i=1}^{j} p_i}{1-\sum_{i=1}^{j-1} p_i}\right)\right\}^{-r-1}.$$

As we observe, both the Pólya urn and blocked Gibbs samplers are computationally intensive for the prior $\text{PKP}^{(r)}(H; L)$. The former depends on a complex predictive distribution, while for the latter, the distribution of p is somewhat complex. We choose to utilize the blocked Gibbs sampler for two main reasons. First, beyond the limitations of the Pólya urn Gibbs sampler discussed in Section 4.3 of [31], Section 6 of the same paper compares the mixing behaviour of the blocked Gibbs sampler with both the Pólya urn Gibbs sampler and its accelerated (modified) version. The results indicate that the blocked Gibbs sampler not only outperforms the standard Pólya urn Gibbs sampler but also exhibits comparable performance to the accelerated version. Second, if we choose to ignore specifying a prior on r and loosely substitute Γ_r during the simulation of V_k, we can continue to use the generalized Dirichlet distribution specified in statement (4.2.2).

The key strategy for achieving an efficient blocked Gibbs sampler for sampling from the nonparametric hierarchical model identified in statement (4.2.1) is to reformulate the model entirely with random variables. Let $p = (p_1, \ldots, p_N)$ and $Z = (Z_1, \ldots, Z_N)$. Then the model can be rewritten as

$$\begin{aligned}(X_i|Z,K) &\sim \mathscr{L}(X_i|Z_{K_i}), \quad i \in \{1,\ldots,n\}\\ (K_i|p) &\sim \sum_{k=1}^{N} p_k\delta_k(\cdot),\\ (p,Z) &\sim \mathscr{L}(p)\mathscr{L}(Z)\end{aligned} \tag{4.2.3}$$

where $K = (K_1, \ldots, K_n)$ and the K_i are conditionally independent classification variables that identify the Z_k associated with each Y_i. Specifically, note that $Y_i = Z_{K_i}$. To extend this model to a mixture of hierarchical models, we introduce a prior for θ in p and include hyperparameters for Z. If we use the representation specified in statement (4.2.3), the model we consider is

$$\begin{aligned}(X_i|Z,K) &\sim \mathcal{N}(Z_{K_i}, \sigma_X),\\ (K_i|p) &\sim \sum_{k=1}^{N} p_k\delta_k(\cdot),\\ (Z_k|\mu,\sigma_Z) &\sim \mathcal{N}(\mu,\sigma_Z),\\ (\mu|\sigma_\mu) &\sim \mathcal{N}(0,\sigma_\mu),\\ (\sigma_Z^{-1}|\tau_1,\tau_2) &\sim \text{Gamma}(\tau_1,\tau_2),\\ (\theta|\nu_1,\nu_2) &\sim \text{Gamma}(\nu_1,\nu_2),\\ (\sigma_X^{-1}|\gamma_1,\gamma_2) &\sim \text{Gamma}(\gamma_1,\gamma_2).\end{aligned} \tag{4.2.4}$$

where $\mathcal{N}(\mu, \sigma)$ denotes a normal distribution with mean μ and variance σ.

To sample the posterior $\mathscr{L}(p, Z, K, \mu, \sigma_Z, \sigma_X, \theta|X)$ that corresponds to the model identified in statement (4.2.4) using blocked Gibbs sampler, we need to complete six updates in each cycle of the sampler:

$$\begin{aligned}&(p, Z|K, \mu, \sigma_Z, \sigma_X, \theta, X),\\ &(K|p, Z, X),\\ &(\mu|Z, \sigma_Z),\\ &(\sigma_Z|Z, \mu),\end{aligned} \tag{4.2.5}$$

$$(\theta|p),$$

$$(\sigma_X|Z, K, X).$$

See Appendix A for the derivation of the above conditional distributions.

The output of the Gibbs sampler can easily be used in density estimation problems in which we estimate a predictive density for a future observation X_{n+1}. Let $f(X_{n+1}|X)$ represent the predictive density for X_{n+1} conditioned on the data X and let Y_{n+1} be the corresponding unobserved Y value. Then

$$\begin{aligned} f(X_{n+1}|X) &= \int f(X_{n+1}|Y_{n+1}, \sigma_X) d\mathscr{L}(Y_{n+1}, \sigma_X|X) \\ &= \int\int f(X_{n+1}|Y_{n+1}, \sigma_X) d\mathscr{L}(Y_{n+1}|P) d\mathscr{L}(P, \sigma_X|X) \end{aligned}$$

For a probability measure $P(\cdot) = \sum_{k=1}^{N} p_k \delta_{Z_k}(\cdot)$,

$$\int f(X_{n+1}|Y_{n+1}, \sigma_X) d\mathscr{L}(Y_{n+1}|P) = \sum_{k=1}^{N} p_k f(X_{n+1}|Z_k, \sigma_X). \tag{4.2.6}$$

Consequently, $f(X_{n+1}|X)$ can be approximated by averaging the mixture of normal densities appearing in expression (4.2.6) over the sampled values $(p^{(b)}, Z^{(b)}, \sigma_X^{(b)})$ obtained from the Gibbs sampler. A predictive density estimate can then be derived by evaluating the averaged density over a refined partition.

4.3 An illustration

To compare the hierarchical nonparametric Bayesian model using the prior $P_{r,L,H}$ (with a gamma Lévy measure) as the mixing measure with the hierarchical nonparametric Bayesian model employing the DP as the mixing measure, we conducted experiments on both simulated data and real data. The simulated data were generated from a mixture of normal distributions with unequal variances and mixing weights

$$f(\cdot) = \sum_{i=1}^{7} \omega_i \mathcal{N}(\cdot|\mu_i, \sigma_i) \tag{4.3.1}$$

where $\omega_1 = \omega_2 = 0.1$, $\omega_3 = 0.15$, $\omega_4 = \omega_5 = 0.125$, $\omega_6 = \omega_7 = 0.2$, $(\mu_1, \ldots, \mu_7) = (-8, -5, 0, 5, 7, 11, 15)$ and $(\sigma_1, \ldots, \sigma_7) = (0.8, 0.5, 2, 0.25, 0.25, 1, 1.25)$.

Figure 4.1 illustrates the predictive density estimates identified in statement (4.2.6) based on 100 simulated observations from the mixture specified in syayement (4.3.1), using the blocked Gibbs sampling method with two different priors: the DP and the prior $P_{r,L,H}$ (incorporating a gamma Lévy measure) with $r = 3$. Each density plot represents the average of 300 samples collected after a burn-in period of 1500 iterations. For all density estimates, we set the truncation level to $N = 70$ and used the hyperparameters $\tau_1 = \tau_2 = \gamma_1 = \gamma_2 = 0.001$, $\nu_1 = 2,\ \nu_2 = 4$, and $\sigma_\mu = 100$. To maintain clarity in Figure 4.1, we only plot the density estimates using the DP and the $P_{r,L,H}$ prior with $r = 3$, as the results for the values of $r \in \{2, 3, 4, 5, 6\}$ are qualitatively similar.

Overall, we observe that employing the $P_{r,L,H}$ prior (with a gamma Lévy measure) for $r \in \{2, 3, 4, 5, 6\}$ produces highly accurate predictive densities compared to the DP prior.

As an example involving actual study data, we reanalyzed the galaxy data in [60], representing the relative velocities of $n = 82$ galaxies from six well-separated conic sections of space. The data have also been studied by [33] using the blocked Gibbs sampler outlined above and also by [15] using a Pólya Gibbs sampler. In order to be able to compare our results with their findings, we used the same values for the hyperparameters.

Figure 4.2 illustrates the predictive density estimates identified in expression (4.2.6) for the galaxy data, using the blocked Gibbs sampling method with three different priors: the DP and the prior $P_{r,L,H}$ (incorporating a gamma Lévy measure) with $r = 3$ and $r = 4$. Each density plot represents the average of 300 samples collected after a burn-in period of 1500 iterations. For all density estimates, we set the truncation level to $N = 50$ and used the hyperparameter values $\tau_1 = \tau_2 = \gamma_1 = \gamma_2 = 0.001$, $\nu_1 = 2,\ \nu_2 = 4$, and $\sigma_\mu = 100$. To maintain clarity in Figure 4.2, we only display the density estimates using the DP and the $P_{r,L,H}$ prior with $r = 3$ and $r = 4$, as the results for the values of $r \in \{2, 3, 4, 5, 6\}$ are qualitatively similar.

Overall, we observe that employing the $P_{r,L,H}$ prior (with a gamma Lévy measure) for $r \in \{2, 3, 4, 5, 6\}$ produces more accurate predictive densities compared to use of the DP prior.

Tables 4.2 and 4.3 summarize the posterior probabilities of the number of distinct Y_i values

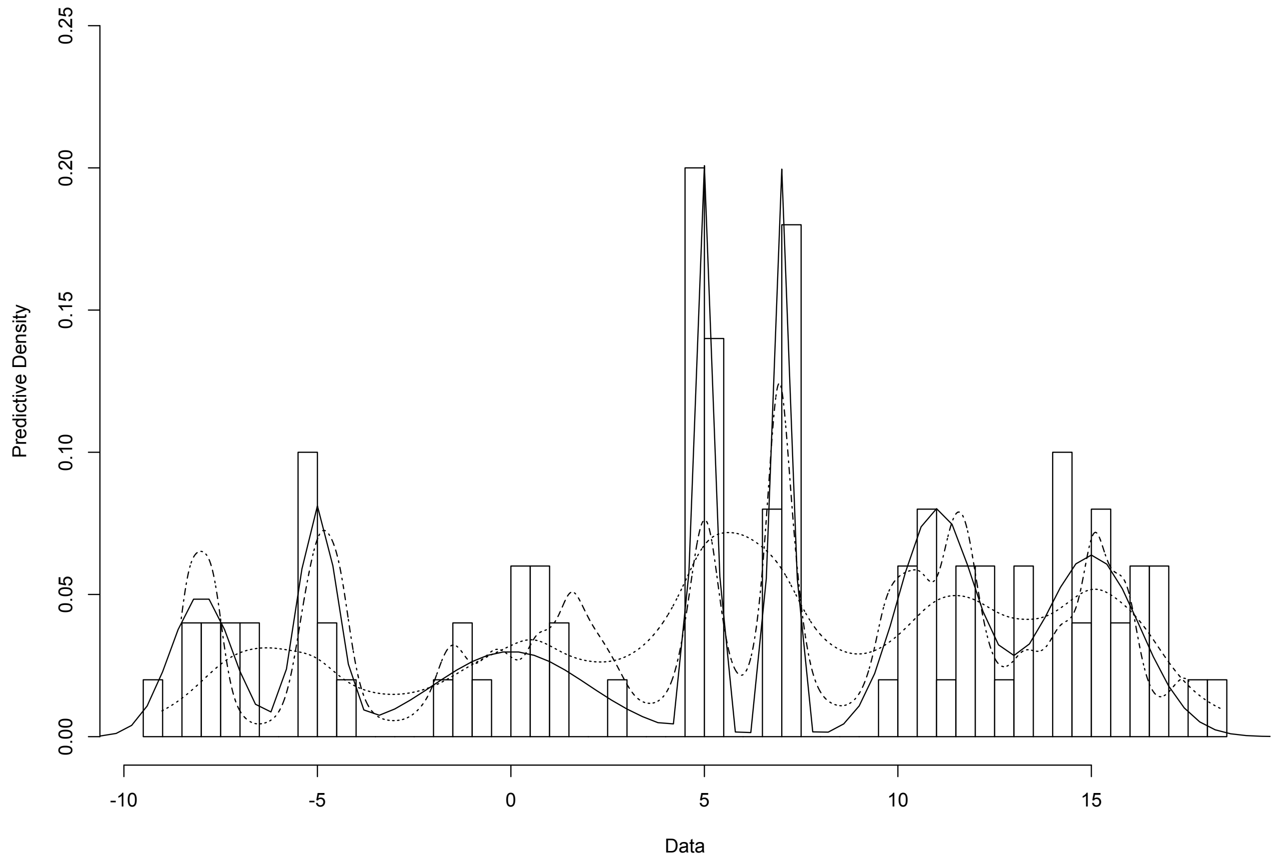

Figure 4.1: Posterior density estimates for the simulated data when we use DP and $P_{r,L,H}$ with L as the gamma Lévy measure as the prior: $\cdot\cdot\cdot\cdot\cdot$, DP,– – – –, $P_{r,L,H}$ with $r = 3$, ——, true model

corresponding to Figures 4.1 and 4.2, respectively. As observed, the tendency for overestimation becomes slightly more noticeable as the value of r increases. Furthermore, the consistency of the posterior probabilities for the number of distinct values can be verified by increasing n. For instance, in our simulated data, when n increased from 100 to 300, we observed a rise in the posterior probabilities for $k \in \{7, 8, 9, 10\}$ for the prior with $r = 3$ and $r = 4$. See Figure 4.3 and Table 4.4.

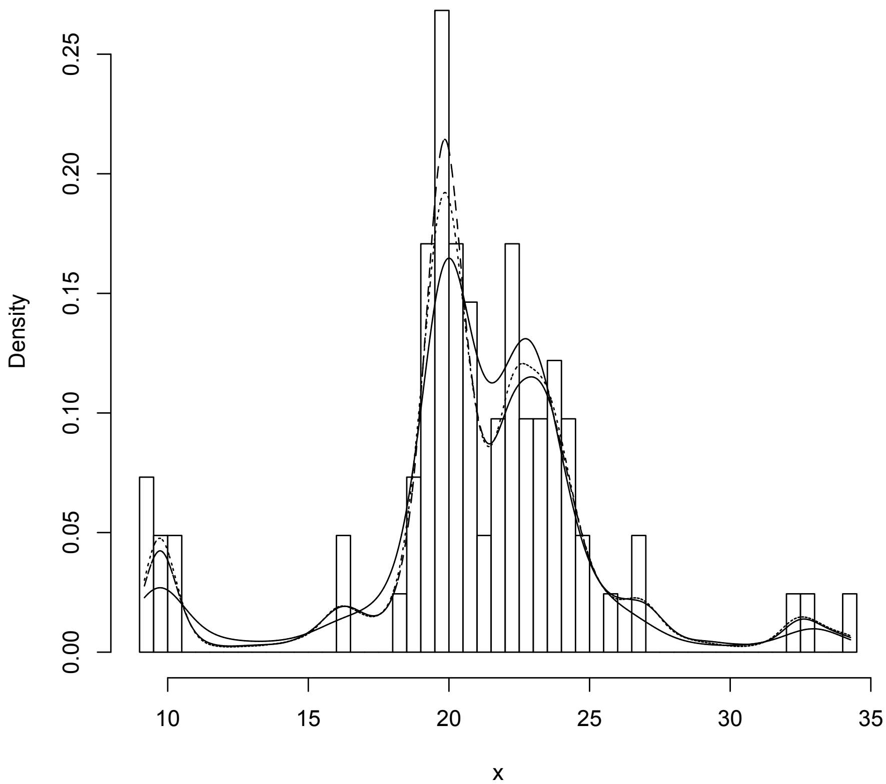

Figure 4.2: Posterior density estimates for the galaxy data when we use DP and $P_{r,L,H}$ with L as the gamma Lévy measure as the prior: $\cdots\cdots$, $P_{r,L,H}$ with $r = 3$,– – –, $P_{r,L,H}$ with $r = 4$, ———, DP

Table 4.2: Posterior probabilities of the number of distinct Y_i values from Fig. 4.1 for the simulated data

	k	≤ 6	7	8	9	10	11	12	13	14	15	16	17	>17
$r = 0$	$\mathbb{P}(K_n = k\|X)$	0.24	0.12	0.21	0.19	0.11	0.08	0.04	0.01					
$r = 3$	$\mathbb{P}(K_n = k\|X)$	0.06	0.08	0.15	0.18	0.17	0.14	0.10	0.06	0.03	0.01	0.01	0.01	
$r = 4$	$\mathbb{P}(K_n = k\|X)$	0.04	0.05	0.09	0.13	0.14	0.14	0.12	0.11	0.06	0.04	0.03	0.02	0.03

Table 4.3: Posterior probabilities of the number of distinct Y_i values from Fig. 4.2 for the galaxy data

	k	≤ 5	6	7	8	9	10	11	12	13	14	15	16	>16
$r=0$	$\mathbb{P}(K_n=k\|X)$	0.16	0.09	0.18	0.19	0.17	0.11	0.05	0.04	0.01				
$r=3$	$\mathbb{P}(K_n=k\|X)$	0.04	0.05	0.11	0.12	0.14	0.17	0.17	0.15	0.02	0.02	0.01		
$r=4$	$\mathbb{P}(K_n=k\|X)$	0.01	0.03	0.09	0.09	0.12	0.13	0.15	0.16	0.06	0.05	0.04	0.04	0.03

Table 4.4: Posterior probabilities of the number of distinct Y_i values from Figure 4.3 for the simulated data

	k	≤ 6	7	8	9	10	11	12	13	14
$r=0$	$\Pr(K_n=k\|X)$	0.05	0.14	0.22	0.20	0.15	0.12	0.09	0.03	
$r=3$	$\Pr(K_n=k\|X)$	0.02	0.16	0.23	0.20	0.16	0.12	0.08	0.03	
$r=4$	$\Pr(K_n=k\|X)$	0.02	0.15	0.22	0.21	0.14	0.14	0.10	0.01	0.01

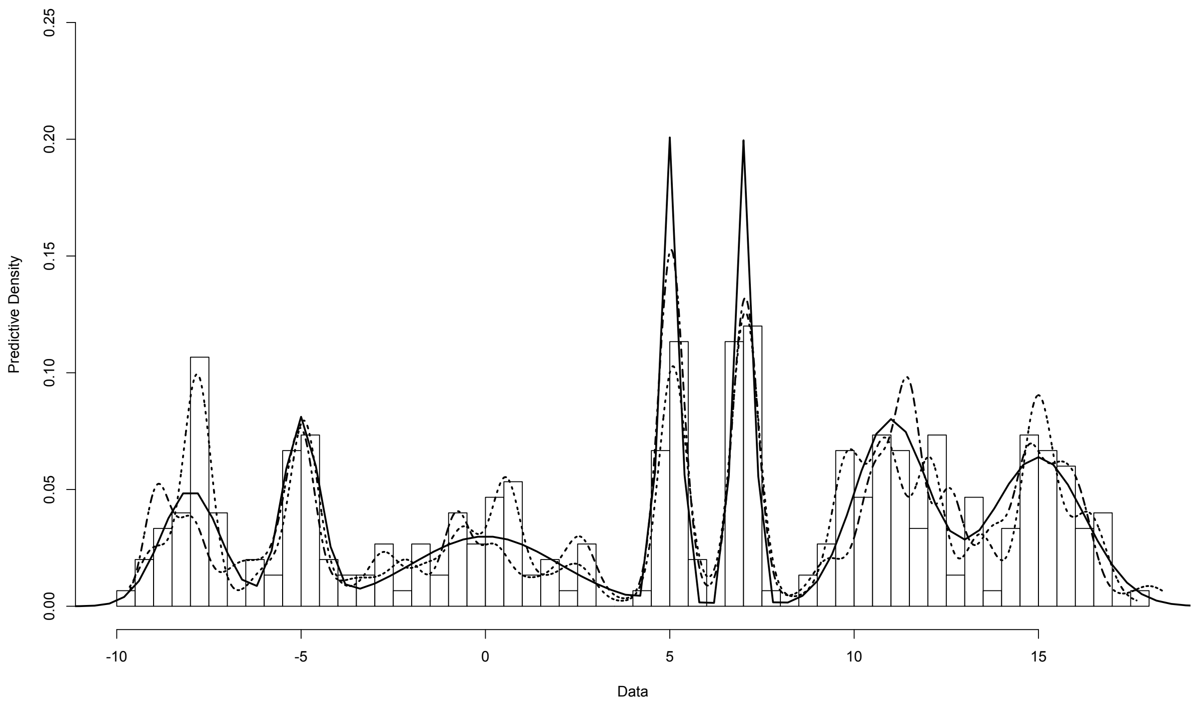

Figure 4.3: Posterior density estimates for the simulated data when we use DP and $P_{r,L,H}$ with L as the gamma Lévy measure as the prior: $\cdot\cdot\cdot\cdot\cdot$, DP,– – – –, $P_{r,L,H}$ with $r = 3$, ——, true model

Chapter 5

On a class of negative binomial processes and their applications

In the literature, the term 'negative binomial process' is employed to describe various processes, each playing a crucial role in stochastic processes and diverse statistical areas. These processes are especially prominent in applications within Bayesian nonparametric. However, ambiguity arises from the presence of conflicting definitions for negative binomial processes across different contexts in the literature. This chapter aims to clarify the confusion by systematically reviewing the various definitions of the negative binomial process and highlighting their distinctions. It provides a comprehensive account to help practitioners distinguish between them and avoid potential misunderstandings. Additionally, we extend the univariate negative binomial process from Chapter 3 to a bivariate form.

5.1 Introduction

The negative binomial process appears across diverse domains, including stochastic processes, statistics, machine learning, and computer science. Its versatility is reflected in the various definitions and representations it assumes. Originating from the negative binomial distribution, characterized by two parameters, the process takes on distinct forms when these parameters are substituted with positive finite measures, yielding alternative definitions.

The negative multinomial distribution denoted by $\mathrm{NMn}(r, \boldsymbol{p})$ is given by the probability distribution

$$P(\boldsymbol{X} = \boldsymbol{x}) = P(X_1 = x_1, \ldots, X_n = x_n) = \frac{\Gamma(r + \sum_{i=1}^{n} x_i)}{\Gamma(r) \prod_{i=1}^{n} x_i!} p_0^r \prod_{i=1}^{n} p_i^{x_i},$$

and probability generating function

$$G(s_1, \ldots, s_n) = \left(\frac{p_0}{1 - \sum_{i=1}^{n} p_i s_i} \right)^r, \quad (s_1, \ldots, s_n) \in [0, 1]^n,$$

where

$$r > 0; \quad \boldsymbol{p} = (p_0, p_1, \ldots, p_{n-1}), \; p_n = 1 - \sum_{i=0}^{n-1} p_i, \; p_i > 0, \; i = 0, 1, 2, \ldots, n;$$

$$x_i = 0, 1, 2, \ldots \quad i = 1, 2, \ldots, n.$$

See [64] for more on NMn. The univariate case of NMn, when $n = 1$, is the negative binomial distribution denoted by $\mathrm{NB}(r, p)$ with $p = p_0$.

[68] define a particular type of a negative binomial process by replacing the dispersion parameter r of a negative binomial distribution with a positive finite measure. In fact, in this case, it is not necessary to find the finite dimensional distributions of this process and only one dimensional negative binomial distribution will be sufficient. They also introduce a random sum representation for this process and use it in joint count and mixture modeling, applied to topic modeling problems and Poisson factor analysis. For other numerous applications see also [46] and references therein.

An entirely different definition for the negative binomial process is given in [23] by introducing this process as a mixed Poisson random measure. The finite dimensional distribution of this process follows a negative multinomial distribution with certain p_i's. Later, [27] employing trimmed subordinators, provide a point process representation for this process. Then, they use this representation to define a generalized Poisson-Kingman distribution and obtain its related Ewens' sampling formula to employ in a genetic diversity study in species sampling, see [30]. Another series representation, known as Ferguson and Klass [20] series representation, is our proposed point process (3.2.1) for this definition of the negative binomial process.

In the literature, various definitions and representations exist for the negative binomial process. Consequently, it is necessary to present a comprehensive review and clarification of these conflicting definitions and representations. This effort aims to reduce confusion among researchers, particularly in the fields of Bayesian nonparametric and stochastic processes.

5.2 The first definition for the negative binomial process

In the following definition, we provide the official description of the negative binomial process as utilized in [68] .

Definition 5.2.1. Let G be a Radon measure on $(\mathbb{E}, \mathscr{E})$ and $p \in (0, 1)$. Then the point process ξ is a negative binomial process, denoted by $\xi \sim \mathcal{NBP}^{(1)}(G, p)$, if $\xi(A) \sim \text{NB}(G(A), p)$ for a subset A of $\mathbb{E}$, and the number of points in disjoint subsets of $\mathbb{E}$ are independent random variables.

Remark 5.2.2. We can extend the Definition 5.2.1 by replacing p with a vector of probabilities $\boldsymbol{p} = (p_0, p_1, \dots, p_{n-1})$ and $p_n = 1 - \sum_{i=0}^{n-1} p_i$. In this case, we write $\xi \sim \mathcal{NBP}^{(1)}(G, \boldsymbol{p})$.

The negative binomial distribution is widely recognized as a gamma-Poisson (mixture) distribution. In this context, if we consider a Poisson distribution with a parameter λ, where λ is a random variable distributed according to a gamma distribution with shape parameter r and rate $p/(1-p)$, integrating λ out results in the negative binomial distribution $\text{NB}(r, p)$. Similarly, the negative binomial process outlined in Definition 5.2.1 is also characterized as a gamma-Poisson (mixture) process. The exploration of this association is presented in the subsequent theorem. However, before presenting the theorem, a brief setup is warranted.

If for a $c > 0$ and a base measure G_0 on $\mathbb{E}$, we take

$$L(x) = G_0(\mathbb{E}) \int_x^\infty u^{-1} e^{-cu} \mathrm{d}u$$

then the random measure given by

$$Q = \sum_{i=1}^{\infty} L^{-1}(\Gamma_i) \delta_{\omega_i},$$

defines a gamma process on $\mathbb{E} \times \mathbb{R}_+$ denoted by $\mathrm{GaP}(G_0, c)$. Also, the atoms ω_i's are i.i.d. from $G_0/G_0(\mathbb{E})$ and independent of Γ_i's. This implies that for disjoint sets $A_1, \ldots, A_k$, the random variables $\{Q(A_i)\}_{1 \le i \le k}$ are independent and $Q(A_i)$ has a gamma distribution with shape parameter $G_0(A_i)$ and scale parameter of $1/c$. Independence follows since Q is a pure jump Lévy process. See [34] for more details.

Furthermore, it is important to observe that, for consistency with our notations, the parameter p in [68] should be modified to $1-p$.

Theorem 5.2.3. For $p \in (0,1)$ and the base measure G_0 on $\mathbb{E}$, if $\zeta | G \sim \mathrm{PRM}(G)$ and $G|(G_0, p) \sim \mathrm{GaP}(G_0, p/(1-p))$, then $\zeta|(G_0, p) \sim \mathcal{NBP}^{(1)}(G_0, p)$.

Proof. For a positive measurable function f on $\mathbb{E}$, we derive the Laplace functional of ζ

$$\begin{aligned}
\Psi_\zeta(f) = E\left(e^{-\zeta(f)}\right) &= E\left(\exp\left\{-\int_{\mathbb{E}} (1 - e^{-f(x)}) G(\mathrm{d}x)\right\}\right) \\
&= E\left(\exp\left\{-\int_{\mathbb{E}} (1 - e^{-f(x)}) \sum_{i=1}^{\infty} L^{-1}(\Gamma_i) \delta_{\omega_i}(\mathrm{d}x)\right\}\right) \\
&= E\left(\exp\left\{-\sum_{i=1}^{\infty} (1 - e^{-f(\omega_i)}) L^{-1}(\Gamma_i)\right\}\right)
\end{aligned}$$

To evaluate the above expectation, we use the recursive method introduced in [5]. Therefore, let for any $t \ge 0$

$$\begin{aligned}
M(t) &= E\left(\exp\left\{-\sum_{i=1}^{\infty} (1 - e^{-f(\omega_i)}) L^{-1}(\Gamma_i + t)\right\}\right) \\
&= E\left[E\left(\exp\left\{-\sum_{i=1}^{\infty} (1 - e^{-f(\omega_i)}) L^{-1}(\Gamma_i + t)\right\} \middle| \Gamma_1 = u, \omega_1 = \omega\right)\right] \\
&= \int_{\mathbb{E}} \int_0^{\infty} \exp\left\{-(1 - e^{-f(\omega)}) L^{-1}(u + t)\right\} M(t+u) e^{-u} \mathrm{d}u \frac{G_0(\mathrm{d}\omega)}{G_0(\mathbb{E})}
\end{aligned}$$

Substitute $v = u + t$

$$e^{-t} M(t) = \int_{\mathbb{E}} \int_t^{\infty} \exp\left\{-(1 - e^{-f(\omega)}) L^{-1}(v)\right\} M(v) e^{-v} \mathrm{d}v \frac{G_0(\mathrm{d}\omega)}{G_0(\mathbb{E})}$$

Now, differentiate both sides with respect to t

$$-e^{-t} M(t) + e^{-t} \frac{\partial M(t)}{\partial t} = -\int_{\mathbb{E}} \exp\left\{-(1 - e^{-f(\omega)}) L^{-1}(t)\right\} M(t) e^{-t} \frac{G_0(\mathrm{d}\omega)}{G_0(\mathbb{E})}$$

$$M(t) = \exp\left[-\int_t^\infty \int_{\mathbb{E}} \left(1 - \exp\left\{-(1-e^{f(\omega)})L^{-1}(s)\right\}\right) \frac{G_0(\mathrm{d}\omega)}{G_0(\mathbb{E})}\mathrm{d}s\right]$$

Therefore, we have

$$\Psi_\zeta(f) = M(0) = \exp\left[-\int_{\mathbb{E}} \int_0^\infty \left(1 - \exp\left\{-(1-e^{f(\omega)})L^{-1}(s)\right\}\right) \frac{G_0(\mathrm{d}\omega)}{G_0(\mathbb{E})}\mathrm{d}s\right].$$

Now, take $z = L^{-1}(s)$ which also gives $\mathrm{d}s = L(\mathrm{d}z) = G_0(\mathbb{E})z^{-1}e^{-\frac{p}{1-p}z}\mathrm{d}z$ and then use the Frullani integral to get

$$\Psi_\zeta(f) = \exp\left[\int_{\mathbb{E}} \ln\left(\frac{p}{1-(1-p)e^{-f(\omega)}}\right) G_0(\mathrm{d}\omega)\right]. \tag{5.2.1}$$

∎

Remark 5.2.4. Instead of Laplace functional, the probability generating function is commonly used by authors. To derive the probability generating function, simply take $f(\omega) = -\ln\left(s^{I_A(\omega)}\right)$ in (5.2.1) for a subset A of $\mathbb{E}$. Hence,

$$\begin{aligned} E\left(s^{\zeta(A)}\right) &= \exp\left\{\int_{\mathbb{E}} \ln\left(\frac{p}{1-(1-p)e^{\ln(s^{I_A(\omega)})}}\right) G_0(\mathrm{d}\omega)\right\} \\ &= \exp\left\{\ln\left(\frac{p}{1-(1-p)s}\right)\int_A G_0(\mathrm{d}\omega)\right\} = \left(\frac{p}{1-(1-p)s}\right)^{G_0(A)}. \end{aligned}$$

In the forthcoming theorem, we derive the Laplace functional of the random sum representation introduced by [68] to determine its equivalence with Definition 5.2.1. First, we establish a clear understanding of logarithmic distribution. The multivariate logarithmic distribution denoted by Log($\boldsymbol{p}$) is given by the probability distribution

$$P(X_1 = x_1, \ldots, X_n = x_n) = -\frac{(\sum_{i=1}^n x_i - 1)!}{\ln(p_0)} \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!},$$

and the probability generating function

$$G(s_1, \ldots, s_n) = \frac{\ln(1 - \sum_{i=1}^n p_i s_i)}{\ln(p_0)}$$

where

$$\boldsymbol{p} = (p_0, p_1, \ldots, p_{n-1}),\ p_n = 1 - \sum_{i=0}^{n-1} p_i,\ p_i > 0,\ i = 0, 1, 2, \ldots, n;$$

$$x_i = 0, 1, 2, \dots \quad i = 1, 2, \dots, n, \quad \sum_{i=1}^{n} x_i = 1, 2, \dots .$$

The univariate case of the logarithmic distribution, when $n = 1$, is denoted by $\text{Log}(p)$ with $p = p_0$.

Theorem 5.2.5. For $p \in (0, 1)$ and the base measure G_0 on $\mathbb{E}$, let N follow a Poisson distribution with parameter $-G_0(\mathbb{E})\ln(p)$ and let $(Y_i)_{i\geq 1}$ be an i.i.d. sequence of random variables distributed as $\text{Log}(p)$. Also, independent of $(Y_i)_{i\geq 1}$, let $(\omega_i)_{i\geq 1}$ be an i.i.d. sequence sampled from $G_0/G_0(\mathbb{E})$. Then the random measure $\xi = \sum_{i=1}^{N} Y_i \delta_{\omega_i}$ follows $\mathcal{NBP}^{(1)}(G_0, p)$.

Proof. We use the Laplace functional to prove the theorem. The Laplace functional of ξ is

$$\begin{aligned}
\Psi_\xi(f) &= E\left(e^{-\xi(f)}\right) = E\left(e^{-\sum_{i=1}^{N} Y_i f(\omega_i)}\right) \\
&= E\left[E\left(e^{-\sum_{i=1}^{N} Y_i f(\omega_i)} | N\right)\right] = E\left(\left[E\left(e^{-Y_1 f(\omega_1)}\right)\right]^N\right).
\end{aligned}$$

However,

$$E\left(e^{-Y_1 f(\omega_1)}\right) = E\left[E\left(e^{-Y_1 f(\omega_1)} | Y_1\right)\right] = E\left(g(Y_1)\right)$$

where

$$g(Y_1) := E\left(e^{-Y_1 f(\omega_1)} | Y_1\right) = \int_{\mathbb{E}} e^{-Y_1 f(\omega_1)} \frac{G_0(\mathrm{d}\omega_1)}{G_0(\mathbb{E})}.$$

Therefore,

$$\begin{aligned}
\Psi_\xi(f) &= E\left(\left[E\left(e^{-Y_1 f(\omega_1)}\right)\right]^N\right) = E\left(\left[E\left(g(Y_1)\right)\right]^N\right) \\
&= \exp\left\{-G_0(\mathbb{E}) \ln(p) [E\left(g(Y_1)\right) - 1]\right\} \\
&= \exp\left\{G_0(\mathbb{E}) \ln(p) \sum_{y=1}^{\infty} (g(y) - 1) \frac{(1-p)^y}{y \ln(p)}\right\} \\
&= \exp\left\{G_0(\mathbb{E}) \sum_{y=1}^{\infty} \frac{(1-p)^y}{y} \int_{\mathbb{E}} \left(e^{-y f(\omega_1)} - 1\right) \frac{G_0(\mathrm{d}\omega_1)}{G_0(\mathbb{E})}\right\} \\
&= \exp\left\{\int_{\mathbb{E}} \left[\sum_{y=1}^{\infty} \frac{((1-p)e^{-f(\omega_1)})^y}{y} - \sum_{y=1}^{\infty} \frac{(1-p)^y}{y}\right] G_0(\mathrm{d}\omega_1)\right\} \\
&= \exp\left\{\int_{\mathbb{E}} \left[-\ln\left(1 - (1-p)e^{-f(\omega_1)}\right) + \ln(p)\right] G_0(\mathrm{d}\omega_1)\right\}
\end{aligned}$$

$$= \exp\left\{ \int_{\mathbb{E}} \ln\left(\frac{p}{1-(1-p)e^{-f(\omega_1)}} \right) G_0(\mathrm{d}\omega_1) \right\}.$$

which is the same as in (5.2.1). ∎

We can generalize the random sum ξ given in Theorem 5.2.5 by replacing p with a vector of probabilities $\boldsymbol{p} = (p_0, p_1, \ldots, p_{n-1})$ and $p_n = 1 - \sum_{i=0}^{n-1} p_i$. Let N follow a Poisson distribution with parameter $-G_0(\mathbb{E})\ln(p_0)$ and let $(Y_i)_{i\geq 1}$ be an i.i.d. sequence of random variables distributed as $\mathrm{Log}(\boldsymbol{p})$. Also, independent of $(Y_i)_{i\geq 1}$, let $(\omega_i)_{i\geq 1}$ be an i.i.d. sequence sampled from $G_0/G_0(\mathbb{E})$. Then the random measure $\xi = \sum_{i=1}^{N} Y_i \delta_{\omega_i}$ follows the negative binomial process introduced in Remark 5.2.2, i.e. $\xi \sim \mathcal{NBP}^{(1)}(G_0, \boldsymbol{p})$.

5.3 The alternative negative binomial process

As highlighted in [64], we can generate an NMn random variable through compounding of independent Poisson random variables with a gamma distribution. Let $X_1, \ldots, X_n | m$ be independent Poisson random variables,

$$P(X_1 = x_1, \ldots, X_n = x_n | m) = \prod_{i=1}^{n} e^{-m\lambda_i} (m\lambda_i)^{x_i} / x_i!.$$

Placing gamma prior on m as follows

$$f(m) = \frac{1}{a\Gamma(r)} \left(\frac{m}{a} \right)^{r-1} e^{-m/a}, \quad a, r > 0,$$

implies

$$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{\Gamma(r + \sum_{i=1}^{n} x_i)}{\Gamma(r) \prod_{i=1}^{n} x_i!} \left(\frac{1}{1 + a\sum_{i=1}^{n} \lambda_i} \right)^{r} \prod_{i=1}^{n} \left(\frac{a\lambda_i}{1 + a\sum_{i=1}^{n} \lambda_i} \right)^{x_i}.$$

Similar to above, a well-known definition of the negative binomial process is given in [23] in terms of a mixed Poisson random measure. For $r > 0$, let Γ_r be a $\mathrm{Gamma}(r, 1)$ random variable. Also, let ν be a Radon measure. Then the negative binomial process κ on $(\mathbb{M}, \mathscr{M})$ with parameters r and ν, denoted by $\mathcal{NBP}^{(2)}(r, \nu)$, is defined and characterized uniquely by a mixed PRM with a mixing randomized mean measure $\nu\Gamma_r$ with the following Laplace functional

$$\Psi_\kappa(f) = \int_0^\infty \exp\left\{ -\int_{\mathbb{E}} (1 - e^{-f(x)}) u\nu(\mathrm{d}x) \right\} P(\Gamma_r \in \mathrm{d}u)$$

$$= \left(1 + \int_{\mathbb{E}} (1 - e^{-f(x)})\nu(\mathrm{d}x)\right)^{-r}. \tag{5.3.1}$$

The terminology "negative binomial" arises from the finite-dimensional distribution associated with this process. If $B_1, \ldots, B_n$ are pairwise disjoint bounded Borel sets of $\mathbb{E}$, then the numbers of points of $\mathcal{NBP}^{(2)}(r,\nu)$ in $B_1, \ldots, B_n$ follow a negative multinomial distribution with parameters r and probabilities

$$p_0 = \frac{1}{1 + \nu(\bigcup_{i=1}^n B_i)}, \quad p_i = \frac{\nu(B_i)}{1 + \nu(\bigcup_{i=1}^n B_i)}, \quad i = 1, \ldots, n.$$

In Chapter 3, we presented two point process representations for $\mathcal{NBP}^{(2)}(r,\nu)$, one defined in (3.2.1) and the other discussed in Remark 3.2.2.

5.3.1 Asymptotic results when $r \to \infty$

An $\mathcal{NBP}^{(2)}(r,\nu)$ with $r = 1$ is typically called geometric process and we denote it by GeoP(ν). We can see that an $\mathcal{NBP}^{(2)}(r,\nu)$ is distributed as sum of r iid GeoP(ν). Therefore, we may write the following central limit theorem for $\mathcal{NBP}^{(2)}(r,\nu)$, when $r \to \infty$.

Theorem 5.3.1. If κ follows an $\mathcal{NBP}^{(2)}(r,\nu)$ and $(\eta_i)_{i\geq 1}$ be an iid sequence from GeoP(ν), then as $r \to \infty$, the set indexed process

$$\frac{\kappa - r\nu}{\sqrt{r}} \stackrel{d}{=} \frac{\sum_{i=1}^r \eta_i - rE(\eta_1)}{\sqrt{r}} \stackrel{d}{\to} G_\nu$$

where G_ν is a Gaussian process (a signed measure) with mean zero and covariance matrix with components

$$Cov(G_\nu(A_i), G_\nu(A_j)) = \nu(A_i \cap A_j) + \nu(A_i)\nu(A_j)$$

for any $A_i, A_j \in \mathscr{E}$.

The proof straightforwardly follows an argument similar to that in [70], Section 7.8.

For the sake of simplicity, we skip the discussion of the functional central limit theorem, as done by [70]. Note also that, using the Laplace functional of the negative binomial process and Poisson process, we can see that $\mathcal{NBP}^{(2)}(r,\nu/r)$ approaches to PRM(ν) when $r \to \infty$.

5.4 The bivariate negative binomial process

There are past efforts to generalize a univariate negative binomial random variable to a bivariate negative binomial random variable, see [14, 50, 44]. In [41], they use Lévy copulas to construct bivariate Lévy jump processes. In this section, with a different approach, we extend the univariate negative binomial process outlined in Section 5.3 to include a correlated bivariate negative binomial process. To that end, we first need to define a bivariate PRM.

Theorem 5.4.1. let $\{(U_{i,1}, U_{i,2})\}$ be an i.i.d. sequence of random vectors on $(0,\infty)\times(0,\infty)$, with joint distribution $H(\cdot,\cdot)$, constructed so as to be independent of $\{\Gamma_i\}$. Then the point process

$$\sum_{i=1}^{\infty} \delta_{(U_{i,1}\Gamma_i, U_{i,2}\Gamma_i)}$$

is a PRM with the following Laplace functional

$$\exp\left(-\int_0^\infty\int_0^\infty\int_0^\infty \left(1-e^{-f(u_1x,u_2x)}\right) dH(u_1,u_2)dx\right). \tag{5.4.1}$$

Similarly, for any Lévy measure ν_1 and ν_2 satisfying (2.1.4), the point process

$$^2\zeta = \sum_{i=1}^{\infty} \delta_{(\nu_1^{-1}(U_{i,1}\Gamma_i), \nu_2^{-1}(U_{i,2}\Gamma_i))}$$

is a PRM with the following Laplace functional

$$\exp\left(-\int_0^\infty\int_0^\infty\int_0^\infty \left(1-e^{-f(\nu_1^{-1}(u_1x),\nu_2^{-1}(u_2x))}\right) dH(u_1,u_2)dx\right). \tag{5.4.2}$$

Proof. We only prove (5.4.2). The proof of (5.4.1) follows analogously. For $t>0$, define

$$^2\zeta_t = \sum_{i=1}^{\infty} \delta_{\left(\nu_1^{-1}(U_{i,1}(\Gamma_i+t)), \nu_2^{-1}(U_{i,2}(\Gamma_i+t))\right)}.$$

Then,

$$\Psi_{^2\zeta_t}(f) = E(e^{-{}^2\zeta_t(f)}) = E\left(e^{-\sum_{i=1}^\infty f\left(\nu_1^{-1}(U_{i,1}(\Gamma_i+t)),\nu_2^{-1}(U_{i,2}(\Gamma_i+t))\right)}\right)$$

$$= E\left(E\left(e^{-\sum_{i=1}^\infty f\left(\nu_1^{-1}(U_{i,1}(\Gamma_i+t)),\nu_2^{-1}(U_{i,2}(\Gamma_i+t))\right)}|\Gamma_1=w, U_{1,1}=u_1, U_{1,2}=u_2\right)\right)$$

$$= \int_0^\infty\int_0^\infty\int_0^\infty e^{-f\left(\nu_1^{-1}(u_1(w+t)),\nu_2^{-1}(u_2(w+t))\right)}\Psi_{^2\zeta_{w+t}}(f)e^{-w}dH(u_1,u_2)dw.$$

Use the change of variable $x = w + t$ and differentiate both sides with respect to t. Now, applying a similar approach to the proof of Theorem 5.2.3 and solving the derived simple linear differential equation implies (5.4.2) (need to evaluate the solution at $t = 0$). ∎

The following theorem introduces a point process which follows a correlated bivariate negative binomial process with the required marginals.

Theorem 5.4.2. Let $\{(U_{i,1}, U_{i,2})\}$ be an i.i.d. sequence of random vectors on $(0,\infty)\times(0,\infty)$, with joint distribution $H(\cdot,\cdot)$, constructed so as to be independent of $\{\Gamma_i\}$. Then for any Lévy measure ν_1 and ν_2 satisfying (2.1.4), the point process

$$^2\kappa = \sum_{i=1}^{\infty} \delta_{\left(\nu_1^{-1}\left(\frac{U_{i,1}\Gamma_{r+i}}{\Gamma_r}\right), \nu_2^{-1}\left(\frac{U_{i,2}\Gamma_{r+i}}{\Gamma_r}\right)\right)}$$

follows a bivariate negative binomial process which we denote it by $\mathcal{BNBP}^{(2)}(r, \nu_1, \nu_2, H)$. The Laplace functional of $^2\kappa$ is given by

$$\left(1 + \int_1^\infty \int_0^\infty \int_0^\infty \left(1 - e^{-f(\nu_1^{-1}(u_1 x), \nu_2^{-1}(u_2 x))}\right) dH(u_1, u_2) dx\right)^{-r}.$$

Proof. First, note that $(^2\kappa|\Gamma_r = z)$ is a PRM with the following Laplace functional

$$\exp\left\{-z \int_1^\infty \int_0^\infty \int_0^\infty \left(1 - e^{-f(\nu_1^{-1}(u_1 x), \nu_2^{-1}(u_2 x))}\right) dH(u_1, u_2) dx\right\}.$$

Therefore,

$$\begin{aligned}
E(e^{-^2\kappa(f)}) &= E\left[E(e^{-^2\kappa(f)}|\Gamma_r = z)\right] \\
&= \int_{z>0} E(e^{-^2\kappa(f)}|\Gamma_r = z) P(\Gamma_r \in \mathrm{d}z) \\
&= \left(1 + \int_1^\infty \int_0^\infty \int_0^\infty \left(1 - e^{-f(\nu_1^{-1}(u_1 x), \nu_2^{-1}(u_2 x))}\right) dH(u_1, u_2) dx\right)^{-r}.
\end{aligned}$$

∎

The marginal process of a bivariate negative binomial process is itself a univariate negative binomial process, as stated in the following theorem. The proof relies on the invariance property of PRM (see [35] for details). Specifically, the invariance property states that if

ν is a positive and continuous Lévy measure satisfying (2.1.4), and if $\{U_i\}$ are iid positive random variables independent of $\{\Gamma_i\}$, then

$$\sum_{i=1}^{\infty} \delta_{\nu^{-1}(U_i\Gamma_i)} \stackrel{d}{=} \sum_{i=1}^{\infty} \delta_{\nu^{-1}\left(\frac{\Gamma_i}{E(U_1^{-1})}\right)}.$$

Theorem 5.4.3. Suppose that ν is a positive and continuous Lévy measure satisfying (2.1.4). If $\{U_i\}$ are iid positive random variables independent of $\{\Gamma_i\}$, then

$$\sum_{i=1}^{\infty} \delta_{\nu^{-1}\left(\frac{U_i\Gamma_{r+i}}{\Gamma_r}\right)} \stackrel{d}{=} \sum_{i=1}^{\infty} \delta_{\nu^{-1}\left(\frac{\Gamma_{r+i}}{E(U_1^{-1})\Gamma_r}\right)} \sim \mathcal{NBP}^{(2)}(r, E(U_1^{-1})\nu). \tag{5.4.3}$$

Proof. Note that, conditional on $\Gamma_r = w$, the left and right hand side point processes in (5.4.3) have the same distribution using the invariance property of PRM. They both are PRM with mean measure $wE(U_1^{-1})\nu$ on $(0, \nu^{-1}(1))$. Now, if we denote the left hand side point process in (5.4.3) by τ, we can write its Laplace functional as follows

$$\begin{aligned}
E(e^{-\tau(f)}) &= E\left[E(e^{-\tau(f)}|\Gamma_r = w)\right] \\
&= \int_{w>0} E(e^{-\tau(f)}|\Gamma_r = w)P(\Gamma_r \in \mathrm{d}w) \\
&= \int_{w>0} \exp\left\{-w\int_0^{\nu^{-1}(1)} (1 - e^{-f(x)})E(U_1^{-1})\nu(\mathrm{d}x)\right\} \frac{w^{r-1}e^{-w}}{\Gamma(r)}\mathrm{d}w \\
&= \left(1 + \int_0^{\nu^{-1}(1)} (1 - e^{-f(x)})E(U_1^{-1})\nu(\mathrm{d}x)\right)^{-r}
\end{aligned}$$

which is the Laplace functional of $\mathcal{NBP}^{(2)}(r, E(U_1^{-1})\nu)$. ∎

If ν_1 and ν_2 are selected to be gamma Lévy measures as defined in (2.2.6) but with different parameters θ, and if we let $(Y_i)_{i\geq 1}$ be an i.i.d. sequence of random variables taking values in $\mathbb{E}$ with a common distribution P_0, and further assume these are independent of the Γ_i's and the pairs $\{(U_{i,1}, U_{i,2})\}$, then we can generalize the bivariate gamma process as defined in [35] by

$$\boldsymbol{G}^{(r)}(\cdot) = (G_1^{(r)}(\cdot), G_2^{(r)}(\cdot)) = \sum_{i=1}^{\infty} \left(\nu_1^{-1}\left(\frac{U_{i,1}\Gamma_{r+i}}{\Gamma_r}\right), \nu_2^{-1}\left(\frac{U_{i,2}\Gamma_{r+i}}{\Gamma_r}\right)\right)\delta_{Y_i}(\cdot).$$

In other words, by setting $r = 0$ and $\Gamma_0 = 1$, we recover the bivariate gamma process as defined in [35]. In this case, the Dirichlet bi-measure defined in [35] which was applied to an image enhancement problem using blocked Gibbs sampling, is extended accordingly by

$$(\mathscr{P}_1^{(r)}(\cdot), \mathscr{P}_2^{(r)}(\cdot)) = \left(\frac{G_1^{(r)}(\cdot)}{G_1^{(r)}(\mathbb{E})}, \frac{G_2^{(r)}(\cdot)}{G_2^{(r)}(\mathbb{E})} \right).$$

Chapter 6

The Liouville process

The Liouville distribution is a well-known conjugate prior for the multinomial distribution. In this chapter, we introduce a discrete random probability measure constructed from a random vector following a Liouville distribution and subsequently derive its weak limit to define our proposed Liouville process. The resulting process is a spike-and-slab process, where the Dirichlet process serves as the slab and a single point from its mean acts as the spike. These two components are linearly combined using a random weight governed by the Liouville distribution. By using the Liouville process as a prior on the space of probability measures, we derive the corresponding posterior process as well as the predictive distribution.

6.1 Introduction

The Dirichlet distribution is widely recognized as a popular prior for the multinomial distribution. If the vector of proportions $(p_1, \ldots, p_{n+1})$ has Dirichlet distribution with density function (2.2.1), then we can write

$$p_1 = \pi_1, \quad p_i = \pi_i \prod_{j=1}^{i-1} (1 - \pi_j), \quad i = 2, \ldots, n+1,$$

where π_i's are independent and $\pi_i \sim Beta(a_i, \sum_{j=i+1}^{n+1} a_j)$ for $i = 1, \ldots, n$ and $\pi_{n+1} = 1$.

When the proportions p_i's represent failure times, imposing independent priors on the hazards π_i's (as done by the Dirichlet prior) often does not realistically reflect prior beliefs about the hazard function in many real-world cases. For further discussion, see [1], [10], and [40] and references therein.

Another limitation of the Dirichlet distribution is that it always imposes a negative correlation structure on the proportions, regardless of the values of the hyperparameters. In cases such as failure time analysis, where the sample space is the positive real line, it is quite plausible that there could be a positive correlation between the probabilities of failure in intervals that are close to each other. For further discussion, see [12, p. 834], and [57].

The Liouville distribution (see also [25]) serves as an alternative prior that not only addresses the limitations of the Dirichlet prior but also offers a broader class of conjugate priors for the multinomial distribution.

A random vector $(q_1, \ldots, q_{n+1})$ has a Liouville distribution denoted by $(q_1, \ldots, q_n) \sim \mathcal{L}(\psi, a_1, \ldots, a_n)$ if

$$f(q_1, \ldots, q_n) = A \prod_{i=1}^{n} q_i^{a_i - 1} \psi \left(\sum_{i=1}^{n} q_i \right) I_{\mathbb{S}_n}(q_1, \ldots, q_n),$$

where $q_1 + \cdots + q_{n+1} = 1$, $a_i > 0$ for all $i = 1, 2, \ldots, n$, and $\psi : (0, 1) \to (0, \infty)$ is a continuous function such that

$$\int_0^1 \psi(t) t^{\sum_{i=1}^n a_i - 1} \, dt < \infty.$$

Also,

$$A = A(\psi, a_1, \ldots, a_n) = \left(\frac{\prod_{i=1}^n \Gamma(a_i)}{\Gamma\left(\sum_{i=1}^n a_i\right)} \int_0^1 \psi(u) u^{\sum_{i=1}^n a_i - 1} \, du \right)^{-1}.$$

It is easily seen that $\psi(t) = (1 - t)^{a_{n+1} - 1}, 0 < t < 1, a_{n+1} > 0$, represents the Dirichlet distribution (Dirichlet distribution is a member of the Liouville family). There is also another type of Liouville family corresponding to Inverted Dirichlet distribution which is not of our concern here.

Additionally, for each $i \in \{1, \ldots, n\}$,

$$E(q_i) = \frac{A(\psi, a_1, \ldots, a_n)}{A(\psi, a_1, \ldots, a_i + 1, \ldots, a_n)} = \frac{a_i}{\sum_{j=1}^n a_j} \frac{\int_0^1 \psi(t) t^{\sum_{j=1}^n a_j} dt}{\int_0^1 \psi(t) t^{\sum_{j=1}^n a_j - 1} dt}.$$

Now, suppose that $(X_1, X_2, \ldots, X_n) \sim \text{Multi}(m, q_1, q_2, \ldots, q_n)$ where $m = \sum_{i=1}^{n} X_i$ and $(q_1, q_2, \ldots, q_{n-1}) \sim \mathcal{L}(\psi, a_1, \ldots, a_{n-1})$. Then for posterior distribution we have

$$(q_1, q_2, \ldots, q_{n-1})|(x_1, \ldots, x_n) \sim \mathcal{L}(\phi, a_1 + x_1, \ldots, a_{n-1} + x_{n-1})$$

where $\phi(t) = \psi(t)(1-t)^{x_n}$.

For an example of positive correlation, see [57], where for $n = 2$, they consider $\psi(u) = e^{\beta e^{-u}}$, with $u = q_1 + q_2$, $a_1 = 3$, and $a_2 = 5$. The covariance of q_1 and q_2 is calculated for values of β in $[-10, 100]$. It is observed that when β exceeds approximately 25, the covariance between q_1 and q_2 becomes positive.

It is important to note that the limitations discussed above for the Dirichlet distribution also extend to the Dirichlet process. In the next chapter, we develop the Liouville process as an alternative to the Dirichlet process to address these restrictions.

6.2 The Liouville process

In this section, we give a constructive definition of the Liouville process. The Liouville process will be a random convex combination of a Dirichlet process and a random point mass. Recall that the Dirichlet process has a stick-breaking representation. For $a > 0$ and probability measure H on the measurable space $(\mathbb{E}, \mathscr{E})$, the Dirichlet process $P \sim DP(a, H)$ can be written as follows

$$P(\cdot) = (1 - V_1)\delta_{Y_1}(\cdot) + \sum_{i=2}^{\infty}(1 - V_i)\prod_{j=1}^{i-1} V_j \delta_{Y_i}(\cdot),$$

where $V_i \overset{iid}{\sim} Beta(a, 1)$ is stochastically independent of $Y_i \overset{iid}{\sim} H$. Note that the Dirichlet process P is uniquely characterized by the following distributional equality

$$P \overset{d}{=} VP + (1 - V)\delta_Y, \tag{6.2.1}$$

where on the right hand side, V is independent of P, $V \sim Beta(a, 1)$, and $Y \sim H$. Now, we define the Liouville process as follows.

Definition 6.2.1. Let $P \sim DP(a, H)$, $V \sim \mathcal{L}(\psi, a)$, and $Y \sim H$, where V, P, and Y are mutually independent. Then, define

$$Q = VP + (1 - V)\delta_Y, \tag{6.2.2}$$

as a Liouville process with parameters ψ, a, H, and denote $Q \sim \mathcal{LP}(\psi, a, H)$.

Remark 6.2.2. As seen in Definition 6.2.1, the construction of the Liouville process Q still follows the stick-breaking approach. However, in this case, each stick length is scaled by a random variable V, which follows a Liouville distribution and is independent of the stick-breaking sequence V_i. As a consequence, the resulting stick lengths in the Liouville process become dependent.

Note that in the above definition, when $\psi(t) = 1$ for $t \in (0, 1)$, the random variable V is distributed as $Beta(a, 1)$ distribution, then Q will be a Dirichlet process as shown in (6.2.1). The intuition behind the definition of the Liouville process in equation (6.2.2) stems from a connection between the Liouville and Dirichlet distributions, as established in [25, Theorem 3.1], along with the finite-dimensional Dirichlet prior (2.2.4).

The following theorem from [25, Theorem 3.1] offers a representation of a general Liouville distribution in terms of a Dirichlet distribution and a univariate Liouville random variable.

Theorem 6.2.3. Let $(q_1, \ldots, q_n) \sim \mathcal{L}(\psi, a_1, \ldots, a_n)$. Then

$$(q_1, \ldots, q_n) \stackrel{d}{=} (p_1, \ldots, p_n)t_n,$$

where $(p_1, \ldots, p_n)$ and t_n are mutually independent, $(p_1, \ldots, p_n) \sim \mathrm{D}(a_1, \ldots, a_n)$, and $t_n \stackrel{d}{=} \sum_{i=1}^n q_i \sim \mathcal{L}(\psi, \sum_{i=1}^n a_i)$.

The following theorem provides a key theoretical justification for our proposed Liouville process, as introduced in Definition 6.2.1. Specifically, it shows that the Liouville process Q arises as the weak limit of a sequence of the point processes Q_n, defined below.

Theorem 6.2.4. Using the notations from Theorem 6.2.3, let $a_i = a/n, i = 1, \dots, n$, and q_i's are independent of $Y_i \overset{iid}{\sim} H$, $i = 0, 1, \dots, n$. Define

$$Q_n = \sum_{i=1}^{n} q_i \delta_{Y_i} + (1 - \sum_{i=1}^{n} q_i)\delta_{Y_0} \overset{d}{=} t_n \sum_{i=1}^{n} p_i \delta_{Y_i} + (1 - t_n)\delta_{Y_0} = t_n P_n + (1 - t_n)\delta_{Y_0}.$$

Then for each real-valued measurable function g, integrable with respect to H,

$$Q_n(g) = \int g \, \mathrm{d}Q_n \xrightarrow{d} Q(g) = \int g \, \mathrm{d}Q.$$

Proof. From [34], for a real valued measurable function g that is integrable with respect to H, we have

$$Q_n(g) = t_n P_n(g) + (1 - t_n)g(Y_0) \xrightarrow{d} Q(g) = tP(g) + (1 - t)g(Y)$$

where Q is as defined in Definition 6.2.1. Also, notice the fact that the distribution of $t_n = t$ is free from n and $t \overset{d}{=} V = V_a \sim \mathcal{L}(\psi, a)$ (V as defined in Definition 6.2.1) and $Y \overset{d}{=} Y_0$. ∎

To analyze the asymptotic behaviour of Q when $a \to \infty$, we calculate the mean and variance of $Q(A)$ for any measurable set $A \in \mathscr{E}$ here. The mean is given by

$$E(Q(A)) = E(V_a)H(A) + (1 - E(V_a))H(A) = H(A).$$

Moreover, the second moment of $Q(A)$ is given as follows

$$\begin{aligned}
E(Q^2(A)) &= E\left((V_a P(A) + (1 - V_a)\delta_Y(A))^2\right) \\
&= E[V_a^2 P^2(A) + (1 - V_a)^2 \delta_Y^2(A) + 2V_a(1 - V_a)P(A)\delta_Y(A)] \\
&= E(V_a^2)\frac{H(A)(1 + aH(A))}{1 + a} + E\left[(1 - V_a)^2\right] H(A) + 2E(V_a - V_a^2)H^2(A).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(Q(A)) &= E(Q^2(A)) - E^2(Q(A)) \\
&= \left(\frac{E(V_a^2)}{1 + a} + 1 + E(V_a^2) - 2E(V_a)\right) H(A) \\
&+ \left(\frac{aE(V_a^2)}{1 + a} + 2E(V_a) - 2E(V_a^2) - 1\right) H^2(A).
\end{aligned}$$

Notice that, as $a \to \infty$, for any continuous function ψ,

$$E(V_a^k) = \frac{\int_0^1 v^{a+k-1}\psi(v)\mathrm{d}v}{\int_0^1 v^{a-1}\psi(v)\mathrm{d}v} \longrightarrow 1. \tag{6.2.3}$$

To verify (6.2.3), we only need to check for ψ of the form $\psi(v) = v^m$, for some non-negative integer m, since ψ can be uniformly approximated by a polynomial on $[0,1]$ (Stone-Weierstrass approximation, see [61]). Therefore, $\mathrm{Var}(Q(A)) \to 0$ as $a \to \infty$. Consequently $Q(A)$ converges to $H(A)$ in probability as $a \to \infty$.

Remark 6.2.5. Assume $V_a \sim \mathcal{L}(\psi, a)$, with ψ be a continuous function free from a on $[0,1]$, then we have $V_a \xrightarrow{p} 1$ as $a \to \infty$. Therefore, $Q = Q_a$ approaches weakly to the Dirichlet process as $a \to \infty$. Since all asymptotic results are already established for the Dirichlet process, they can be easily extended to Q_a as $a \to \infty$. See [36] for more details.

6.3 Posterior and predictive distributions

In this section, we will study the posterior and predictive probability measures when the non-parametric prior is assumed to be the Liouville process $Q = VP + (1-V)\delta_Y$, introduced in Section 6.2. Thus, the goal is to derive both $Q|\boldsymbol{X}$ and $E(Q|\boldsymbol{X})$ where $X_1, \ldots, X_m$ be a random sample of size m from Q.

First, note that when sampling from the Liouville process Q, we are effectively drawing sample from the Dirichlet process P with probability V and from the mean measure of P, denoted by H, with probability $1-V$. We follow the same steps as in [32] to derive the posterior. Define the latent indicators $K_i \overset{iid}{\sim} \mathrm{Bernoulli}(V)$ for $i \in \{1, \ldots, m\}$, i.e.,

$$Pr(K_i = 0|V) = 1 - V, \text{ and } Pr(K_i = 1|V) = V.$$

Therefore, when $K_i = 1$, which occurs with probability $V \sim \mathcal{L}(\psi, a)$, we have $X_i \sim P$. Conversely, when $K_i = 0$, which occurs with probability $1-V$, $X_i = Y \sim H$.

If we define $W_0 = 1 - V$, $W_1 = V$, and $\boldsymbol{K} = (K_1, \ldots, K_m)$, then $(W_1|\boldsymbol{K}) \sim \mathcal{L}(\phi, a + m_1)$ where $\phi(t) = (1-t)^{m_0}\psi(t)$, $m_j = \sum_{i=1}^m I(K_i = j)$ for $j \in \{0,1\}$, where I denotes the indicator function.

Theorem 6.3.1. Suppose $X_1, \ldots, X_m$ is an iid sample from $Q = VP + (1-V)\delta_Y$ where $V \sim \mathcal{L}(\psi, a)$, $P \sim DP(a, H)$, and $Y \sim H$, with V, P, and Y mutually independent. Then, the posterior process Q given $\boldsymbol{X} = (X_1, \ldots, X_m)$ is characterized by

$$\mathscr{L}(Q|\boldsymbol{X}) = \sum_{\boldsymbol{K}} \mathscr{L}(Q|\boldsymbol{X}, \boldsymbol{K})\mathrm{Pr}(\boldsymbol{K}), \tag{6.3.1}$$

where the sum is over all $\boldsymbol{K}$, where

$$\mathrm{Pr}(\boldsymbol{K}) = \mathrm{Pr}(K_1)\prod_{i=2}^{m}\mathrm{Pr}(K_i|K_1, \ldots, K_{i-1}) = E(W_{K_1})\prod_{i=2}^{m}E(W_{K_i}|K_1, \ldots, K_{i-1}) \tag{6.3.2}$$

is the prior for $\boldsymbol{K}$ and

$$\mathscr{L}(Q|\boldsymbol{X}, \boldsymbol{K}) = \mathscr{L}\left\{V^*P^* + (1-V^*)\delta_{X^*}\right\}, \tag{6.3.3}$$

in which $V^* \stackrel{d}{=} (W_1|\boldsymbol{K})$ and P^* is the posterior of P based on the observations when $K_i = 1$. Therefore, $P^* \sim DP(a^*, H^*)$, where

$$a^* = a + m_1, \text{ and, } H^* = \frac{a}{a+m_1}H + \frac{1}{a+m_1}\sum_{i=1}^{m}K_i\delta_{X_i}.$$

In (6.3.3), X^* represents the observations from the Dirac probability measure δ_Y. When $m_0 = 0$, we can express δ_{X^*} as δ_Y.

Proof. The equations (6.3.1) and (6.3.2) follow directly from the definitions, so it remains to prove Equation (6.3.3). Conditional on $\boldsymbol{K}$, we know which observations originate from P and which from δ_Y, allowing the posterior to simplify to

$$Q \mid \boldsymbol{X}, \boldsymbol{K} \sim V^*P^* + (1-V^*)\delta_{Y^*},$$

where

$$V^* \stackrel{d}{=} V \mid \boldsymbol{K} \sim \mathcal{L}(\phi, a+m_1), \text{ with } \phi(t) = (1-t)^{m_0}\psi(t),$$
$$P^* \stackrel{d}{=} P \mid \boldsymbol{X}, \boldsymbol{K} \sim \mathrm{DP}\left(a+m_1, \frac{aH + \sum_{i=1}^{m}K_i\delta_{X_i}}{a+m_1}\right),$$
$$Y^* \stackrel{d}{=} Y \mid \boldsymbol{X}, \boldsymbol{K} \sim H(\cdot \mid \{X_i : K_i = 0\}).$$

This completes proof. ∎

Note that for any integer $i \geq 2$, $(W_1|K_1,\ldots,K_{i-1}) \sim \mathcal{L}(\phi, a+m^*_{1,i})$ where $\phi(t) = (1-t)^{m^*_{0,i}}\psi(t)$ and $m^*_{k,i} = \sum_{j=1}^{i-1} I(K_j = k)$ for $k \in \{0,1\}$. Taking

$$S(\psi,\alpha,\beta) = \int_0^1 \psi(t)t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t,$$

we get

$$E(W_k) = \frac{S(\psi, a+k, -k+2)}{S(\psi,a,1)},$$
$$E(W_k|\boldsymbol{K}) = \frac{S(\psi, a+m_1+k, m_0-k+2)}{S(\psi, a+m_1, m_0+1)},$$

and

$$E(W_{K_i}|K_1,\ldots,K_{i-1}) = \frac{S(\psi, a+m^*_{1,i}+K_i, m^*_{0,i}-K_i+2)}{S(\psi, a+m^*_{1,i}, m^*_{0,i}+1)}.$$

Therefore,

$$\begin{aligned}\Pr(\boldsymbol{K}) &= E(W_{K_1})\prod_{i=2}^{m} E(W_{K_i}|K_1,\ldots,K_{i-1})\\ &= \frac{S(\psi, a+K_1, -K_1+2)}{S(\psi,a,1)}\prod_{i=2}^{m}\frac{S(\psi, a+m^*_{1,i}+K_i, m^*_{0,i}-K_i+2)}{S(\psi, a+m^*_{1,i}, m^*_{0,i}+1)}.\end{aligned}$$

To derive the predictive probability measure, for a measurable set $A \in \mathscr{E}$, notice that

$$\begin{aligned}\Pr(X_{m+1}\in A \mid \boldsymbol{X}) &= E(Q(A)\mid \boldsymbol{X})\\ &= \sum_{\boldsymbol{K}} E(Q(A)\mid \boldsymbol{X},\boldsymbol{K})\Pr(\boldsymbol{K})\\ &= \sum_{\boldsymbol{K}} E\left(V^*P^*(A)+(1-V^*)\delta_{X^*}(A)\right)\Pr(\boldsymbol{K})\\ &= \sum_{\boldsymbol{K}} \left(E(V^*)E(P^*(A))+(1-E(V^*))E(\delta_{X^*}(A))\right)\Pr(\boldsymbol{K})\\ &= \sum_{\boldsymbol{K}}\left(\frac{S(\psi,a+m_1+1,m_0+1)}{S(\psi,a+m_1,m_0+1)}H^*(A)\right.\\ &\qquad\left.+\frac{S(\psi,a+m_1,m_0+2)}{S(\psi,a+m_1,m_0+1)}E(\delta_{X^*}(A))\right)\Pr(\boldsymbol{K})\end{aligned}$$

where $E(\delta_{X^*}(A)) = \delta_{X^*}(A)$ if $m_0 \neq 0$, and $E(\delta_{X^*}(A)) = H(A)$ if $m_0 = 0$.

Remark 6.3.2. We observe that the general form of the posterior distribution of Q is a mixture of mixtures. However, conditional on $\boldsymbol{K}$, the posterior of Q retains the same structural form as the prior.

Chapter 7

Conclusions and future work

7.1 Conclusions

One of our main contribution in Chapter 3 was to define a new family of priors in Bayesian nonparametric inference based on the negative binomial process. We first derived a new representation for the negative binomial process directly as a functional of the Poisson random measure. Then using this representation of the negative binomial process, we provided a family of generalized Poisson-Kingman distribution and its associated random discrete probability measure which contains many well-known priors in nonparametric Bayesian analysis such as the Dirichlet process, the Poisson-Dirichlet process, the normalized generalized gamma process, etc. A natural extension of the Dirichlet process was formulated as a functional of the proposed series representation for the negative binomial process. We also provided an almost sure convergent approximation for this extended Dirichlet process. Another by-product of our proposed series representation for the negative binomial process was a new series representation for the Poisson-Dirichlet process. It has been shown that an approximation based on this new representation for the Poisson-Dirichlet process is very efficient, as illustrated in a simulation study.

In Chapter 4, we demonstrated that the new family of priors introduced in Chapter 3 offers greater flexibility and yields more accurate predictive densities in Bayesian hierarchical models compared to the traditional Dirichlet process prior. To illustrate this, our simulation study employed both real-world and simulated datasets.

In the literature, the term "negative binomial process" is used to describe various processes, each playing a vital role in stochastic processes and different areas of statistics. These processes are particularly prominent in Bayesian nonparametric applications. However, ambiguity arises from the existence of conflicting definitions across different contexts. In Chapter 5, we addressed this confusion by systematically reviewing the various definitions of the negative binomial process and highlighting their key distinctions. Our aim was to provide a comprehensive overview that enables practitioners to recognize their differences and avoid potential misunderstandings. Furthermore, we extended the univariate negative binomial process introduced in Chapter 3 to a bivariate form.

In Chapter 6, after discussing some of the limitations of the Dirichlet distribution such as its tendency to impose negative correlations, we found that the Liouville distribution offers an effective alternative to address these shortcomings. We then introduce a discrete random probability measure built from a random vector following a Liouville distribution and then derived its weak limit to define the Liouville process. The resulting process was a spike-and-slab model, where the Dirichlet process acts as the slab and a single point from its mean represents the spike. These two components are then combined linearly with a random weight that follows a Liouville distribution. Finally, by placing the Liouville process as a prior over the space of probability measures, we derived both posterior and predictive distributions of this process.

7.2 Research extensions

The research contained in this thesis can be extended in various directions. Some of these are:

1. Develop a stick-breaking representation for the prior $P_{r,L,H}$ defined in (3.2.4) for Lévy measures L other than the gamma and α-stable cases.

2. Incorporate Metropolis-Hastings steps to sample from the posterior distributions of p and r in the block Gibbs sampler used in the simulation study in Chapter 4.

3. Implement the Pólya urn Gibbs sampler as an alternative to the block Gibbs sampler in the simulation study conducted in Chapter 4.

4. Derive the posterior and predictive distributions of the extended bivariate Dirichlet process obtained using the bivariate negative binomial process introduced in Chapter 5.

5. Apply the extended bivariate Dirichlet process, derived from the bivariate negative binomial process discussed in Chapter 5, to real-world applications.

6. Implement the Liouville process, defined in Chapter 6, in real-world applications.

Appendix A

Conditional distributions for the blocked Gibbs sampler

The conditional distributions of $Z, K, \mu, \sigma_Z, \sigma_X, p,$ and θ required for the blocked Gibbs sampler in Chapter 4, are provided below:

1. The conditional distribution of Z is

$$f(Z|K,\mu,\sigma_Z,X) \propto \left\{ \prod_{j=1}^{m} f(Z_{K_j^*}|\mu,\sigma_Z) \prod_{\{i:K_i=K_j^*\}} f(X_i|Z_{K_j^*}) \right\} f(Z^K|\mu,\sigma_Z).$$

   The first term corresponds to the product of conditional normal

$$(Z_{K_j^*}|K,\mu,\sigma_Z,X) \sim \mathcal{N}(\mu_j^*, \sigma_{Z_j}^*),$$

   where

$$\mu_j^* = \sigma_{Z_j}^* \left(\mu/\sigma_Z + \sum_{\{i:K_i=K_j^*\}} X_i/\sigma_X \right),$$

   $\sigma_{Z_j}^* = (n_j/\sigma_X + 1/\sigma_Z)^{-1}$, and $K_1^*, \ldots, K_m^*$ represent the unique set of K_i's and n_j is the number of times K_j^* occurs in K. Also, Z^K corresponds to those values in Z excluding $Z_K = (Z_{K_1^*}, \ldots, Z_{K_m^*})$.

2. The conditional distribution of K_i is

$$(K_i|p, Z, X) \sim \sum_{k=1}^{N} p^*_{k,i}\delta_k(\cdot),$$

where

$$(p^*_{1,i}, \ldots, p^*_{N,i}) \propto \left(p_1 \exp\left\{\frac{-1}{2\sigma_X}(X_i - Z_1)^2\right\}, \ldots, p_N \exp\left\{\frac{-1}{2\sigma_X}(X_i - Z_N)^2\right\}\right).$$

3. The conditional distribution of μ is

$$(\mu|Z, \sigma_Z, \sigma_\mu) \sim \mathcal{N}(\mu^*, \sigma^*_\mu),$$

where

$$\mu^* = \frac{\sigma^*_\mu}{\sigma_Z}\sum_{k=1}^{N} Z_k, \quad 1/\sigma^*_\mu = N/\sigma_Z + 1/\sigma_\mu.$$

4. The conditional distribution of σ_Z is

$$(\sigma_Z^{-1}|Z, \mu) \sim \text{Gamma}\left(\tau_1 + N/2, \tau_2 + \sum_{k=1}^{N}(Z_k - \mu)^2/2\right).$$

5. The conditional distribution of σ_X is

$$(\sigma_X^{-1}|X, Z, K) \sim \text{Gamma}\left(\gamma_1 + n/2, \gamma_2 + \sum_{i=1}^{n}(X_i - Z_{K_i})^2/2\right).$$

6. The conditional distribution of p is

$$(p|K, \theta) \sim \mathscr{G}(a^*_1, b^*_1, \ldots, a^*_{N-1}, b^*_{N-1}),$$

with

$$a^*_k = 1 + m_k, \quad b^*_k = \theta + \sum_{j=k+1}^{N} m_j, \quad k \in \{1, \ldots, N-1\}$$

where m_k is the number of K_i's which equal k.

7. The conditional distribution of θ is

$$(\theta|p) \sim \text{Gamma}\left(N + \nu_1 - 1, \nu_2 - \sum_{k=1}^{N-1}\log(1 - V_k)\right).$$

Appendix B

R programs

Here is the R codes used in Chapter 4.

```
rm(list = ls())
library(LaplacesDemon)
library(ggplot2)

 mix_pdf <- function(x, loc, scale, weights) {
   d <- rep(0, length(x))
   for (i in seq_along(loc)) {
     d <- d + weights[i] * dnorm(x, mean = loc[i], sd = scale)
   }
   return(d)
 }


 n<-82
x<-c(9.172,9.350,9.483,9.558,9.775,10.227,10.406,16.084,16.170,18.419,18.552,
18.600,18.927,19.052,19.070,19.330,19.343,19.349,19.440,
19.473,19.529,19.541,19.547,19.663,19.846,19.856,19.863,19.914,
19.918,19.973,19.989,20.166,20.175,20.179,20.196,20.215,20.221,20.415,
20.629,20.795,20.821,20.846,20.875,20.986,21.137,21.492,21.701,21.814,21.921,
```

```
21.960,22.185,22.209,22.242,22.249,22.314,22.374,22.495,
22.746,22.747,22.888,22.914,23.206,23.241,23.263,23.484,23.538,23.542,
23.666,23.706,23.711,
24.129,24.285,24.289,24.366,24.717,24.990,
25.633,26.960,26.995,32.065,32.789,34.279)

 grid <- seq(min(x), max(x), 0.001)

N<-50
#N<-150
tu1<-0.001
tu2<-0.001
s.teta<-100
nu1<-2
nu2<-4
gamma1<-0.001
gamma2<-0.001

n.sims<-300



mixDensity<-NULL
samp.K.after.burn<-NULL
number.of.clusters<-NULL
   for (s in 1:n.sims){

# initializing the parameters
n.burn<-1501
start.p<-rep(1/N,N)
#start.p<-as.vector(rdirichlet(1,rep(2,N)))
start.K<-sample(1:N, size = n, prob = start.p, replace=TRUE)
start.teta<-mean(rnorm(10,0,10))
```

```
start.sx<-mean(rinvgamma(10,shape=0.1,scale=0.1))
start.sz<-mean(rinvgamma(10,shape=0.1,scale=0.1))
start.Z<-rnorm(N,start.teta,(start.sz)^0.5)
start.alpha<-mean(rgamma(10,shape=2,scale=4))


     samp.p<-matrix(0,nrow=n.burn,ncol=N)
  samp.Z<-matrix(0,nrow=n.burn,ncol=N)
  samp.K<-matrix(0,nrow=n.burn,ncol=n)
  samp.teta<-c()
  samp.sx<-c()
  samp.sz<-c()
  samp.alpha<-c()
  samp.mixDensity<-matrix(0,nrow=n.burn,ncol=length(grid))

  samp.p[1,]<-start.p
  samp.Z[1,]<-start.Z
  samp.K[1,]<-start.K
  samp.teta[1]<-start.teta
  samp.sx[1]<-start.sx
  samp.sz[1]<-start.sz
  samp.alpha[1]<-start.alpha

  for (i in 2:n.burn){

   # Updating p
              mk<-c()
                 for (k in 1:N){
                     counter<-0
                     for (j in 1:n) {if (k== samp.K[i-1,j]) counter<-counter+1}
                         mk[k]<-counter
                                           }
```

```
          Mk<-rev(cumsum(rev(mk)))
          Mk<-Mk[-1]
          V<-c()
          V2<-c()
           for (t in 1:(N-1)) {V[t]<-rbeta(1,1+mk[t],(1*samp.alpha[i-1])+Mk[t])}
                    W<-cumprod(1-V)
                    W<-c(1,W)
                    V2<-c(V,1)
                    p<-V2*W
samp.p[i,]<-p

# Updating Z
  for (j in 1:N) {
    kseq.p <- which(samp.K[i-1,] == j)
    sigmaz.3 <- 1 / (length(kseq.p) / samp.sx[i-1] + 1 / samp.sz[i-1])
    th.3 <- sigmaz.3 * (sum(x[kseq.p]) / samp.sx[i-1] + samp.teta[i-1] / samp.sz[i-1])
    samp.Z[i,j] <- rnorm(1, th.3, sd = sqrt(sigmaz.3))
  }

# Updating K
  for (j in 1:n) {
    pr <- p * dnorm(x[j], mean = samp.Z[i,], sd = sqrt(samp.sx[i-1]))
    samp.K[i,j] <- sample(1:N, size = 1, prob = pr / sum(pr))
  }

 # Updating theta
  sigma.c3 <- 1 / (N / samp.sz[i-1] + 1 / s.teta)
  samp.teta[i] <- rnorm(1, sum(samp.Z[i,]) * sigma.c3 / samp.sz[i-1], sqrt(sigma.c3))

# Updating sigmaz
samp.sz[i]<-rinvgamma(1,tu1+N/2,tu2+0.5*sum((samp.Z[i,]-samp.teta[i])^2))
```

```
  # Updating sigmax
  samp.sx[i]<-rinvgamma(1,gamma1+n/2,gamma2+0.5*sum((x-samp.Z[i,][samp.K[i,]])^2))

  # Updating alpha
  samp.alpha[i]<-rgamma(1,N+nu1-1,nu2-sum(log(1-V)))
 }
 number.of.clusters[s]<-length(unique(samp.K[n.burn, ]))
mixDensity<-rbind(mixDensity, mix_pdf(grid, samp.Z[n.burn, ],
 sqrt(samp.sx[n.burn]), samp.p[n.burn, ]) )
  }
mixDensity<-apply(mixDensity,2,mean,na.rm=T)

 hist(x, breaks = 40, prob = TRUE, col="white" ,main = "")
 lines(grid, mixDensity, lwd = 1,lty="solid", ylim=c(0,0.9))
```

Bibliography

[1] J. Aitchison. Inverse distributions and independent gamma-distributed products of random variables. *Biometrika*, 50:505–509, 1963.

[2] L. Al-Labadi. On new constructive tools in bayesian nonparametric inference. Unpublished Ph.D. thesis, Department of Mathematics and Statistics, University of Ottawa, Ottawa, 2012.

[3] L. Al-Labadi and M. Zarepour. On simulations from the two-parameter poisson-dirichlet process and the normalized inverse-gaussian process. *Sankhyā*, 76-A(1):158–176, 2014.

[4] D. Applebaum. *Lévy processes and Stochastic Calculus.* Campridge University Press, Cambridge, 2004.

[5] D. Banjevic, H. Ishwaran, and M. Zarepour. A recursive method for functionals of poisson processes. *Bernoulli*, 8(3):295–311, 2002.

[6] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

[7] P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero. Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE transactions on pattern analysis and machine intelligence*, 37(2):212–229, 2013.

[8] T. Broderick, L. Mackey, J. Paisley, and M. I. Jordan. Combinatorial clustering and the beta negative binomial process. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):290–306, 2015.

[9] M. A. Carlton. Application of the two-parameter poisson-dirichlet distribution. Unpublished Ph.D. thesis, Department of Statistics, University of California, Los Angeles, 1999.

[10] R. J. Connor and J. E. Mosimann. Concepts of independent for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64:194–206, 1969.

[11] R. Cont and P. Tankov. *Financial Modelling with Jump Processes*. Chapman and Hall, New York, first edition, 2004.

[12] J. J. Deely and D. V. Lindley. Bayes empirical bayes. *Journal of the American Statistical Association*, 76(376):833–841, 1981.

[13] K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2:183–201, 1974.

[14] C. B. Edwards and J. Gurland. A class of distributions applicable to accidents. *Journal of the American Statistical Association*, 56:503–503, 1961.

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

[16] S. Favaro, A. Lijoi, R. Mena, and I. Prünster. Bayesian nonparametric inference for species variety with a two parameter poisson-dirichlet process prior. *Journal of the Royal Statistical Society Series B*, 71:993–1008, 2009.

[17] S. Feng. *The Poisson-Dirichlet Distribution and Related Topics*. Springer-Verlag, New York, 2010.

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

[19] T. S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2:615–629, 1974.

[20] T. S. Ferguson and M. J. Klass. A representation of independent increment processes without gaussian components. *Annals of Mathematical Statistics*, 43:1634–1643, 1972.

[21] T. S. Ferguson and E. G. Phadia. Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, 7:163–186, 1979.

[22] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer, New York, 2003.

[23] G. Gregoire. Negative binomial distributions for point processes. *Stochastic Process and their Applications*, 16(2):179–188, 1984.

[24] R. C. Griffiths, R. A. Maller, and S. Shemehsavar. A gibbs sampling scheme for a generalised poisson-kingman class. *arXiv:2402.11563*, 2023.

[25] R. Gupta and Donald St. P. Richards. Multivariate liouville distributions. *Journal of Multivariate Analysis*, 23(2):233–256, 1987.

[26] N. L. Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics,*, 18:1259–1294, 1990.

[27] Y. F. Ipsen and R. A. Maller. Negative binomial construction of random discrete distributions on the infinite simplex. *Theory of Stochastic Processes*, 22:34–46, 2017.

[28] Y. F. Ipsen, R. A. Maller, and S. Shemehsavar. Limiting distributions of generalised poisson-dirichlet distributions based on negative binomial processes. *Journal of Theoretical Probability*, 33:1974–2000, 2020.

[29] Y. F. Ipsen, R. A. Maller, and S. Shemehsavar. A generalised dickman distribution and the number of species in a negative binomial process model. *Advances in applied probability*, 53(2):370–399, 2021.

[30] Y. F. Ipsen, S. Shemehsavar, and R. A. Maller. Species sampling models generated by negative binomial processes. Preprint. Available at https://arxiv.org/abs/1904.13046, 2018.

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

[32] H. Ishwaran and L. F. James. Some further developments for stick-breaking priors: Finite and infinite clustering and classification. *Sankhyā: The Indian Journal of Statistics*, 65(3):577–592, 2003.

[33] H. Ishwaran and M. Zarepour. Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

[34] H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the dirichlet process. *The Canadian Journal of Statistics*, 30:269–283, 2002.

[35] H. Ishwaran and M. Zarepour. Series representations for multivariate generalized gamma processes via a scale invariance principle. *Statistica Sinica*, 19(4):1665–1682, 2009.

[36] L. F. James. Large sample asymptotics for the two parameter poisson-dirichlet process. *In Pushing the limits of contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, Volume 3, Eds. Clarke, B. and Ghosal, S. IMS*, pages 187–199, 2008.

[37] L. F. James, A. Lijoi, and I. Prünster. Conjugacy as a distinctive feature of the dirichlet process. *Scandinavian Journal of Statistics*, 33(1):105–120, 2006.

[38] L. F. James, A. Lijoi, and I. Prünster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

[39] G. H. Jang, J. Lee, and S. Lee. Posterior consistency of species sampling priors. *Statistica Sinica*, 20:581–593, 2010.

[40] J. Kalbfleisch and R. Prentice. *The statistical analysis of failure time data.* John Wiley and Sons, New York, 1981.

[41] J. Kallsen and P. Tankov. Characterization of dependence of multidimensional lévy processes using lévy copulas. *Journal of Multivariate Analysis*, 97(7):1551–1572, 2006.

[42] J. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21:59–78, 1967.

[43] J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society. Series B*, 37(1):1–22, 1975.

[44] B. Kopociński. Multivariate negative binomial distributions generated by multivariate exponential distributions. *Applicationes Mathematicae*, 25(4):463–472, 1999.

[45] R. M. Korwar and M. Hollander. Contributions to the theory of dirichlet processes. *Annals of Probability*, 1:705–711, 1973.

[46] T. J. Kozubowski and K. Podgŕski. Distributional properties of the negative binomial lévy process. *Probability and Mathematical Statistics*, 29:43–71, 2009.

[47] A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse gaussian priors. *Journal of the American Statistical Association*, 100:1278–1291, 2005.

[48] A. Lijoi, R. H. Mena, and I. Prünster. Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society. Series B*, 69(4):715–740, 2007.

[49] A. Y. Lo. On a class of bayesian nonparametric estimates: I, density estimates. *The Annals of Statistics*, 12:351–357, 1984.

[50] A. W. Marshall and I. Olkin. A family of bivariate distributions generated by the bivariate bernoulli distribution. *Journal of the American Statistical Association*, 80(390):332–338, 1985.

[51] P. Muliere and L. Tardella. Approximating distributions of random functionals of ferguson-dirichlet prior. *The Canadian Journal of Statistics*, 53(2):370–399, 1998.

[52] A. Ongaro and C. Cattaneo. Discrete random probability measures: a general framework for nonparametric bayesian inference. *Statistics and Probability Letters*, 67:33–45, 2004.

[53] J. Pitman. Some developments of the blackwell–macqueen urn scheme. *In Statistics, Probability and Game Theory (T. S. Ferguson, L. S. Shapley and J. B. MacQueen, eds.)*, pages 245–267, 1996.

[54] J. Pitman. Poisson-kingman partitions. In *Lecture Notes-Monograph Series*, volume 40, pages 1–34. Institute of Mathematical Statistics, 2003.

[55] J. Pitman. Combinatorial stochastic processes. In *Lecture Notes in Mathematics*. Springer-Verlag, 2006.

[56] J. Pitman and M. Yor. The two-parameter poisson–dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[57] W. S. Rayens and C. Srinivasan. Dependence properties of generalized liouville distributions on the simplex. *Journal of the American Statistical Association*, 89(428):1465–1470, 1994.

[58] E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585, 2003.

[59] S. I. Resnick. *Extreme values, regular variation, and point processes*. Springer-Verlag, New York, 1987.

[60] K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85:617–624, 1990.

[61] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill, New York, 3rd edition, 1976.

[62] K. I. Sato. *Lévy processes and Infinitely Divisible Distributions*. Campridge University Press, Cambridge, 1999.

[63] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[64] M. Sibuya, I. Yoshimura, and R. Shimizu. Negative multinomial distribution. *Annals of the Institute of Statistical Mathematics*, 16:409–426, 1964.

[65] Y. W. Teh. A hierarchical bayesian language model based on pitman-yor processes. *In proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, 2006.

[66] S. Walker and P. Muliere. Beta-stacy processes and a generalisation of the polya-urn scheme. *The Annals of Statistics,*, 25:1762–1780, 1997.

[67] M. Zarepour and L. Al-Labadi. On a rapid simulation of the dirichlet process. *Statistics and Probability Letters*, 82:916–924, 2012.

[68] M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):307–320, 2015.

[69] M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and poisson factor analysis. *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 1462–1471, 2012.

[70] K. Ziegler. Functional central limit theorems for triangular arrays of function-indexed processes under uniformly integrable entropy conditions. *Journal of Multivariate Analysis*, 62:233–272, 1997.