Nonparametric Bayesian modelling in Machine Learning

Nada Habli

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in
Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Nonparametric Bayesian inference has widespread applications in statistics and machine learning. In this thesis, we examine the most popular priors used in Bayesian non-parametric inference. The Dirichlet process and its extensions are priors on an infinite-dimensional space. Originally introduced by Ferguson (1983), its conjugacy property allows a tractable posterior inference which has lately given rise to a significant developments in applications related to machine learning. Another yet widespread prior used in nonparametric Bayesian inference is the Beta process and its extensions. It has originally been introduced by Hjort (1990) for applications in survival analysis. It is a prior on the space of cumulative hazard functions and it has recently been widely used as a prior on an infinite dimensional space for latent feature models.

Our contribution in this thesis is to collect many diverse groups of nonparametric Bayesian tools and explore algorithms to sample from them. We also explore machinery behind the theory to apply and expose some distinguished features of these procedures. These tools can be used by practitioners in many applications.

Acknowledgements

There are many people who provided much support and guidance throughout the lengthy course of this thesis to whom I am thankful.

First and foremost I am deeply indebted to my supervisor Dr. Mahmoud Zarepour who guided and inspired my work. I would like to express my sincere gratitude to him for redirecting my interest of research and introducing me to machine learning. It turns out to be a perfect fit to my background study in Computer Science. I attribute the level of my Masters degree to his encouragement and effort and without him this thesis would not have been completed or written. It was a privilege to have studied and researched under the guidance of this word-class scientist and professor.

Besides my supervisor, I am grateful to Dr. Luai Al Labadi for his continuous support and advice. He was a past PhD student of Dr. Zarepour and most of his contributions are in the area of nonparametric Bayesian inference. I sincerely appreciate his support over the phone and the back and forth emails we shared together. Without his support I would not be able accomplish this much in my thesis.

To my dear colleagues in lab B03 of the department of Mathematics, in particular Maryam, Jo-Ann, Farid, Sheikh, Ervé, Rachid, Hicham, Jason and Ibrahim. I thank you for the opportunity of getting to know such broad-minded, wise and fun researchers and for your assistance throughout my studies. I have been blessed with a friendly and cheerful group of fellow students who provided a much needed form of escape from my studies and for helping me keep things in perspective.

I acknowledge and greatly appreciate the financial support from the University of Ottawa for the Admission Scholarship as well as the Swartzen Memorial Scholarship offered by the department of Mathematics.

Last but not least, I would like especially to thank my family which consists of my mother, my father, my brother and my sisters. My hard-working parents have sacrificed their lives for my siblings and me while providing unconditional love and care. I love them so much, and I would not have made it this far without them. My brother, Omar and his wife Irialis have been my greatest support during the difficult moment in my life and I love them dearly. They have given me their unequivocal support throughout for which my sincere expression of thanks likewise does not suffice. I know I always have my family to count on when times are rough.

Dedication

To my gorgeous lovely children, Maya, Adam, Jad and Issam.

I love you all deeply and dearly.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In most data analysis that involves statistical inference, we often observe some set of data where we wish to fit a statistical model to be able to infer about its characteristic. Such characteristic can be as simple as estimating the mean of the data or as complex as estimating its entire distribution. Regardless of the complexity of the information we want to extract from the data, we need to construct a statistical model that fits the data. This requires estimating a set of parameters that govern the underlying physical setting of the measured data. There exist two main and distinct approaches to tackle this problem, namely Frequentist and Bayesian statistics.

The Frequentist approach to statistics considers probability as a limiting long run frequency. In particular, data are a repeatable random sample where we believe that the underlying parameters remain constant (or fixed) during this repeatable process. On the other hand, the Bayesian approach to statistics considers the parameters ϕ as being random, hence they are assigned a prior distribution $p(\phi)$. The observed data X is then used to update our prior belief for each unknown parameter via the Bayes rule $p(\phi|X) \propto p(X|\phi)p(\phi)$, where $p(\phi|X)$ is known as the posterior distribution. A parametric Bayesian inference is used when the set of parameters governing the data is finite. However, this could be restrictive as a model when we observe more and

more data. We want to have a model that grows in complexity when we observe more data. One way to overcome this problem is to use a non-parametric approach. In the Bayesian framework, this approach allows us to put a prior on an infinite dimensional parameter space. The choice of a prior distribution has been carefully and widely discussed in Bayesian inference. The main reason is that we need to construct a prior which will lead to predictive models such that we know how to sample from the prior and posterior distribution. The most practical and useful priors are conjugate priors. We say that the prior is a conjugate prior for the likelihood if the posterior has the same distributional form as the prior distribution. The conjugacy property is very useful from a computation point of view because it will be straight forward to sample from the posterior.

In this thesis, we highlight many diverse groups of nonparametric Bayesian priors and explore algorithms to sample from these them. The outline of this thesis is as follows: In Chapter 2 we review some preliminary theory for a Lévy random variable and its characteristic function. We introduce an example of Lévy random variables along with their Lévy measures. In Chapter 3 we introduce two well known priors; the Gamma process and the Dirichlet process. The former has been recently applied to exchangeable models of sparse graphs in Caron & Fox (2014) and for non-parametric ranking models in Caron et al. (2013). It also has been used as a prior for infinite-dimensional latent indicator matrices in Titsias (2008). The latter application is one of the earliest Bayesian non-parametric approach to infer on latent (or hidden) feature models, in particular when each feature occurs multiple times for a data point, in contrast of being simply binary. The Dirichlet process is a prior on an infinite-dimensional space. It has commonly been used as prior on latent class models, in particular in clustering and mixture models. For more details refer to Teh and Jordan (2013) and Teh (2010). In Chapter 4, we describe in detail the two-parameter Poisson Dirichlet process and the normalized inverse Gaussian process. The former is also known as the Pitman-Yor process and is a generalization of the Dirichlet process.

Unlike the Dirichlet and Beta process, the two-parameter Dirichlet process does not have the conjugacy property. We present the main contribution of this thesis in Chapter 5. We first introduce a different approximation of the Beta process along with algorithms to sample from it. We then introduce an extension to the Beta process known as the Beta-Bernoulli process. We describe two methodologies to sample from the Beta-Bernoulli process where one of them has been known in the Computer Science community by the Indian buffet process (IBP) (Ghahramani & Griffiths, 2005) and the other method is our contribution in this thesis. We describe a new sampling technique which focuses on the accuracy and efficiency of the approximated Beta-Bernoulli process. Finally, we present the sampling technique of the Beta-Bernoulli process on a simulated example.

Chapter 2

Lévy random variables and processes

2.1 Lévy process

In this chapter, we present some preliminary discussion on Lévy processes and their applications in nonparametric Bayesian inference.

Definition 2.1.1 (Lévy process) *A stochastic process* $X = \{X_t : t \geq 0\}$ *defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is called a* Lévy process *if the following properties hold:*

1. *The paths of* X *are* $\mathbb{P}$*-right continuous with left limits;*
2. X_t *starts at* 0*, i.e.,* $\mathbb{P}(X_0 = 0) = 1$ *a.s.*
3. X_t *has independent increments, i.e., the random variables* $X_{t_0}, X_{t_1} - X_{t_0}$*, …,* $X_{t_n} - X_{t_{n-1}}$ *are independent for all* $0 \leq t_0 < t_1 < \cdots < t_n$*, for* $n \geq 1$*;*
4. X_t *has stationary increments, i.e.,* $X_{t+s} - X_t \stackrel{d}{=} X_s$ *for all* $s, t \geq 0$*; and*

5. *X_t is stochastically continuous, i.e., for all $s \geq 0$, $X_t \xrightarrow{p} X_s$ as $t \to s$, or equivalently*

$$\lim_{t \to s} Pr\{|X_t - X_s| > \epsilon\} = 0$$

.

Throughout this thesis, " $\stackrel{d}{=}$ " and " $\xrightarrow{p}$ " denote equality in distribution and convergence in probability, respectively. Note that the fifth property does not imply that the sample paths are continuous.

Definition 2.1.2 (Subordinators) *Let $X = \{X_t : t \geq 0\}$ be a Lévy process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then X is called a* subordinator *if the following properties hold:*

1. *X_t is a Lévy process defined on $\mathbb{R}^+$;*
2. *X_t is a.s. non-negative; and*
3. *X_t is a.s. non-decreasing.*

2.2 Characteristic function

2.2.1 Infinitely divisibility

Definition 2.2.1 (Infinitely Divisible) *A real-valued random variable Θ has an* infinitely divisible *distribution if for each $n = 1, 2, \ldots$ there exist a sequence of i.i.d. random variable $\Theta_{1,n}, \ldots, \Theta_{n,n}$ such that*

$$\Theta \stackrel{d}{=} \Theta_{1,n} + \cdots + \Theta_{n,n}.$$

Based on this definition, one way to determine whether a given random variable has an infinitely divisible distribution is by checking its characteristic exponent. Assuming Θ has the characteristic exponent $\psi(u) := -\log \mathbb{E}(e^{iu\Theta})$ for all $u \in \mathbb{R}$, then

Θ is an infinitely divisible distribution if, for all $n \geq 1$, there exist a characteristic exponent of a probability distribution ψ_n such that $\psi(u) = n\psi_n(u)$ for all $u \in \mathbb{R}$.

From the definition of a Lévy process, we conclude that for any $t > 0$, X_t is a random variable belonging to the class of infinitely divisible distributions. This follows from the fact that X_t has stationary independent increments and therefore for any $n = 1, 2, \ldots,$

$$X_t \stackrel{d}{=} X_{t/n} + (X_{2t/n} - X_{t/n}) + \cdots + (X_t - X_{(n-1)t/n}).$$

Suppose now that for all $u \geq 0$ and $t \geq 0$, we define

$$\psi_t(u) = -\log \mathbb{E}(e^{iuX_t}). \tag{2.2.1}$$

Then, using (2.2.1), for any $m, n \geq 0$, we can easily get,

$$n\psi_1(u) = \psi_n(u) = m\psi_{n/m}(u).$$

Hence, for any $t > 0$ rational,

$$\psi_t(u) = t\psi_1(u), \tag{2.2.2}$$

where ψ_1 is the characteristic exponent of X_1. If t is a irrational, choosing a decreasing sequence of rational $\{t_n : n \geq 1\}$ such that $t_n \downarrow t$ as $n \to \infty$ along with the a.s. right continuity of X_t implies right continuity of $\exp\{-\psi_t(u)\}$. In which case, (2.2.2) holds for all $t \geq 0$.

2.2.2 Lévy-Khintchine definition

Theorem 1 (Lévy-Khintchine Theorem) *[Fristedt and Gray, 1996] There is a one-to-one correspondence between all infinitely divisible distributions (and therefore Lévy processes) X_t and the set of triples (a, σ, ν) where $a \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, and ν is a measure concentrated on $\mathbb{R}\backslash\{0\}$ satisfying $\int (1 \wedge x^2)\nu(dx) < \infty$, such that for all*

$$\mathbb{E}[e^{iuX_t}] = \exp(-\psi(u, t)),$$

where

$$\psi(u,t) = -iaut + t\frac{\sigma^2 u^2}{2} + t\int_{\mathbb{R}\setminus\{0\}} (1 - e^{iux} + itu1_{(|x|<1)})\nu(dx). \tag{2.2.3}$$

In (2.2.3) ν *is called the Lévy measure. For subordinator, there is a one-to-one correspondence between all infinitely divisible functions on* $\mathbb{R}^+$ *and pairs* (b,ν) *such that for all* t*, the Laplace transform of* X_t *is*

$$\mathbb{E}[e^{-uX_t}] = \exp[-but - t\int_0^\infty (1 - e^{-ux})\nu(dx)],$$

where $b \in \mathbb{R}^+$ *and* ν *is a measure on* $\mathbb{R}^+$ *satisfying* $\int (x \wedge 1)\nu(dx) < \infty$*.*

Here and throughout this thesis, we define $(x \wedge 1) := \min(x, 1)$.

2.2.3 Lévy-Itô definition

Theorem 2 (Lévy-Itô decomposition) *[Fristedt and Gray, 1996] Let* X_t *be a Lévy process on* $\mathbb{R}$ *with triple* (a,σ,ν) *as described in Definition 1. Let* (Y,W) *be an independent pair where* W *is standard Brownian motion and* Y *is a Poisson point process in* $(0,\infty)\times(\mathbb{R}\setminus\{0\})$ *whose intensity measure is* $\lambda\times\nu$ *where* λ *is the Lebesgue measure. Then, there exists a sequence of* $\epsilon_k \downarrow 0$ *such that*

$$X_t \stackrel{d}{=} at + \sigma W_t + \lim_{k\to\infty}\left[\int_{(-\infty,-\epsilon_k)\cup[\epsilon_k,\infty)} yY((0,t]\times dy) - \int_{(-\infty,-\epsilon_k]\cup[\epsilon_k,\infty)} y\nu(dy)\right].$$

When X_t *is a subordinator, then* ν *is a measure satisfying* $\int_0^\infty (x\wedge 1)\nu(dx) < \infty$*, and the Lévy-Itô decomposition simplifies to*

$$X_t \stackrel{d}{=} bt + \int yY((0,t]\times dy), \tag{2.2.4}$$

where the pair (b,ν) *is described in Definition 1 and* Y *is a Poisson process in* $(0,\infty)\times(0,\infty)$ *whose intensity measure is* $\lambda\times\nu$*.*

We reinterpret X_t in (2.2.4) to be $X(0,t]$, the measure assigned to the interval $(0,t]$. Since X_t can be constructed from a non-negative Poisson process, X is a random measure. In general, we define $X(A) = \int_A dX_t$ for any Borel set A.

Also, using the convention of the nonparametric Bayesian community, we redefine ν to be what was previously the product measure $\lambda \times \nu$. By doing so, we can allow λ to be different from the Lebesgue measure on some space Ω. If λ is not a multiple of Lebesgue measure, the resulting process will not be a Lévy process since it will not have stationary increments. Instead, it will defined to be a *completely random measure*.

Definition 2.2.2 (Completely random measure) *[Kingman, 1967] A random measure Θ is a* completely random measure *if, for any finite collection $A_1, \ldots, A_n$ of disjoint sets, the random variables $\Theta(A_1), \ldots, \Theta(A_n)$ are independent.*

Definition 2.2.3 (Random measure) *[Resnick, 1987] Let E be a Polish space and $\mathscr{B}(E)$ be the Borel σ-algebra generated by the open sets in E. A measure μ is called Radon if $\mu(K) < \infty$ for any compact set K in E. Let $M_+(E)$ be the space of Radon measures in $\mathbb{E}$. Let $\mathscr{M}_+(E)$ be the smallest σ-algebra of subsets of $M_+(E)$ making the maps $\mu \to \mu(f) = \int f(x)d\mu(x)$ from $M_+(E)$ to $\mathbb{R}$ measurable for all functions $f \in C_K^+(E)$, where $C_K^+(E)$ denotes the set of continuous functions $f : E \to [0, \infty)$ with compact support. Note that, $\mathscr{M}_+(E)$ is the Borel σ-algebra generated by the topology of vague convergence. A random measure on E is any measurable map ξ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(M_+(E), \mathscr{M}_+(E))$.*

Definition 2.2.4 (Point process) *Let E be a locally compact space with a countable basis. Let $\mathscr{E}$ be a Borel σ-algebra of subsets of E. Let $(\theta_i)_{i \geq 1}$ be a countable collection of not necessary distinct points of E. A point measure on E is a measure m of the following form:*

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

where, δ_{θ_i} denotes the Dirac measure at θ_i, i.e., $\delta_{\theta_i}(A) = 1$ if $\theta_i \in A$ and 0 otherwise for a set $A \in \mathscr{E}$. If $K \in \mathscr{E}$ is compact then $m(K) < \infty$ (i.e., m is Radon meaning the measure of compact sets is always finite). Take $M_p(E)$ as the space of all point

measures defined on $\mathbb{E}$ *and* $\mathscr{M}_p(E)$ *be the smallest* σ*-algebra containing all sets of the form* $\{m \in M_p(E) : m(A) \in B\}$ *for* $A \in \mathscr{E}$, $B \in \mathscr{B}([0,\infty))$. *Alternatively,* $\mathscr{M}_p(E)$ *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* E *is a measurable map from the probability space* $(\Omega, \mathcal{F}, \mathbb{P}) \to (M_p(E), \mathscr{M}_p(E))$. *Therefore, a point process is a random element of* $M_p(E)$. *The probability law of the point process* ξ *is the measure* $P \circ \xi^{-1} = P[\xi \in \cdot]$ *on* $\mathscr{M}_p(E)$.

The Laplace functional is a useful tool to determine the distribution of point processes. It is important to notice that the Laplace functional of a random measure uniquely determines the distribution of any random measure.

Definition 2.2.5 (Laplace Functional of Point Process) *Let* Q *be a probability measure on* $(M_p(E), \mathscr{M}_p(E))$ *(where* $M_p(E)$ *and* $\mathscr{M}_p(E)$ *are as constructed in Definition 2.2.3). The Laplace transform of* Q *is the map* ψ *which takes non-negative Borel functions on* E *into* $[0,\infty)$ *defined by*

$$\psi(f) = \int_{M_p(E)} \left(\exp\left\{ -\int_E f(x) m(dx) \right\} \right) Q(dm).$$

If $\xi : (\Omega, \mathcal{F}) \to (M_p(E), \mathscr{M}_p(E))$ *is a point process, the* Laplace functional *of* ξ *is the Laplace transform of the law of* $\xi(f)$*:*

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

Definition 2.2.6 (Poisson random measure) *Let* μ *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* μ, *denoted by* $PRM(\mu)$, *if* ξ *satisfies:*

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

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

2. *For any $k \geq 1$, if $A_1, \ldots, A_k$ are mutually disjoint sets in $\mathscr{E}$, then $\xi(A_1), \ldots, \xi(A_k)$ 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 $\mu(A)$ and the number of points in disjoint sets are independent random variables.

Proposition 1 *Let ξ a $PRM(\mu)$, the Laplace functional of $PRM(\mu)$ uniquely determines the law of ξ. It is given for any measurable positive function, $f \geq 0$, by*

$$\psi_\xi(f) = \exp\left\{ -\int_E (1 - e^{-f(x)})\mu(dx) \right\}.$$

For proof, see proposition 3.6 of Resnick (1987).

Proposition 2 *Let ξ_1 and ξ_2 be two independent Poisson random measures on $(E, \mathscr{B}(E))$ with mean measure μ_1 and μ_2 respectively. Then, the random measure $\xi = \xi_1 + \xi_2$ is also a Poisson random measure with mean measure $\mu = \mu_1 + \mu_2$.*

Proof:

$$\begin{aligned} E(e^{-\xi(f)}) &= E(e^{-\xi_1(f)-\xi_2(f)}) \\ &= E(e^{-\xi_1(f)})E(e^{-\xi_2(f)}) \\ &= \exp\left\{ -\int_E (1 - e^{-f(x)})\mu_1(dx) \right\} \exp\left\{ -\int_E (1 - e^{-f(x)})\mu_2(dx) \right\} \\ &= \exp\left\{ -\int_E (1 - e^{-f(x)})(\mu_1 + \mu_2)(dx) \right\}. \end{aligned}$$

∎

Throughout this thesis, we define

$$\Gamma_i = E_1 + \cdots + E_i, \tag{2.2.5}$$

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

Theorem 3 *Let $\xi \sim PRM(\lambda)$ where λ is the Lebesgue measure on $\mathbb{R}$. Then ξ can be written as follows*

$$\xi = \sum_{i=1}^{\infty} \delta_{\Gamma_i}.$$

Proof: Using the recursive technique in Banjevic & Zarepour (2002), we have for any non-negative function f

$$\begin{aligned}
\xi(f) &= \sum_{i=1}^{\infty} f(\Gamma_i + t) \\
M(u,t) &= E[e^{-u\xi}] = E(e^{-u\sum_{i=1}^{\infty} f(\Gamma_i+t)}) \\
&= E(E(e^{-u\sum_{i=1}^{\infty} f(\Gamma_i+t)}|\Gamma_1 = s)) \\
&= \int_t^{\infty} e^{-uf(s+t)} E(e^{-u\sum_{i=1}^{\infty} f(\Gamma_i+s+t)}) e^{-s} ds \\
&= \int_t^{\infty} e^{-uf(s+t)} M(u, t+s) e^{-s} ds.
\end{aligned}$$

Now using the change of variable $s + t = v$ and multiplying both sides by e^{-t}, we get

$$\begin{aligned}
M(u,t) &= \int_t^{\infty} e^{-uf(v)} M(u,v) e^{-(v-t)} dv \\
e^{-t} M(u,t) &= \int_t^{\infty} e^{-uf(v)} M(u,v) e^{-v} dv.
\end{aligned}$$

Differentiating both sides with respect to t, we get

$$\begin{aligned}
-e^{-t} M(u,t) + DM(u,t) e^{-t} &= -e^{-uf(t)} M(u,t) e^{-t} \\
DM(u,t) &= (1 - e^{-uf(t)}) M(u,t) \\
\frac{DM(u,t)}{M(u,t)} &= (1 - e^{-uf(t)}) \\
M(u,t) &= \exp\left(- \int_t^{\infty} (1 - e^{-uf(s)}) ds \right).
\end{aligned}$$

Now take $t = 0$ to get

$$M(u, 0) = \exp\bigg(- \int_0^\infty (1 - e^{-uf(s)})ds \bigg).$$

∎

2.2.4 Transformation of Poisson Random measure

The next proposition shows that mapping points of a Poisson point process yields new Poisson point process with a certain representation for its mean measure.

Proposition 2.2.7 (Proposition 3.7 of Resnick, 1987) *Let $\mathbb{E}_i$, $i = 1, 2$ be two locally compact spaces with countable bases. Let $\mathscr{E}_i$, $i = 1, 2$ be the associated σ-fields. Let $T : (\mathbb{E}_1, \mathscr{E}_1) \to (\mathbb{E}_2, \mathscr{E}_2)$ be measurable. If ξ is a $PRM(\mu)$ 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*

$$\xi = \sum_{i=1}^{\infty} \delta_{X_i}$$

then

$$\tilde{\xi} = \xi \circ T^{-1} = \sum_{i=1}^{\infty} \delta_{T(X_i)}.$$

The next proposition shows that starting from PRM, we may construct a new PRM whose points live in a higher dimensional space.

Proposition 2.2.8 (Proposition 3.8 of Resnick, 1987) *Let $\mathbb{E}_i$, $i = 1, 2$ be two locally compact spaces with countable bases. Suppose*

$$\sum_{i=1}^{\infty} \delta_{X_i}$$

is a PRM(μ) on $\mathbb{E}_1$. Suppose $(\theta_i)_{i\geq 1}$ are i.i.d. random elements on $\mathbb{E}_2$ with common probability distribution F, and suppose the Poisson process and $(\theta_i)_{i\geq 1}$ are defined on

the same probability space and are independent. Then the point process on $\mathbb{E}_1 \times \mathbb{E}_2$,

$$\sum_{i=1}^{\infty} \delta_{(X_i,\theta_i)},$$

is a PRM with mean measure $\mu \times F$.

Let $\xi = \sum_{i=1}^{\infty} \delta_{(\Gamma_i,\theta_i)}$ such that $\theta_i \overset{i.i.d.}{\sim} H$ and Γ_i is independent of θ_i then for A and B in Ω, we have

$$\mathbb{E}[e^{-u\xi(A\times B)}] = \exp\bigg(- \int_B \int_A (1 - e^{-ux})\nu(dx)dH(\theta)\bigg).$$

Theorem 4 (Banjevic & Zarepour, 2002) *Suppose* X *is a subordinator with characteristic function*

$$\Psi(x) = \exp\bigg\{ - \int_0^{\infty} (1 - \exp(ixu))d\nu(u)\bigg\}, \qquad -\infty < x < \infty$$

where ν *is a positive, continuous and non-increasing Lévy measure defined on* $(0, \infty)$ *such that* $\nu(x) = \int_x^{\infty} d\nu(u)$ *and*

$$\int_{\epsilon}^{\infty} \nu^{-1}(u)du < \infty, \qquad \text{for each } \epsilon > 0$$

in which $\nu^{-1}(u) = \sup\{x : \nu(x) \leq u\}$. *Then* X *has the almost sure representation*

$$X = \sum_{i=k}^{\infty} \nu^{-1}(\Gamma_k).$$

Theorem 5 (Campbell's Theorem) *[Kingman, 1993] Let* Y *be a Poisson process on* $\Omega \times (0, \infty)$ *with mean measure* ν. *Then* Y *has a sum representation*

$$Y = \sum_{i=1}^{\infty} x_i \delta_{\theta_i}.$$

We have that Y *is absolutely convergent if and only if*

$$\int_{\Omega} \int_0^{\infty} x.\nu(d\theta, dx) < \infty.$$

2.3 Lévy random variables

In this section, we introduce some examples of Lévy random variables and their characteristic.

2.3.1 Poisson random variable

Let ξ be a Poisson random variable with probability distribution defined as follows, for each $\lambda > 0$

$$\mu_\lambda(\{k\}) = e^{-\lambda}\lambda^k/k! \qquad k = 0, 1, 2, \ldots.$$

Using the definition of characteristic function, an easy calculation shows that

$$\begin{aligned}
\mathbb{E}[e^{iu\xi}] &= \sum_{k\geq 0} e^{iuk}\mu_\lambda(\{k\}) \\
&= \sum_{k\geq 0}[e^{iu}]^k e^{-\lambda}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{iu})^k}{k!} \\
&= e^{-\lambda}e^{\lambda e^{iu}} \\
&= e^{-\lambda(1-e^{iu})} \\
&= \left[e^{-\frac{\lambda}{n}(1-e^{iu})}\right]^n \qquad (2.3.1)
\end{aligned}$$

In here, (2.3.1) is the characteristic function of the sum of n independent Poisson processes, each of which with parameter λ/n. Therefore, the Lévy-Khintchine definition states that there exist a triple (a, σ, ν) such that the characteristic exponent $\psi(u)$ satisfies the equation in (2.2.3). Indeed, we see that the characteristic exponent of the Poisson random variable has the same form of (2.2.3) with $a = \sigma = 0$ and $\nu = \lambda\delta_1$, where δ_1 is the Dirac measure at 1. The Poisson random variable can be written as

$$\xi(\cdot) = \sum_{i=1}^{\infty}\delta_{\Gamma_i}(\cdot).$$

Refer to Theorem 3 for the proof.

2.3.2 Gamma random variable

Let X be a Gamma random variable with probability measure defined as follows, for $\alpha, \beta > 0$

$$\mu_{\alpha,\beta}(dx) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx$$

Using the characteristic function we get,

$$\begin{aligned} \mathbb{E}[e^{iuX}] &= \int_0^\infty e^{iux} \mu_{\alpha,\beta}(dx) \\ &= \frac{1}{(1 - iu/\beta)^\alpha} \\ &= \left(\frac{1}{(1 - iu/\beta)^{\alpha/n}} \right)^n \end{aligned} \tag{2.3.2}$$

From (2.3.2) we can conclude that a Gamma random variable is infinitely divisible and therefore the Lévy-Khintchine definition state that there exists a triple (a, σ, ν) such that $\psi(u)$ satisfies (2.2.3). The following helps us find the values of (a, σ, ν).

Theorem 6 (Frullani integral) *Let $a, b > 0$ such that $a < b$. If $f'(x)$ is continuous and the integral converges, then*

$$\int_0^\infty \frac{f(ax) - f(bx)}{x} dx = [f(0) - f(\infty)] \log\left(\frac{b}{a}\right).$$

Lemma 2.3.1 *For all $\alpha, \beta > 0$ and $z \in \mathbb{C}$ such that the real part of z is in $(-\infty, 0)$ we have*

$$\frac{1}{(1 - z/\beta)^\alpha} = \exp\left[- \int_0^\infty (1 - e^{zx}) \alpha x^{-1} e^{-\beta x} dx \right].$$

Proof: Using the Frullani integral with $f(x) = \alpha e^{-x}$, it follows

$$\exp\left\{ - \int_0^\infty (1 - e^{zx}) \alpha x^{-1} e^{-\beta x} dx \right\} = \exp\left\{ - \int_0^\infty \frac{\alpha e^{-\beta x} - \alpha e^{-(\beta - z)x}}{x} \right\}$$

$$
\begin{aligned}
&= \exp\left\{ -\int_0^\infty \frac{f(\beta x) - f((\beta - z)x)}{x} \right\} \\
&= \exp\left\{ -\left([f(0) - f(\infty)] \log\left(\frac{\beta - z}{\beta} \right) \right) \right\} \\
&= \exp\left\{ -\alpha \log\left(\frac{\beta - z}{\beta} \right) \right\} \\
&= \exp\left\{ \log\left(\frac{\beta}{\beta - z} \right)^\alpha \right\} \\
&= \left(\frac{1}{1 - z/\beta} \right)^\alpha
\end{aligned}
$$

∎

For simplicity and without loss of generality, in this thesis, we take $\beta = 1$. The characteristic exponent of a Gamma random variable X is

$$
\begin{aligned}
\psi(u) &= -\log \mathbb{E}(e^{iuX}) \\
&= -\log\left(\frac{1}{(1 - iu)^\alpha} \right) \\
&= -\log(e^{-\int_0^\infty (1 - e^{iux})\alpha x^{-1} e^{-x} dx}) \\
&= \alpha \int_0^\infty (1 - e^{iux}) \frac{1}{x} e^{-x} dx \quad \text{for } \theta \in \mathbb{R}.
\end{aligned}
$$

Therefore,

$$\sigma = 0,$$

and for ν concentrated on $(0, \infty)$, we have

$$\nu(dx) = \alpha x^{-1} e^{-x} dx \tag{2.3.3}$$

$$a = -\int_0^1 x\nu(dx) \tag{2.3.4}$$

The choice of a in the Lévy-Khintchine formula is the necessary quantity to cancel the term coming from $\int iu1_{(|x|<1)}\nu(dx)$ in (2.2.3). We can show that any Gamma random variable can be written as

$$X(\cdot) = \sum_{i=1}^\infty \nu^{-1}(\Gamma_i)\delta_{\theta_i}(\cdot).$$

See Ferguson and Klass (1972) and Banjevic & Zarepour (2002) for detail and proof.

2.3.3 Stable random variable

Definition 2.3.2 *A random variable X is said to have a stable distribution if for all $n \geq 2$, $a_n > 0$ and $b_n \in \mathbb{R}$,*

$$X_1 + X_2 + \cdots + X_n \stackrel{d}{=} a_n X + b_n, \tag{2.3.5}$$

where $X_1, X_2, \ldots, X_n$ are independent copies of X.

It can be proven that $a_n = n^{1/\alpha}$ for $0 < \alpha \leq 2$. The value α is known as the index of stability. Any stable random variable is infinitely divisible, this follows by subtracting b_n/n from each of the independent copies $X_1, X_2, \ldots, X_n$

$$X_1 + X_2 + \cdots + X_n \stackrel{d}{=} a_n X + b_n$$
$$\sum_{k=1}^{n} \left(\frac{X_k - b_n/n}{a_n} \right) \stackrel{d}{=} X$$

Definition 2.3.3 *A stable random variable denoted by $S_\alpha(c, \beta, \mu)$, with an index of stability $\alpha \in (0, 2]$, $c \geq 0$, $-1 \leq \beta \leq 1$, $\mu \in \mathbb{R}$ has a characteristic exponents as follows:*

$$\psi(u) = \begin{cases} -c^\alpha |u|^\alpha (1 - i\beta (sign\ u) \tan \frac{\pi\alpha}{2}) + i\mu u & \text{if } \alpha \neq 1 \\ c|u|(1 + i\beta \frac{2}{\pi} (sign\ u) \ln |u| + i\mu u & \text{if } \alpha = 1 \end{cases}$$

where

$$sign(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{if } u = 0 \\ -1 & \text{if } u < 0 \end{cases}$$

To make a connection with the Lévy-Khintchine formula, one should have $\sigma = 0$, $a = \mu - \int_{\mathbb{R}} x 1_{(|x|<1)} \nu(dx)$ in (2.2.3) and

$$\nu(dx) = \begin{cases} \frac{P}{x^{1+\alpha}} dx & \text{if } x \in (0, \infty) \\ \frac{Q}{|x|^{1+\alpha}} dx & \text{if } x \in (-\infty, 0) \end{cases}$$

where for $P, Q \geq 0$

$$c = P + Q,$$

and

$$\beta = \begin{cases} \frac{P-Q}{P+Q} & \text{if } \alpha \in (0,1) \cup (1,2) \\ 0 & \text{if } \alpha = 1 \quad (P = Q). \end{cases}$$

A random variable with symmetric α-stable distribution, denoted by X_α can be written as a series representation as follows, with $0 < \alpha < 2$:

$$X_\alpha(\cdot) = \sum_{i=1}^{\infty} \epsilon_i \Gamma_i^{-1/\alpha} \delta_{\theta_i}(\cdot),$$

where $\{\epsilon_i\}$ is an i.i.d. sequence with

$$p(\epsilon_i = 1) = 1/2$$
$$p(\epsilon_i = -1) = 1/2.$$

Chapter 3

Gamma and Dirichlet Process

3.1 Gamma Process

The Gamma process plays a crucial role in nonparametric Bayesian inference. It has gained widespread adoption when computational techniques allowed it to be more practically applicable. It has been used as prior to many applications in different fields such as exchangeable models of sparse graphs (Caron, François and Fox, Emily B, 2014), nonparametric ranking models (Caron, François and Teh, 2013) and infinite-dimensional latent indicator matrices (Titsias, Michalis L, 2008). A normalized Gamma process which will be introduced in future sections, called the Dirichlet process has played a central role in nonparametric Bayesian inference.

3.1.1 Definition of Gamma process

Definition 3.1.1 (Gamma process) *The Gamma process, denoted by* $\mathcal{G} \sim GP(a, H)$*, is a completely random measure on* $[0, \infty] \times \Omega$ *with Lévy measure*

$$\rho(dx, d\theta) = N(dx)dH(\theta),$$

where

$$N(dx) = a\frac{e^{-ax}}{x}dx.$$

Here, $a > 0$ is called the concentration parameter and H the base measure.

Recall from Chapter 2 that the Lévy measure of the Gamma random variable is defined as follows

$$N(x) = a\int_x^\infty u^{-1}e^{-u}du, \text{ for } x > 0. \tag{3.1.1}$$

Note that $N(dx) := dN(x)$.

The Gamma process is the conjugate prior for the non-negative integer valued-Poisson random measure. Suppose we observe m Poisson random measures such that

$$P_1, \ldots, P_m|\mathcal{G} \sim PRM(\mathcal{G})$$
$$\mathcal{G} \sim GP(a, H).$$

Following Thibaux (2008) notation, we update our posterior belief of $\mathcal{G}$ as follows

$$\mathcal{G}|P_1, \ldots, P_m \sim GP(c^*, H^*),$$

where

$$c^* = c + m$$
$$H^* = \frac{c}{c+m}H + \frac{m}{c+m}\frac{\sum_{i=1}^m P_i}{m}.$$

For more details refer to Wolpert and Ickstadt (1998a).

3.1.2 Series representation of the Gamma process

From Ferguson and Klass (1972), the Gamma process $\mathcal{G} \sim GP(a, H)$, can be written as a sum representation based on the Gamma Lévy measure as follows:

$$\mathcal{G}^{Ferg}(\cdot) = \sum_{i=1}^\infty N^{-1}(\Gamma_i)\delta_{\theta_i}(\cdot), \tag{3.1.2}$$

where $\{\theta_i\}_{i\geq 1}$ is a sequence of i.i.d. random variable with common distribution H independent of $\{\Gamma_i\}_{i\geq 1}$. The sequence $\{\Gamma_i\}_{i\geq 1}$ is defined as in chapter 2.

The terms in the Ferguson and Klass (1972) sum representation are relatively difficult to compute since there is no closed form for the inverse of the Lévy measure. Moreover, there are infinite terms in (3.1.2) to calculate.

3.1.3 Approximation of the Gamma process

Bondesson (1982) introduces a sum representation of the Gamma process which avoids the need to work with Lévy measures. It is shown that

$$\mathcal{G}^{Bond}(\cdot) = \sum_{i=1}^{\infty} \exp\left(-\frac{\Gamma_i}{a}\right) E_i \delta_{\theta_i}(\cdot), \tag{3.1.3}$$

where, $\{E_i\}_{i\geq 1}$ is a sequence of i.i.d. Exponential random variable with mean 1 and $\{\theta_i\}_{i\geq 1}$ is a sequence of i.i.d. random variable with common distribution H. In here, $\{E_i\}_{i\geq 1}$ and $\{\theta_i\}_{i\geq 1}$ are independent. The Bondesson (1982) sum representation of the Gamma process can be approximated by truncating the higher order in (3.1.3). Following the same truncation approach used for the Dirichlet process defined by Muliere & Tradella (1998), we let

$$n = \inf\{i : \exp\left(-\frac{\Gamma_i}{a}\right) E_i < \epsilon\}, \tag{3.1.4}$$

for $\epsilon \in (0, 1)$. The Bondesson (1982) sum approximation is defined as

$$\mathcal{G}^{Bond}(\cdot) = \sum_{i=1}^{n} \exp\left(-\frac{\Gamma_i}{a}\right) E_i \delta_{\theta_i}(\cdot). \tag{3.1.5}$$

Note that the weights in (3.1.5) are not monotonically decreasing almost surely. This phenomena is more obvious by looking at Figure 3.1. Note that the vertical lines in this figure represent the weights calculated in (3.1.5) which are clearly not monotonically decreasing. We will discuss the efficiency of different algorithm in further sections.

Figure 3.1 depicts one sample path of the Gamma process $\mathcal{G}^{Bond}((-\infty, t]) \sim GP(a, H)$ for $t \in \Omega = \mathbb{R}$ using Bondesson (1982) approximation. We choose $a = 5$ and $H \sim t(5)$, a t-distribution with five degrees of freedom. By choosing $\epsilon = 0.0001$, we get $n = 30$. The value of n is calculated based on the stopping rule in (3.1.4). The plot shows as well the weights in (3.1.5) at locations $t = \theta_{(i)}$ as vertical lines. The more intense the weight is, the higher the vertical line becomes. Note that a jump in the Gamma process occurs at $t = \theta_{(i)}$ every time there is a visible weight at the same location $t = \theta_{(i)}$, for $i = 1, \ldots, n$. Thus, a more intense weight results in a higher jump in the Gamma process. Note that $\theta_{(1)} < \cdots < \theta_{(n)}$ are the ordered statistics for $\theta_1, \ldots, \theta_n$, such that $\theta_i \overset{i.i.d.}{\sim} t(5)$.

Zarepour and Al Labadi (2012) derive a finite sum approximation of the Gamma process $\mathcal{G}_n$, which converges a.s. to the Ferguson and Klass (1972) sum representation. To see this, let $G_n \sim \text{Gamma}(a/n, 1)$, i.e:

$$G_n(x) = Pr(X_n > n) = \int_x^\infty \frac{1}{\Gamma(a/n)} e^{-t} t^{a/n-1} dt, \tag{3.1.6}$$

and

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

Using the fact that $n/\Gamma(\alpha/n) = \alpha/\Gamma(\alpha/n+1)$ and $n/\Gamma(\alpha/n) \to \alpha$, we have for $x > 0$,

$$nG_n(x) = \frac{n}{\Gamma(\alpha/n)} \int_x^\infty e^{-t} t^{\alpha/n-1} dt \to \alpha \int_x^\infty e^{-t} t^{-1} dt = N(x).$$

Note that for every $x > 0$, $nG_n(x)$ is a sequence of monotone functions converging to a continuous monotone function, therefore

$$G_n^{-1}(x/n) \to N^{-1}(x).$$

By taking $x = \Gamma_i$ and from the fact that $\Gamma_{n+1}/n \to 1$ almost surely as $n \to \infty$, we have

$$G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right) \overset{a.s.}{\to} N^{-1}(\Gamma_i).$$

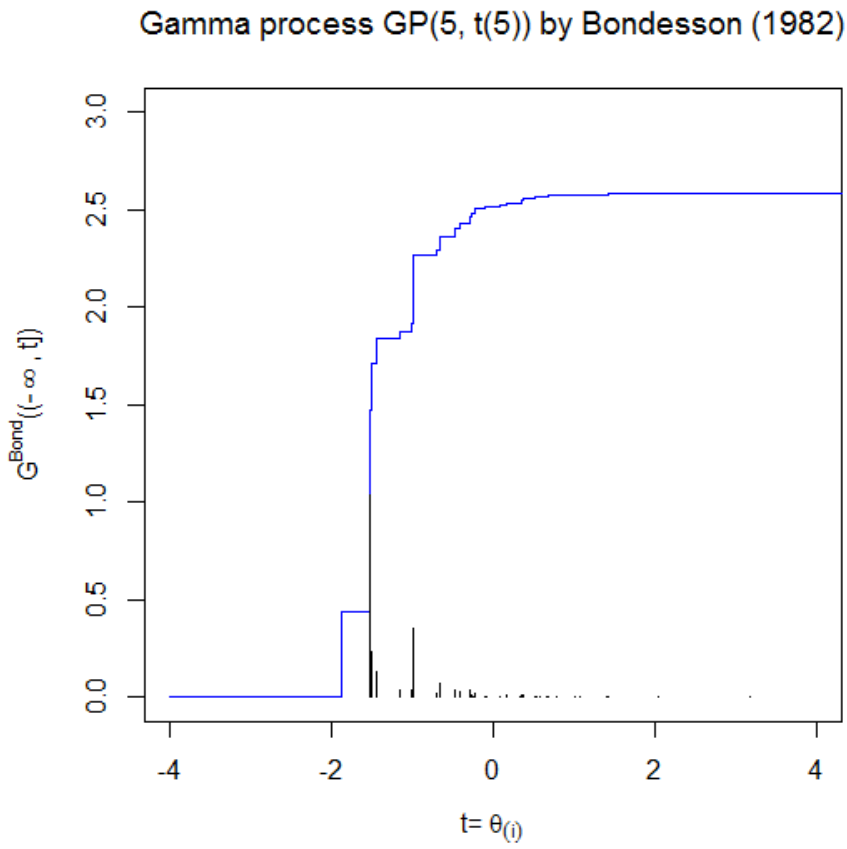

Figure 3.1: One sample path of the Gamma process, $G^{Bond} \sim GP(a, H)$ using Bondeson (1982) approximation. We choose $a = 5$ and $H \sim t(5)$. By choosing $\epsilon = 0.0001$, and following the stopping rule in (3.1.4), we get $n = 30$. The x-axis represents the set of i.i.d. atoms generated from $\theta_i \sim H$ in increasing order and the y-axes represents the corresponding Gamma process. We display in the same plot, the weights in (3.1.5) as vertical lines at the corresponding atoms $t = \theta_i$ for $i = 1, \ldots, n$.

The Gamma process can be approximated as follows:

$$\mathcal{G}_n^{Zar\&Al\_Lab}(\cdot) = \sum_{i=1}^{n} G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)\delta_{\theta_i}(\cdot) \overset{a.s.}{\to} \mathcal{G}(\cdot) = \sum_{i=1}^{\infty} N^{-1}(\Gamma_i)\delta_{\theta_i}(\cdot). \tag{3.1.7}$$

The next algorithm describes the set of rules to sample from the Zarepour and Al Labadi (2012) approximation of the Gamma process.

Algorithm A: An approximation of the Gamma process

1. Choose a relatively large positive integer n.
2. Generate $\theta_i \overset{i.i.d.}{\sim} H$, for $i = 1, \ldots, n$.
3. Generate $n+1$ independent exponential distributions with mean 1, $E_i \overset{i.i.d.}{\sim} \text{Exp}(1)$. Set $\Gamma_i = \sum_{k=1}^{i} E_k$, for $i = 1, \ldots, n+1$.
4. Computing $G_n^{-1}(\Gamma_i/\Gamma_{n+1})$ is equivalent as computing the quantile of the gamma distribution, Gamma$(a/n, 1)$ evaluated at $(1 - \Gamma_1/\Gamma_{n+1})$.

Note that the approximation of Zarepour and Al Labadi (2012) is not based on a stopping rule. It is more an asymptotic result. Nevertheless, a suggested stopping rule for n is proposed for comparison purposes as follows:

$$n = \inf \left\{ j : G_j^{-1} \left(\frac{\Gamma_j}{\Gamma_{j+1}} \right) < \epsilon \right\}. \tag{3.1.8}$$

Figure 3.2 shows one sample path of the Gamma process using Zarepour & Al Labadi (2012) in (3.1.7). We take $a = 5$ and the base measure $H \sim t(5)$. By choosing $\epsilon = 0.0001$ and using the stopping rule defined in (3.1.8), we get $n = 14$. Figure 3.2 plots as well the corresponding weights calculated in (3.1.7) which are represented in vertical lines at the corresponding location $t = \theta_i$. Notice that the weights are now monotonically decreasing almost surely.

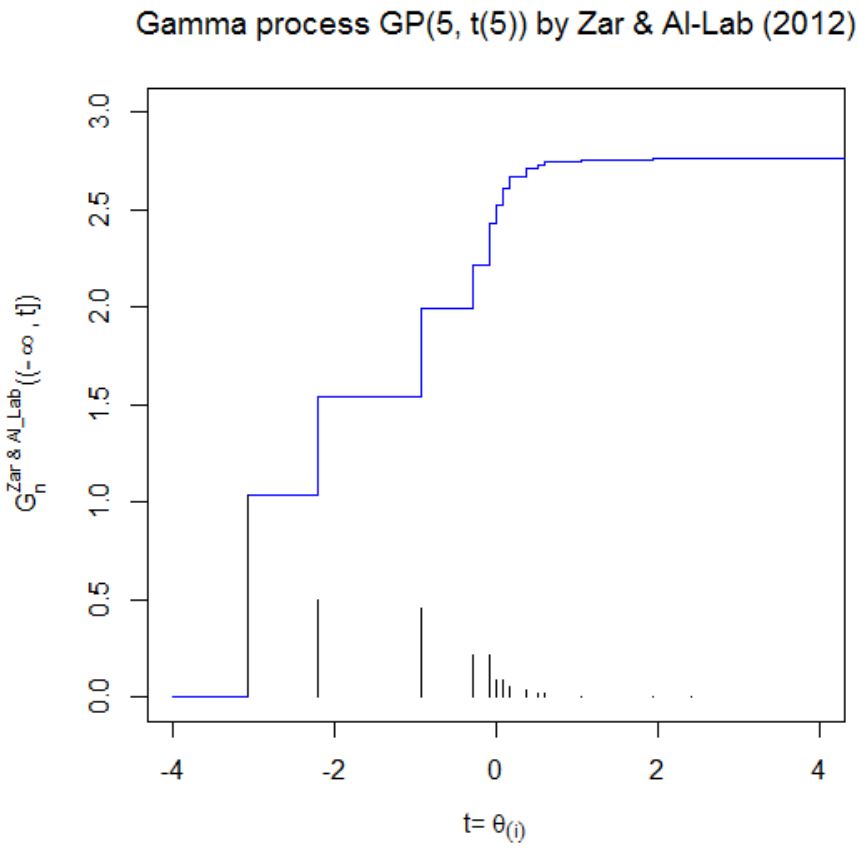

Figure 3.2: One sample path of the Gamma process using Zarepour & Al Labadi (2012) approximation. For comparison purposes we use the same parameter used in Figure 3.1, particularly we choose $a = 5$ and $H \sim t(5)$. Choosing $\epsilon = 0.0001$ and using the stopping rule in (3.1.8), we get $n = 14$. The plot shows as well the weights calculated in (3.1.7) as vertical lines.

It is worth mentioning that the approximation of Zarepour & Al Labadi (2012) for the Gamma process is very efficient from a computation point of view. Moreover, the algorithm will be as efficient as Bondesson (1982) with much smaller (almost half) values of n. Nevertheless, the weights are monotonically decreasing which ensure that there will not be an intense weight after the last negligible weights calculated (or observed). This phenomena is not guaranteed in the Bondesson (1982) Gamma process approximation.

The plot at the top of Figure 3.3 shows ten different paths of the Gamma process approximated using the Bondesson (1982) sum approximation with concentration parameter $a = 5$ and base measure $H \sim t(5)$. For each path, the values of the truncation n is equal to $45, 25, 42, 40, 39, 41, 7, 52, 39$ and 27. We use the truncation rule in (3.1.4) with $\epsilon = 0.0001$. The plot at the bottom of Figure 3.3 shows ten different paths of the Gamma process with the same parameters, a and H, but this time based on the Zarepour and Al Labadi (2012) Gamma process approximation. For each path, the values $n = 14, 10, 10, 10, 11, 12, 8, 11, 10$ and 9 are calculated based on the stopping rule in (3.1.8) with the same tolerance value ϵ. As discussed earlier, the weights of the Zarepour and Al Labadi (2012) approximation of the Gamma process are decreasing almost surely, thus the plot at the bottom shows decreasing jumps (more intense jumps at the beginning and gradually decreases toward the end).

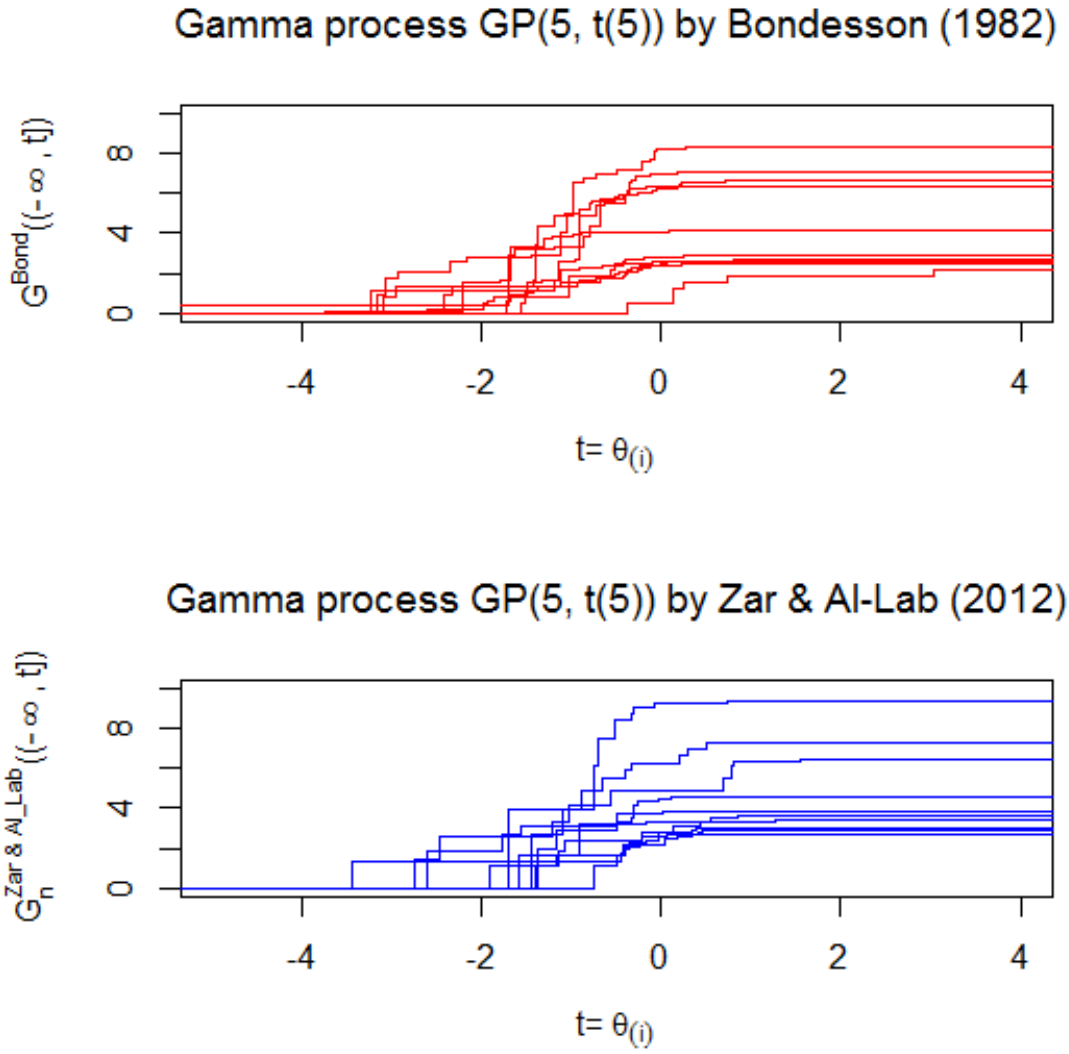

Figure 3.3: Ten sample paths of the Gamma process approximation with $a = 5$, $H = t(5)$ and $\epsilon = 0.0001$. The plot at the top of the figure uses the Bondesson (1982) Gamma approximation and the plot at the bottom uses the Zarepour and Al Labadi (2012) Gamma process approximation. The truncation values for Bondeson (1982) are $n = 45, 25, 42, 40, 39, 41, 7, 52, 39$ and 27, one value for each path. Whereas the truncation values for Zarepour and Al Labadi (2012) are found to be $n = 14, 10, 10, 10, 11, 12, 8, 11, 10$ and 9, for the same tolerance value $\epsilon = 0.0001$.

3.2 Dirichlet process

It is well known that the beta distribution, denoted by $\text{Beta}(a, b)$, is used as a conjugate prior for a binomial model. More concretely, let

$$f(X|p) \sim \text{Binomial}(n, p)$$
$$g(p) \sim \text{Beta}(a, b).$$

By using Bayes' theorem, $f(p|X) \propto f(X|p)g(p)$, the posterior distribution of p becomes

$$f(p|X = x) \sim \text{Beta}(a + x, n + a - x).$$

3.2.1 Definition of the Dirichlet distribution

Definition 3.2.1 (Dirichlet Distribution) *Let $P = (p_1, p_2, \ldots, p_{k-1})$ be a random vector, such that $p_i \geq 0$ for $i = 1, 2, \ldots, k$ and $p_1 + \cdots + p_k = 1$. In addition, suppose that $\boldsymbol{a} = (a_1, \ldots, a_k)$, with $\alpha_i > 0$ for $i = 1, 2, \ldots, k$, and let $a_0 = \sum_{i=1}^{k} a_i$. Then, P is said to have a Dirichlet distribution with parameter $\boldsymbol{a}$, denoted by $P \sim Dir(a_1, \ldots, a_k)$, if its density function is given by*

$$f(p_1, \ldots, p_{k-1}|a_1, \ldots, a_k) = \frac{\Gamma(a_0)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i}^{k-1} p_i^{a_i - 1} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{a_k - 1},$$

over the simplex

$$S = \{(p_1, p_2, \ldots, p_{k-1}) : p_i \geq 0, \sum_{i=1}^{k-1} p_i \leq 1\},$$

where $\Gamma(x)$ denotes the Gamma function.

When $k = 2$, the Dirichlet distribution reduces to the beta distribution, which has the density function

$$f(p; a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1 - p)^{b-1} \quad p \in (0, 1) \quad a, b > 0.$$

Similar to the beta distribution, the Dirichlet distribution $Dir(a_1, \ldots, a_k)$ is used as a conjugate prior for the Multinomial distribution given by

$$f(x_1, \ldots, x_{k-1}|p_1, \ldots, p_{k-1}) = \frac{n!}{x_1! \ldots x_{k-1}!(n - \sum_{i=1}^{k-1} x_i)!} \times p_1^{x_1 - 1} \ldots p_{k-1}^{x_{k-1}-1} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{n - \sum_{i=1}^{k-1} x_i}.$$

In this case, the posterior distribution of $P = (p_1, \ldots, p_{k-1})$ is

$$f(p_1, \ldots, p_{k-1}|X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) \sim Dir(a_1 + x_1, \ldots, a_k + x_k),$$

where $x_k = n - \sum_{i=1}^{k-1} x_i$.

The following is a general review of properties of the Dirichlet distribution.

1. If $\{G_i\}_{1 \le i \le k}$ are independent random variables such that $G_i \sim \text{Gamma}(a_i, 1), i = 1, 2, \cdots, k$, then
$$\left(\frac{G_1}{\sum_{i=1}^k G_i}, \frac{G_2}{\sum_{i=1}^k G_i}, \ldots, \frac{G_{k-1}}{\sum_{i=1}^k G_i}\right) \sim Dir(a_1, \ldots, a_k). \tag{3.2.1}$$

2. If $(p_1, \ldots, p_{k-1}) \sim Dir(a_1, \ldots, a_k)$ and $r_1, \ldots, r_l$ are integers such that $0 < r_1 < \cdots < r_l = k - 1$, then
$$\left(p_{(1,r_1)}, p_{(r_1+1,r_2)}, \ldots, p_{(r_{l-1}+1,r_l)}\right) \sim Dir\left(a_{(1,r_1)}, a_{(r_1+1,r_2)}, \ldots, a_{(r_{l-1}+1,r_l)}, a_k\right),$$
where, for $i, j = 1, \cdots, k$
$$p_{(i,j)} = p_i + p_{i+1} + \cdots + p_j$$
$$a_{(i,j)} = a_i + a_{i+1} + \cdots + a_j.$$

3. If $(p_1, \ldots, p_{k-1}) \sim Dir(a_1, \ldots, a_k)$ and $a_0 = \sum_{i=1}^k a_i$, then
$$E[(p_1, \ldots, p_{k-1})] = \left(\frac{a_1}{a_0}, \ldots, \frac{a_k}{a_0}\right)$$
$$Cov(p_i, p_j) = \begin{cases} \frac{-a_i a_j}{a_0^2(a_0+1)} & \text{for } i \neq j \\ \frac{a_i(a_0 - a_i)}{a_0^2(a_0+1)}. & \text{for } i = j \end{cases}$$

4. If the prior distribution of $(p_1, \ldots, p_{k-1}) \sim Dir(a_1, \ldots, a_k)$ and for the random variable X, let $Pr\{X = j|p_1, \ldots, p_k\} = p_j$ a.s. for $j = 1, \ldots, k$, then the posterior distribution of $(p_1, \ldots, p_{k-1}|X = j) \sim Dir(a_1^j, \ldots, a_k^i)$, where

$$a_i^j = \begin{cases} a_i & \text{if } i \neq j \\ a_i + 1 & \text{if } i = j. \end{cases}$$

Note that this shows that the Dirichlet distribution is a conjugate prior for any random variable with discrete probability distribution and with finite support.

3.2.2 Definition of the Dirichlet process

In this section we discuss how to place a conjugate prior for any probability measure on a general probability space with any support Ω (i.e. $\Omega = \mathbb{R}$, $\Omega = \mathbb{R}^+$ or $\Omega = \mathbb{R}^p$). Ferguson (1973) introduced the Dirichlet process as a class of priors over an arbitrary measurable space Ω, indexed by elements of a Borel σ-algebra $\mathcal{F}$. The Dirichlet process characterized by a stochastic process along with its conjugacy property has gained widespread adoption both in theory and practice.

Definition 3.2.2 (Dirichlet random variable) *Let $(\Omega, \mathcal{F})$ be an arbitrary measurable space and H be a probability measure on $(\Omega, \mathcal{F})$. Let $a > 0$ be arbitrary. A random probability measure $\mathcal{P}$ defined on $\mathcal{F}$ is called a* Dirichlet probability measure *with parameters a and H, denoted by $\mathcal{P} \sim DP(a, H)$, if for any finite measurable partition $\{A_1, \ldots, A_k\}$ of Ω, the joint distribution of the vector $(\mathcal{P}(A_1), \ldots, \mathcal{P}(A_k)) \sim Dir(aH(A_1), \ldots, aH(A_k))$, $k \geq 2$. We assume that if $H(A_j) = 0$, then $\mathcal{P}(A_j) = 0$ with probability one.*

For any measurable set $A \in \mathcal{F}$, $\mathcal{P}(A) \sim \text{Beta}\{aH(A), a(1 - H(A))\}$. Thus,

i)

$$E(\mathcal{P}(A)) = \frac{aH(A)}{aH(A) + a(1 - H(A))} = H(A)$$

$$Var(\mathcal{P}(A)) = \frac{H(A)(1-H(A)}{1+a}.$$

ii) If $\mathcal{P} \sim DP(a, H)$ and $X_1, \ldots, X_m$ is a sample from $\mathcal{P}$, then the posterior distribution of $\mathcal{P}|X_1, \ldots, X_m \sim DP(a^{(1)}, H^{(1)})$, where

$$a^{(1)} = a + m,$$
$$H^{(1)} = \frac{a}{a+m}H + \frac{m}{a+m}\left(\frac{1}{m}\right)\sum_{i=1}^{m}\delta_{X_i}.$$

For more details, interested reader can refer to Ferguson (1973).

From (i), it is clear that H plays the role of the centre of the process and hence is called the *base measure* (or also known as the initial guess). Also, we can see that as a gets larger the variance gets smaller. Therefore the distribution of $\mathcal{P}$ is more tightly concentrated around its mean, H. Hence, the parameter a can be seen as the *concentration parameter*. Note from (ii), the posterior base distribution $H^{(1)}$ is the combination of the base distribution and the empirical distribution. The posterior base distribution approaches the prior base measure H as $a \to \infty$. Also, it approaches the empirical distribution as $a \to 0$. Note that from strong law of large numbers, we get as $m \to \infty$, $H^{(1)} \to F$, where $F = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m}\delta_{X_i}$.

3.2.3 Series representation of the Dirichlet process

Ferguson (1973) showed that the Dirichlet process can be defined by using a sum representation for processes with independent increments. These processes are based on the arrival times of a homogeneous Poisson process. In fact, the Dirichlet process, $\mathcal{P} \sim DP(a, H)$, can be represented as a normalized Gamma process, see similarity of the self normalization in (3.2.1). Ferguson (1973) described the Dirichlet process by normalizing the series representation of a Gamma random measure given in (3.1.2)

as follows

$$\mathcal{P}^{Ferg}(\cdot) = \sum_{i=1}^{\infty} \frac{N^{-1}(\Gamma_i)}{\sum_{i=1}^{\infty} N^{-1}(\Gamma_i)} \delta_{\theta_i}(\cdot), \tag{3.2.2}$$

where N, is defined in (3.1.1) and $(\theta_i)_{i\geq 1}$ is a sequence of i.i.d. random variables with common distribution H. Notice that $\mathcal{P}^{Ferg}$ is a discrete random probability measure.

Similar to the Gamma process, sampling from the Dirichlet process based in (3.2.2) is difficult in practice since there is no closed form for the inverse of the Gamma Lévy measure. Moreover, there are an infinite number of terms in (3.2.2) that must be computed.

3.2.4 Approximation of the Dirichlet process

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

Theorem 7 (Bondesson 1982) *Let $(\theta_i)_{i\geq 1}$ 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 $(\theta_i)_{i\geq 1}$. Then,*

$$\mathcal{P}^{Bond}(\cdot) = \sum_{i=1}^{\infty} \frac{e^{-\Gamma_i/a} E_i}{\sum_{i=1}^{\infty} e^{-\Gamma_i/a} E_i} \delta_{\theta_i}(\cdot). \tag{3.2.3}$$

For more details, see Ishwaran & Zarepour (2002) and Zarepour & Al Labadi (2012). Note that the Bondesson representation overcomes the problem of inverting the Gamma Lévy measure. However, the infinite number of terms to compute in (3.2.3) make it difficult to sample from the Dirichlet process.

One can approximate the Dirichlet process by using a truncation as follows:

$$\mathcal{P}_n^{Bond}(\cdot) = \sum_{i=1}^{n} \frac{e^{-\Gamma_i/a} E_i}{\sum_{i=1}^{n} e^{-\Gamma_i/a} E_i} \delta_{\theta_i}(\cdot). \tag{3.2.4}$$

The choice of n can be selected for a given tolerance value $\epsilon \in (0, 1)$ by

$$n = \inf \left\{ j : \frac{e^{-\Gamma_j/a} E_j}{\sum_{i=1}^{j} e^{-\Gamma_j/a} E_j} < \epsilon \right\}. \tag{3.2.5}$$

Figure 3.4 shows the Bondesson (1982) approximation of Dirichlet process with $a = 5$ and base measure $H \sim t(5)$. It also shows the weights in (3.2.4) represented as vertical lines at different location θ_i. For a tolerance of $\epsilon = 0.0001$ and following the truncation rule in (3.2.5), we get $n = 25$. Note that unlike the Gamma process, the weights of the Dirichlet process sum up to 1. It is worth mentioning that the weights in (3.2.4), represented in the plot by vertical lines, are not monotonically decreasing almost surely.

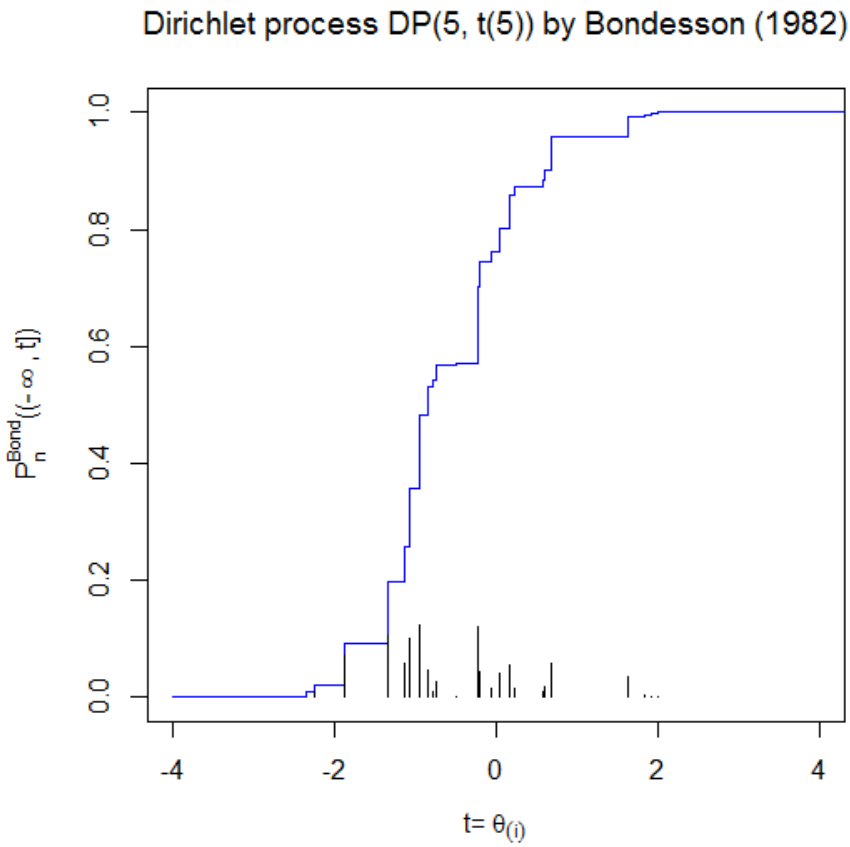

Figure 3.4: One sample path of the Dirichlet process approximated using Bondesson (1982) approximation. We choose $a = 5$ and $H \sim t(5)$. For $\epsilon = 0.0001$, we get $n = 25$. Vertical lines at different location $\theta_{(i)}$ represent the weights in calculated (3.2.4).

Figure 3.5 shows ten different sample paths of the Dirichlet process using Bondesson (1982) approximation. In all paths, we take $a = 5$ and $H \sim t(5)$. For $\epsilon = 0.0001$ the truncation value n are calculated following (3.2.5), thus we get $n = 26, 31, 23, 26, 27, 36, 25, 28, 27$ and 24.

The Ferguson (1973) and Bondesson (1982) sum representation for the Dirichlet process is based on the normalized Gamma process. Sethuraman (1994) defined the

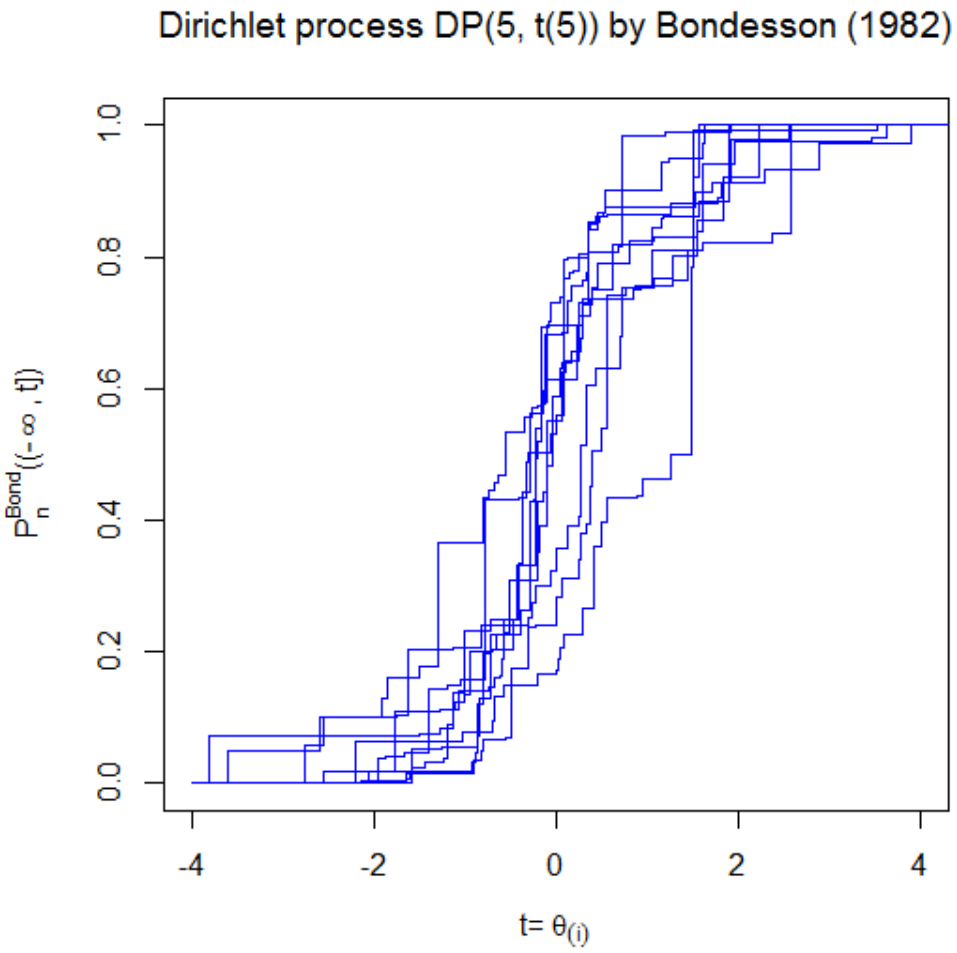

Figure 3.5: Ten sample paths of the Dirichlet process using Bondesson (1982) approximation with $a = 5$, $H \sim t(5)$. For $\epsilon = 0.0001$, we get $n = 26, 31, 23, 26, 27, 36, 25, 28, 27$ and 24.

Dirichlet process by using a stick breaking representation instead. This representation does not involve a normalization. Similar to the Bondesson (1982) sum representation, the stick breaking representation avoids inverting the Lévy measure N in (3.1.1).

Theorem 8 (Sethuraman 1994) *Let $(\mathcal{B}_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with Beta$(1, a)$ distribution. Define*

$$p_1 = \mathcal{B}_1, \quad p_i = \mathcal{B}_i \prod_{k=1}^{i-1}(1 - \mathcal{B}_k), \qquad i \geq 2.$$

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

$$\mathcal{P}^{Seth}(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(\cdot), \tag{3.2.6}$$

is a Dirichlet process with parameter a and H.

The Dirichlet process can be approximated by using Sethurman's stick breaking approximation. This is done by truncating the higher order terms in (3.2.6). Let $(\mathcal{B}_i)_{i\geq 1}$, $(p_i)_{i\geq 1}$ and $(\theta_i)_{i\geq 1}$ be defined in Theorem 8 with the only difference that $\mathcal{B}_n = 1$,

$$\mathcal{P}_n^{Seth}(\cdot) = \sum_{i=1}^{n} p_i \delta_{\theta_i}(\cdot). \tag{3.2.7}$$

Note that by letting $\mathcal{B}_n = 1$, the weights $(p_1, \ldots, p_n)$ sum up to 1 almost surely. For more details, see Ishwaran and James (2001). Muliere and Tradella (1998) proposed a stopping rule for n where, for $\epsilon \in (0, 1)$

$$n = \inf\{i : p_i = (1 - \mathcal{B}_1) \ldots (1 - \mathcal{B}_{i-1})\mathcal{B}_i < \epsilon\}. \tag{3.2.8}$$

Figure 3.6 shows one path of the Dirichlet process using the Sethuraman (1994) approximation. We use $a = 5$ and $H \sim t(5)$. Figure 3.6 depicts also the weights in (3.2.7) at every location $t = \theta_{(i)}$. Those weights are represented in the plot by vertical lines. Using the stopping rule proposed by Muliere and Tradella (1998) in (3.2.8) with $\epsilon = 0.0001$, we get $n = 45$. Similar to the Dirichlet process approximated by Bondeson (1982) (see Figure 3.4), the weights in the stick breaking approach are not strictly decreasing, thus vertical lines in the graph are not strictly decreasing.

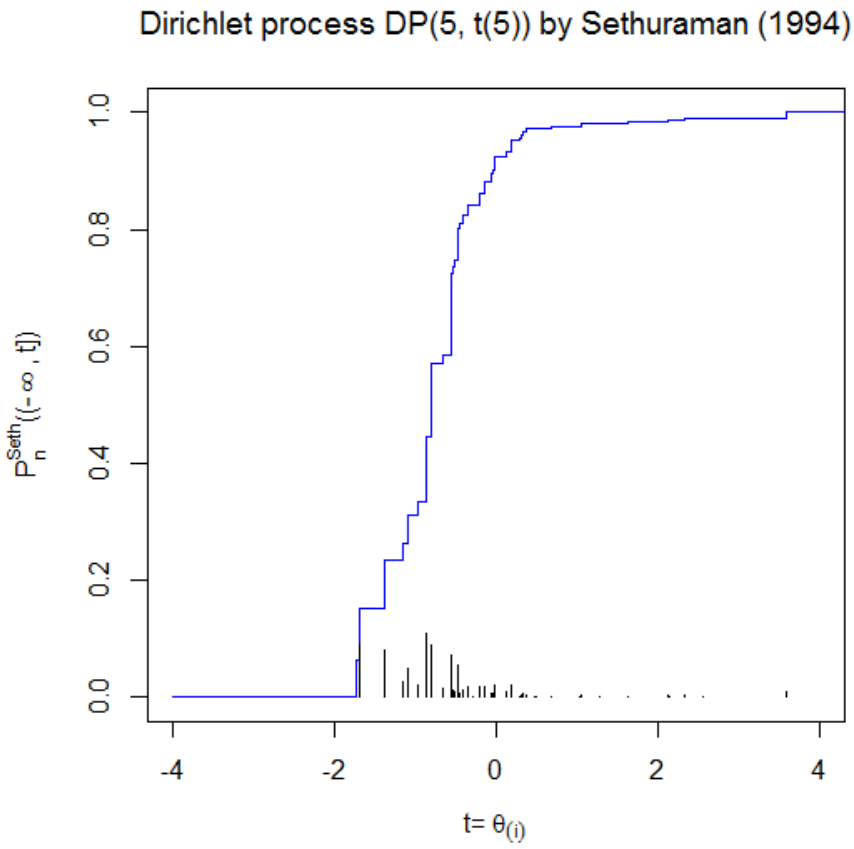

Figure 3.6: One sample path of Sethuraman (1994) Dirichlet process approximation, $P^{Seth}((-\infty, t]) \sim DP(a, H)$. We choose $a = 5$ and $H \sim t(5)$. For $\epsilon = 0.0001$ and using the stopping rule in (3.2.8), we get $n = 45$.

Figure 3.7 shows ten different paths of the Dirichlet process approximated by the stick breaking approach of Sethurman (1994). For each path, the truncation values $n = 30, 38, 57, 44, 54, 52, 21, 31, 43$ and 54 are calculated using (3.2.8). Note that the values of n calculated for each path of the Dirichlet process approximation of Sethuraman (1994) is relatively higher comparing to the Dirichlet process approximated by Bondeson (1982) (see Figure 3.5 for more details). In the next section, we will discuss theoretically why the weights in (3.2.7) are not strictly decreasing. This makes the

stopping rule in (3.2.8) and (3.2.5) inefficient to approximate the Dirichlet process. Indeed, for both approximations, the stopping rules are overestimating the value of n.

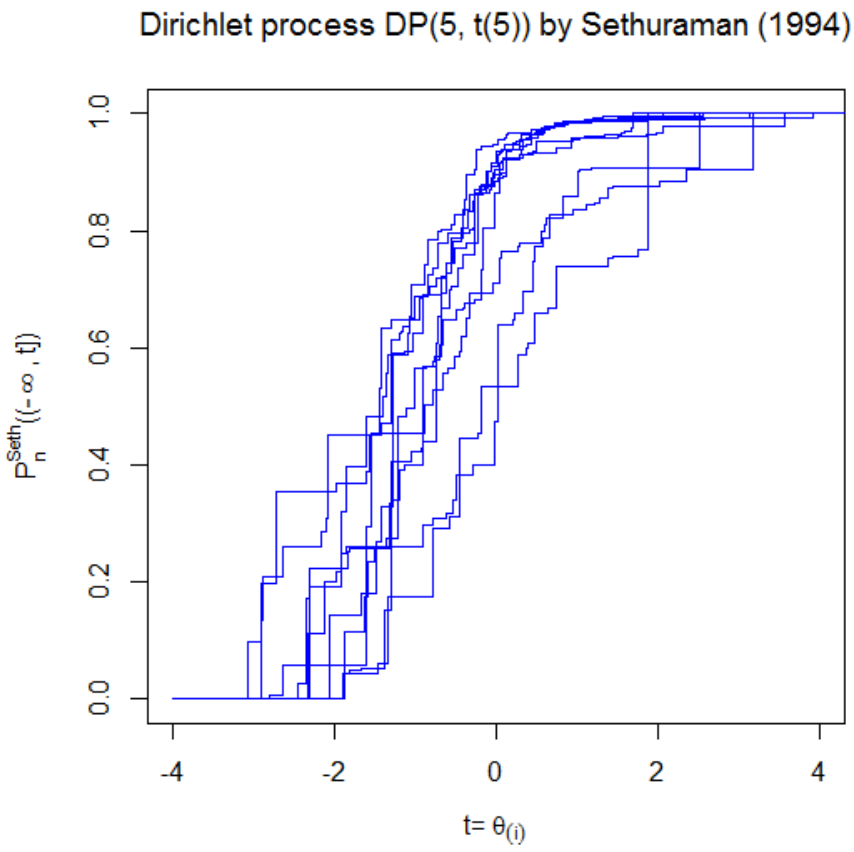

Figure 3.7: Ten sample paths of the Dirichlet process approximated by Sethuraman (1994) stick breaking approach. We choose $a = 5$ and $H \sim t(5)$. For $\epsilon = 0.0001$ and using the stopping rule in (3.2.8), the values of n for each path is $n = 30, 38, 57, 44, 54,\ 52, 21, 31, 43$ and 54.

Zarepour and Al Labadi (2012) prove that the weights in the Bondesson's representation (3.2.3) and in the Sethurman sum representation (3.2.7) are not strictly

decreasing almost surely. They prove that for $i \geq 1$,

$$Pr\left\{\frac{e^{-\Gamma_{i+1}/a}E_{i+1}}{\sum_{i=1}^{\infty}e^{-\Gamma_i/a}E_i} < \frac{e^{-\Gamma_i/a}E_i}{\sum_{i=1}^{\infty}e^{-\Gamma_i/a}E_i}\right\} = a\sum_{k=0}^{\infty}(-1)^k/(k+a),$$

and,

$$Pr\{p_{i+1} < p_i\} = a\sum_{k=0}^{\infty}(-1)^k/(k+a), \tag{3.2.9}$$

where $\{p_i\}_{i\geq 1}$ are as defined in Theorem 8. When $a = 1$ the right hand side of (3.2.9) is equal to 0.6931 and when $a = 10$ the probability is 0.5249. Therefore, for almost all i, with certain values of a there is a non-negligible probability of having non decreasing weights. Therefore, the suggested stopping rules in (3.2.8) and (3.2.5) with respect to the suggested weights are not efficient in simulating the Dirichlet process. The weights are not monotonically decreasing, this phenomena will tend to overestimate the truncation value n. Moreover, there is no guarantee that after choosing the weights up to the value n, there will not be a non negligible weight.

Zarepour and Al Labadi (2012) proposed a monotonically decreasing approximation of the Dirichlet process by normalizing the finite sum $\mathcal{G}_n$ of the Gamma process defined in (3.1.7). This is defined in the next theorem.

Theorem 9 (Zarepour & Al Labadi) *Let $(\theta_i)_{i\geq 1}$ be an i.i.d. sequence of random variables with common distribution H, independent of $(\Gamma_i)_{i\geq 1}$, then as $n \to \infty$*

$$\mathcal{P}_n^{Zar\&Al\_Lab} = \sum_{i=1}^{n}\frac{G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)}{\sum_{i=1}^{n}G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)}\delta_{\theta_i} \overset{a.s}{\to} P^{Ferg}, \tag{3.2.10}$$

where, $G_n(x)$ is as defined in (3.1.6).

Note that since G_n^{-1} is a decreasing function then $G_n^{-1}(\Gamma_i/\Gamma_{n+1}) > G_n^{-1}(\Gamma_{i+1}/\Gamma_{n+1})$. This is coming from the fact that for any $1 \leq i \leq n$, $\Gamma_i/\Gamma_{n+1} < \Gamma_{i+1}/\Gamma_{n+1}$ almost surely. To sample from the Dirichlet process approximation of Zarepour and Al

Labadi (2012), we can use the set of rules describe in Algorithm A. But we would need to normalize the weights of the Gamma process to get the weights of the Dirichlet process.

Although the technique used by Zarepour & Al Labadi (2012) to approximate the Dirichlet process is not based on a truncation method, they propose for comparison purposes a random stopping rule similar to that given in (3.2.5). For a given tolerance value of $\epsilon \in (0, 1)$, n can be calculated as follows:

$$n = \inf \left\{ j : \frac{G_j^{-1}\left(\frac{\Gamma_j}{\Gamma_{j+1}}\right)}{\sum_{i=1}^{j} G_j^{-1}\left(\frac{\Gamma_i}{\Gamma_{j+1}}\right)} < \epsilon \right\}. \tag{3.2.11}$$

Figure 3.8 shows one sample path of the Dirichlet process using Zarepour & Al Labadi (2012) approximation with $a = 5$ and $H \sim t(5)$. The truncation value is calculated following (3.2.11). Thus for $\epsilon = 0.0001$, we get $n = 12$. Vertical lines in the graph state the weights in (3.2.10) at different location $t = \theta_{(i)}$.

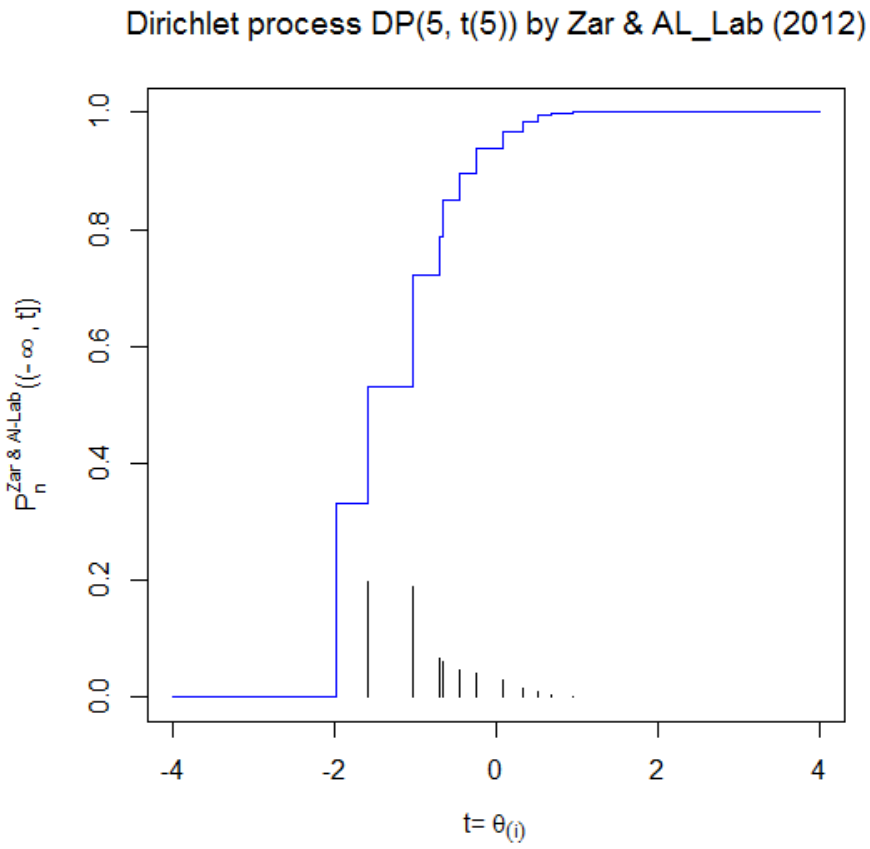

Figure 3.8: One sample path of the Dirichlet process approximated by Zarepour & Al Labadi (2012) approximation of the Dirichlet process with $a = 5$, $H \sim t(5)$. For $\epsilon = 0.0001$, we get $n = 12$

Figure 3.9 shows ten sample paths of the Dirichlet process approximated by Zarepour and Al Labadi (2012). For comparison purposes, parameters $a = 5$, $H \sim t(5)$ are the same as in the Bondesson(1982) and Sethurman (1994) approximation of the Dirichlet process discussed earlier. With $\epsilon = 0.0001$ and using the stopping rule in (3.2.11), we obtain $n = 15, 11, 13, 10, 7, 10, 8, 6, 7$ and 7, one for each sample path. Note that the Zarepour & Al Labadi (2012) approximation of the Dirichlet process uses less weights compared to the other two approximations.

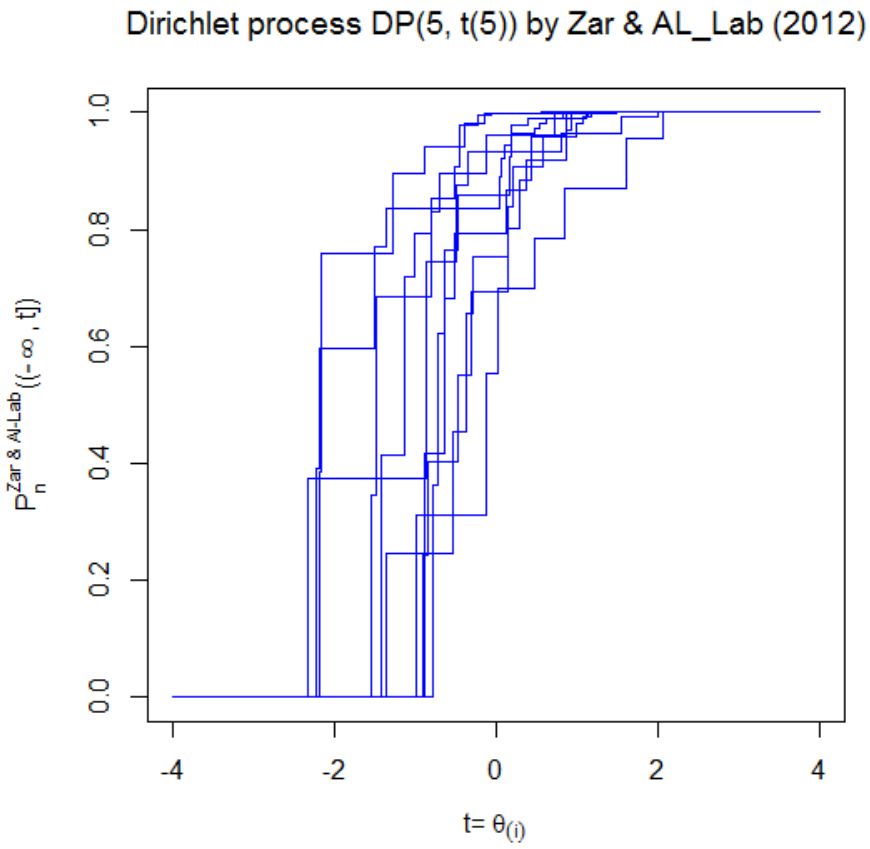

Figure 3.9: Ten paths of Zarepour & Al Labadi (2012) approximation of Dirichlet process with $a = 5$, $H \sim t(5)$. For $\epsilon = 0.0001$ the value of n is equal to $n = 15, 11, 13, 10, 7, 10, 8, 6, 7$ and 7, one for each sample path.

In the following we present a simulated example showing how we can use the Dirichlet process to estimate the distribution of an observed data. To see this more concretely, let us consider an i.i.d. data set coming from a Normal distribution. Let $X_1, \ldots, X_m \sim P$, where $P \sim \text{Normal}(-2, 1)$. Note that in practice the actual distribution of the data is not known in advance. Therefore, we are trying to infer the actual distribution in practice.

As a Bayesian approach, we need to put a prior guess on P. Consider using the following prior

$$P \sim DP(5, t(5)),$$

then, as discussed in (ii), the posterior distribution of P becomes

$$P|X_1, \ldots, X_m \sim DP\left((5+m), \left(\frac{5}{5+m}t(5) + \frac{m}{a+m}\left(\frac{1}{m}\right)\sum_{i=1}^{m}\delta_{X_i}\right)\right).$$

In Figure (3.10), the step functions with solid line describe five sample paths of the Dirichlet process $DP(5, t(5))$ prior guess, approximated by Zarepour & Al Labadi (2012) with $n = 1000$ (we use Algorithm A to sample from the Dirichlet process). In here the solid line represents the actual cumulative distribution of the data set, we have $X_1, \ldots, X_m \sim \text{Normal}(-2, 1)$. The dotted step functions represent five different paths of the posterior distribution of the Dirichlet process approximated after observing m data points. From top to bottom, m is chosen to be $5, 20$ and 200 respectively. We can see how regardless of our prior guess, the posterior distribution will converge to the actual distribution of the data set with increasing number of observation.

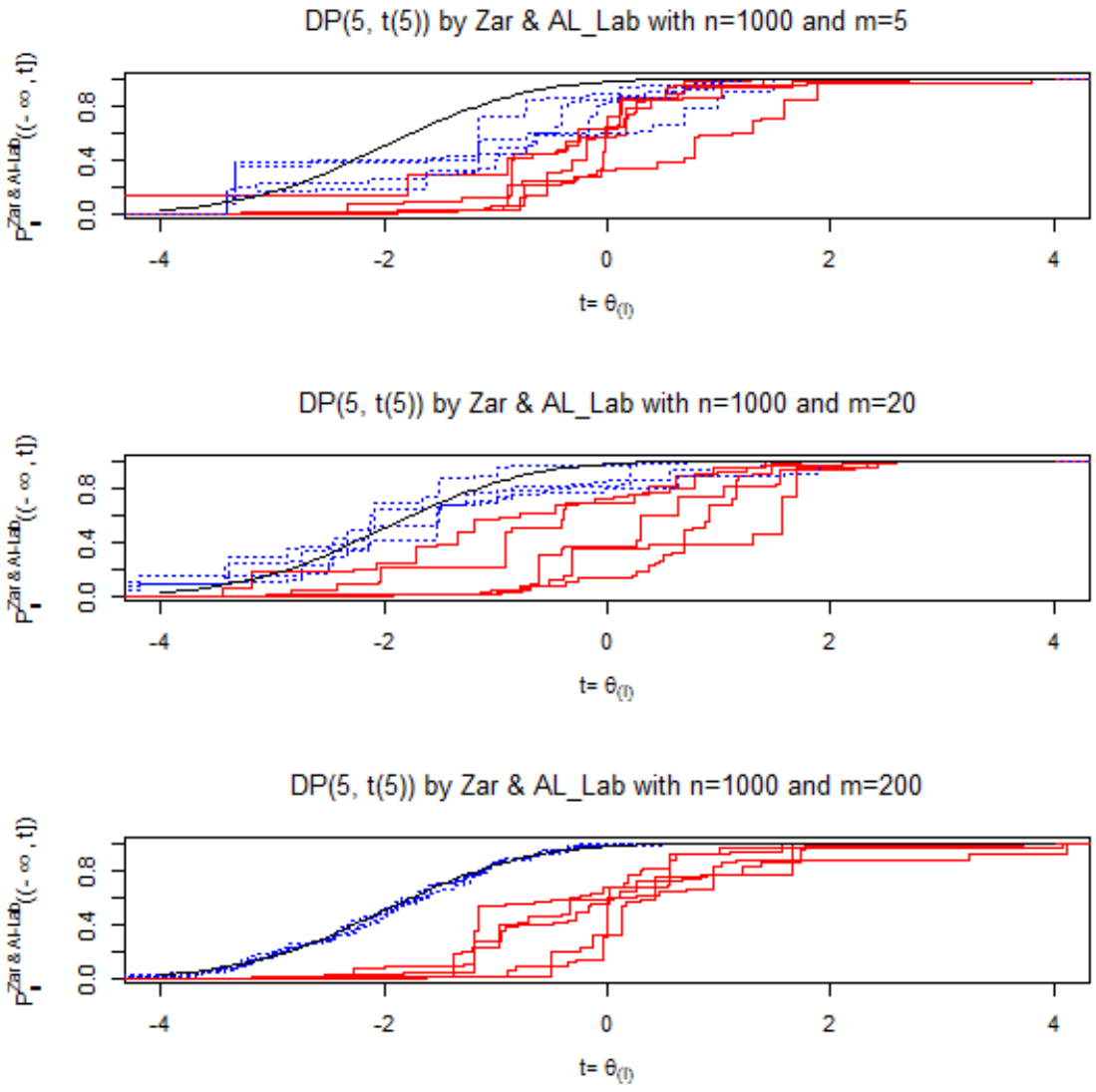

Figure 3.10: Solid step functions show the Dirichlet process prior using the Zarepour & Al Labadi (2012)'s approximation with $a = 5, H \sim t(5)$ and $n = 1000$. The solid line is the actual cumulative distribution of the data set which is in our case Normal$(-2, 1)$. Top to bottom plot shows the posterior distribution of the Dirichlet process after observing $m = 5, 20$ and 200 data points respectively.

Chapter 4

Two-parameter Poisson-Dirichlet and the normalized inverse-Gaussian process

4.1 Two-parameter Poisson-Dirichlet process

The two parameter Poisson-Dirichlet process also known as the Pitman-Yor process is a generalization of the Dirichlet process. It has also been used as prior in non-parametric Bayesian inference. Let $P_{H,\alpha,a} \sim PDP(H;\alpha,a)$ denote a two parameter Poisson-Dirichlet, the probability measure H is called the based measure, where α and a are called the discount parameter and the concentration parameter, respectively.

4.1.1 Definition of the two-parameter Poisson-Dirichlet process

Pitman and Yor (1997) introduce the stick-breaking definition of the two-parameter Poisson-Dirichlet process defined on an arbitrary measurable space $(\Omega, \mathcal{F})$.

Definition 4.1.1 *Let $0 \leq \alpha < 1$, $a > -\alpha$ and $(\beta_i)_{i\geq 1}$ be a sequence of independent random variables with Beta$(1-\alpha, a+i\alpha)$ distribution. Define*

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

Let $(\theta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with common distribution H, independent of $(\beta_i)_{i\geq 1}$ and $p_1 \geq p_2 \geq \cdots$ be the sorted values of $(p_i^{'})_{i\geq 1}$. Then the random probability measure

$$P_{H,\alpha,a}(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(\cdot) \tag{4.1.1}$$

is called a two-parameter Poisson-Dirichlet process with parameters α, a and H.

Note that for the special case when $\alpha = 0$, $(\beta_i)_{i\geq 1}$ become a sequence of independent random variables with Beta$(1, a)$. Thus, the two-parameter Poisson-Dirichlet process $P_{H,0,a}$ becomes simply the Dirichlet process. On the other hand, when $a = 0$, Pitman and Yor (1997) show that the two-parameter Poisson-Dirichlet process $P_{H,\alpha,0}$, becomes the normalized non-negative Stable law process with index $\alpha \in (0, 1)$.

Note that for any measurable subset A of Ω, the two-parameter Poisson-Dirichlet process has the following properties,

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

For more details on the calculation of the moments for the two-parameter Poisson-Dirichlet process, interested reader can refer to Carlton (1999).

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

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

Theorem 10 *Let $X_1, \ldots, X_m$ be a sample from $P_{H,\alpha,a}$. Let n be the number of distinct $X_i's$, $X_j^{'}$ be the j^{th} distinct X_i and m_j be the number of X_i equal to $X_j^{'}$. Then*

$$P_{H,\alpha,a}|X_1, \ldots, X_m \stackrel{d}{=} \sum_{j=1}^{n} W_j \delta_{X_j^{'}} + W_{n+1} P_{H,\alpha,a+n\alpha},$$

where $P_{H,\alpha,a+n\alpha} \sim PDP(H, \alpha, a + n\alpha)$ and $(W_1, \ldots, W_{n+1}) \sim Dir(m_1 - \alpha, \ldots, m_n - \alpha, a + n\alpha)$ and such that they are conditionally independent given $X_1, \ldots, X_m$.

4.1.2 Approximation of the two-parameter Dirichlet process

The infinite sum in (4.1.1) 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 done by truncating the higher order terms in the sum (4.1.1) (Al Labadi and Zarepour (2014)).

Let $(\beta_i)_{i\geq 1}$, $(p_i)_{i\geq 1}$ and α are as defined earlier, with $\beta_n = 1$ (Ishwaran and James, 2001). The random probability measure

$$P_{H,\alpha,a}(\cdot) = \sum_{i=1}^{n} p_i \delta_{\theta_i}(\cdot) \tag{4.1.2}$$

is the finite approximation of the two-parameter Poisson-Dirichlet process. Mimicking the same stopping rule n proposed by Muliere and Tradella (1998) for the Dirichlet process,

$$n = \inf\{i : p_i^{'} = (1 - \beta_1) \cdots (1 - \beta_{i-1})\beta_i < \epsilon\},$$

for $\epsilon \in (0, 1)$.

Al Labadi and Zarepour (2014) show that the weights, $(p_i^{'})_{i\geq 1}$, in (4.1.2) before ordering them are not strictly decreasing almost surely (see Lemma 1 of Al Labadi and Zarepour (2014) for the proof).

Pitman and Yor (1997) propose a different interesting approach to construct the two-parameter Poisson Dirichlet process. This is described in the next proposition.

Proposition 4.1.2 (Pitman and Yor 1997, Proposition 22) *For* $0 < \alpha < 1$ *and* $a > 0$, *suppose* $(p_1(0,a), p_2(0,a), \dots)$ *has distribution* $PD(0,a)$ *and* $(p_1(\alpha,0), p_2(\alpha,0), \dots)$ *has distribution* $PD(\alpha,0)$. *Independent of* $(p_1(0,a), p_2(0,a), \dots)$, *let* $(p_1^i(\alpha,0), p_2^i(\alpha,0), \dots)$, $i = 1,2,\dots$ *be a sequence of independent copies of* $(p_1(\alpha,0), p_2(\alpha,0), \dots)$. *Let* $(p_i)_{i\geq 1}$ *be the descending order statistics of* $\{p_i(0,a)p_j^i(\alpha,0), i,j = 1,2,\dots\}$. *Then* $(p_1, p_2, \dots)$ *has a* $PD(\alpha,a)$ *distribution.*

Note that the weights of the two-parameter Poisson-Dirichlet process is constructed based on two different choices of parameters. One with $\alpha = 0$ which correspond to the Dirichlet process $P_{H,0,a}$ and another when $a = 0$ which correspond to the normalized Stable law process $P_{H,\alpha,0}$. The index of the Stable law process α is in $(0,1)$. Therefore an approximation of the two-parameter Poisson-Dirichlet process would require to draw independently a sample from the Dirichlet process and from the normalized Stable law process.

Pitman and Yor (1997, Proposition 10) prove that the sum representation of the normalized Stable law process can be written as follows

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

Therefore, the approximation of the normalized Stable law process is

$$P_{H,\alpha,0}(\cdot) = \sum_{i=1}^{n} \frac{\Gamma_i^{-1/\alpha}}{\sum_{i=1}^{n} \Gamma_i^{-1/\alpha}} \delta_{\theta_i}(\cdot), \tag{4.1.3}$$

for large enough n. Note that the weights $\left(\Gamma_i^{-1/\alpha} / \sum_{i=1}^{n} \Gamma_i^{-1/\alpha}\right)_{1\leq i\leq n}$ are not necessarily strictly decreasing. Al Labadi and Zarepour (2014) approximate the two-parameter Poisson-Dirichlet process by first sampling a draw from the Dirichlet process given in (3.2.10)(using Algorithm A), then sampling a sample path of the normalized Stable law process in (4.1.3). Al Labadi and Zarepour (2014) compared their approximation with the corresponding stick-breaking approximation given in (4.1.2). Through simulation, Al labadi and Zarepour (2014) show that the two-parameter

Poisson-Dirichlet approximated using their approach concludes more precise results compared to the one obtained using the stick-breaking approximation in (4.1.2).

Figure 4.1 shows one sample path of the two parameter Poisson-Dirichlet process. We take H to be a t-distribution with 5 degrees of freedom, $\alpha = 0.5$ and $a = 1$. The graph shows as well the weights in (4.1.2) as vertical lines at different locations $t = \theta_{(i)}$.

4.2 Normalized inverse-Gaussian process (NIGP)

Analogous to the Dirichlet process, Lijoi, Mena and Prünster (2005) define the normalized inverse Gaussian process as a prior to use in Bayesian nonparametric inference.

Definition 4.2.1 *The random vector $(Z_1, \ldots, Z_m)$ is said to have a normalized inverse-Gaussian distribution with parameters $(\gamma_1, \ldots, \gamma_m)$, where $\gamma_i > 0$ for all i, if it has the joint density function*

$$f(z_1, \ldots, z_m) = \frac{e^{\sum_{i=1}^{\infty} \gamma_i} \prod_{i=1}^{m} \gamma_i}{2^{m/2-1} \pi^{m/2}} \times K_{-m/2}\left(\sqrt{\sum_{i=1}^{m} \frac{\gamma_{i^2}}{z_i}} \right) \times \left(\sum_{i=1}^{m} \frac{\gamma_{i^2}}{z_i} \right)^{-m/4} \times \prod_{i=1}^{m} z_i^{-3/2} \times I_S(z_1, \ldots, z_m),$$

where K is the modified Bessel function of the third type, $S = \{(z_1, \ldots, z_m) : z_i \geq 0, \sum_{i=1}^{m} z_i = 1\}$, and I_S represents the indicator function of the set S.

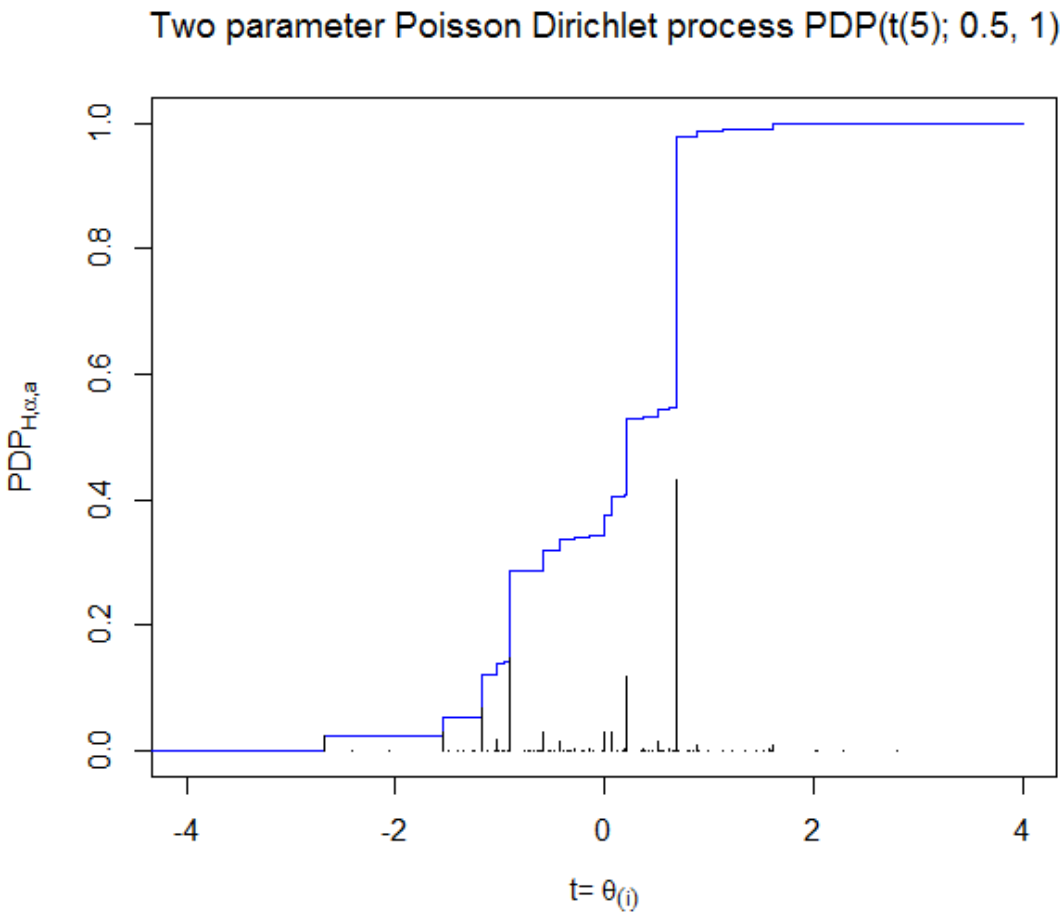

Figure 4.1: One sample path of the two parameter Poisson-Dirichlet process with $H \sim t(5)$, $\alpha = 0.5$ and $a = 1$. the x-axis shows atoms generated from $\theta_i \sim H$ in increasing order and the y-axes represent the resulting two parameter Poisson-Dirichlet process. Vertical lines represent the intensity of the weights calculated from (4.1.2).

4.2.1 Definition of the normalized inverse-Gaussian process (NIGP)

Definition 4.2.2 *Let E be a Polish space and $\mathscr{B}(E)$ (or Ω) be the Borel σ-algebra generated by the open sets in E. A random probability measure $P_{H,a} = \{P_{H,a}(B)\}_{B\in\mathscr{B}(E)}$ is called a normalized inverse-Gaussian process on $(E, \mathscr{B}(E))$ with parameter H (a fixed probability measure) and $a > 0$ (concentration parameter), if for any finite measurable partition $B_1, \ldots, B_m$ of $\mathscr{B}(E)$, the joint distribution of the vector $(P_{H,a}(B_1), \ldots, P_{H,a}(B_m))$ has a normalized inverse-Gaussian distribution with parameter $(aH(B_1), \ldots, aH(B_m))$. We denote the normalized inverse-Gaussian process with parameters a and H by $NIGP(H, a)$.*

Here are some basic properties of the normalized inverse-Gaussian process. For any $B \in \Omega$,

$$E(P_{H,a}(B)) = H(B),$$
$$Var(P_{H,a}(B)) = \frac{H(B)(1 - H(B))}{\xi(a)},$$

where

$$\xi(a) = \frac{1}{a^2 e^a \Gamma(-2, a)}$$

and $\Gamma(-2, a) = \int_a^\infty u^{-3} e^{-u} du$.

Abramowitz and Stegun (1972) show that for large values of a, we have $\xi(a) \approx a$. Therefore, similar to the two-parameter Poisson-Dirichlet process, the base measure H plays the role of the center of the process, while a plays the role of the concentration parameter.

The posterior distribution of the normalized inverse Gaussian process can be found in Lijoi, Mena and Prünster (2005). Note that its posterior distribution is not a conjugacy prior, similar to the two parameter Poisson-Dirichlet process.

4.2.2 Series representation of the NIGP

The normalized inverse-Gaussian process can be written as a series representation. Let $(\theta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with common distribution H, independent of Γ_i. The normalized inverse-Gaussian process has the exact sum representation as follows:

$$P_{H,a}(\cdot) = \sum_{i=1}^{\infty} \frac{L^{-1}(\Gamma_i)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)} \delta_{\theta_i}(\cdot),$$

where

$$L(x) = \frac{a}{\sqrt{2\pi}} \int_x^{\infty} e^{-t/2} t^{-3/2} dt, \text{ for } x > 0, \tag{4.2.1}$$

is the Lévy measure. For more details, the interested reader can refer to Ferguson and Klass (1972) or Nieto-Barajas and Prünster (2009).

4.2.3 Stick-breaking representation of the NIGP

Similar to the Gamma process, inverting the Lévy measure in (4.2.1) is difficult in practice. Favaro, Lijoi and Prünster (2012) use a "stick-breaking" approach to define the normalized inverse-Gaussian process.

Let $(Z_i)_{i\geq 1}$ be i.i.d. random variables where Z_i is 1/2-stable random variable with scale parameter 1. Let $X_1 \sim GIG(a^2, 1, -1/2)$, define

$$V_1 = \frac{X_1}{X_1 + Z_1},$$

for $i = 2, 3, \ldots$, given $V_1, \ldots, V_{i-1}$ and $X_i \sim GIG\left(a^2 / \prod_{j=1}^{i-1}(1 - V_j), 1, -i/2\right)$, for $i \geq 2$

$$V_i = \frac{X_i}{X_i + Z_i},$$

The sequence $(X_i)_{i\geq 1}$ and $(Z_i)_{i\geq 1}$ are independent from each other and GIG denotes the generalized inverse-Gaussian distribution (see equation (2) of Favaro, Lijoi and Prünster (2012)). Let $(\theta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with common distribution H independent of $(V_i)_{i\geq 1}$ and let

$$p_1 = V_1, \quad p_j = V_j \prod_{i=1}^{j-1}(1 - V_i), \quad j \geq 2, \tag{4.2.2}$$

then

$$P_{H,a}(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(\cdot), \tag{4.2.3}$$

is a normalized inverse-Gaussian process with parameters a and H.

4.2.4 Approximation of the NIGP

The normalized inverse-Gaussian process can be approximated by truncating the higher order terms in the sum (4.2.3) as follows:

$$P_{n,H,a}(\cdot) = \sum_{i=1}^{n} p_i \delta_{\theta_i}(\cdot), \tag{4.2.4}$$

where $(V_i)_{i\geq 1}$, $(p_i)_{i\geq 1}$ are as defined in (4.2.2) independent of $(\theta_i)_{i\geq 1}$. Note that $V_n|V_1, \ldots, V_{n-1} = 1$ is necessary to make the weights add to 1 almost surely. A stopping rule for choosing n is similar to the one proposed by Muliere and Tradella (1998) for the Dirichlet process, that is for $\epsilon \in (0, 1)$,

$$n = \inf\{i : p_i = V_j \prod_{i=1}^{j-1}(1 - V_i) < \epsilon\}. \tag{4.2.5}$$

The graph at the top of Figure 4.2 shows one sample path of the normalized inverse-Gaussian process using the stick breaking approach. We take H to be a t-distribution with 5 degrees of freedom, $a = 1$ and $n = 100$. The graph shows as well the weights in (4.2.4) in vertical lines at different locations $t = \theta_{(i)}$. Note that the

choice of n in that graph is a relatively big value (no stopping rule has been applied for this figure). The graph at the bottom shows one same sample path of the normalized inverse-Gaussian process with same parameters a and H but with $n = 3$. The value of n is calculated based on the stopping rule in (4.2.5) for epsilon=0.0001. Note that the stopping rule for n is not efficient because the weights are not monotonically decreasing. Therefore it stops before reaching the full approximation of the NIGP. Moreover, the assumption that $V_n|V_1, \ldots, V_{n-1} = 1$ for making the weights add to 1 leads to an inaccurate approximation of the NIGP.

4.2.5 Al Labadi & Zarepour approximation of the NIGP

Similar to the Dirichlet process, Al Labadi and Zarepour (2014) derive a finite sum representation of the normalized inverse-Gaussian process that converges almost surely to the Ferguson and Klass (1972) representation.

Let $(\theta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with common distribution H, independent of $(\Gamma_i)_{i\geq 1}$, then as $n \to \infty$

$$P_{n,H,a}^{Lab} = \sum_{i=1}^{n} \frac{L_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)}{\sum_{i=1}^{n} L_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)} \delta_{\theta_i} \overset{a.s.}{\to} P_{H,a} = \sum_{i=1}^{\infty} \frac{L^{-1}(\Gamma_i)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)} \delta_{\theta_i}, \tag{4.2.6}$$

where

$$L_n(x) = \int_x^{\infty} \frac{a}{n\sqrt{2\pi}} t^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{a^2}{n^2 t} + t\right) + \frac{a}{n}\right\} dt.$$

In here, $L(x)$ is defined in (4.2.1)

For the same reason discussed in Theorem 9, the weights in the sum approximation of the normalized inverse-Gaussian process (4.2.6) decrease monotonically for any fixed positive integer n. Recall that for any $1 \leq i \leq n$, $\Gamma_i/\Gamma_{n+1} < \Gamma_{i+1}/\Gamma_{n+1}$

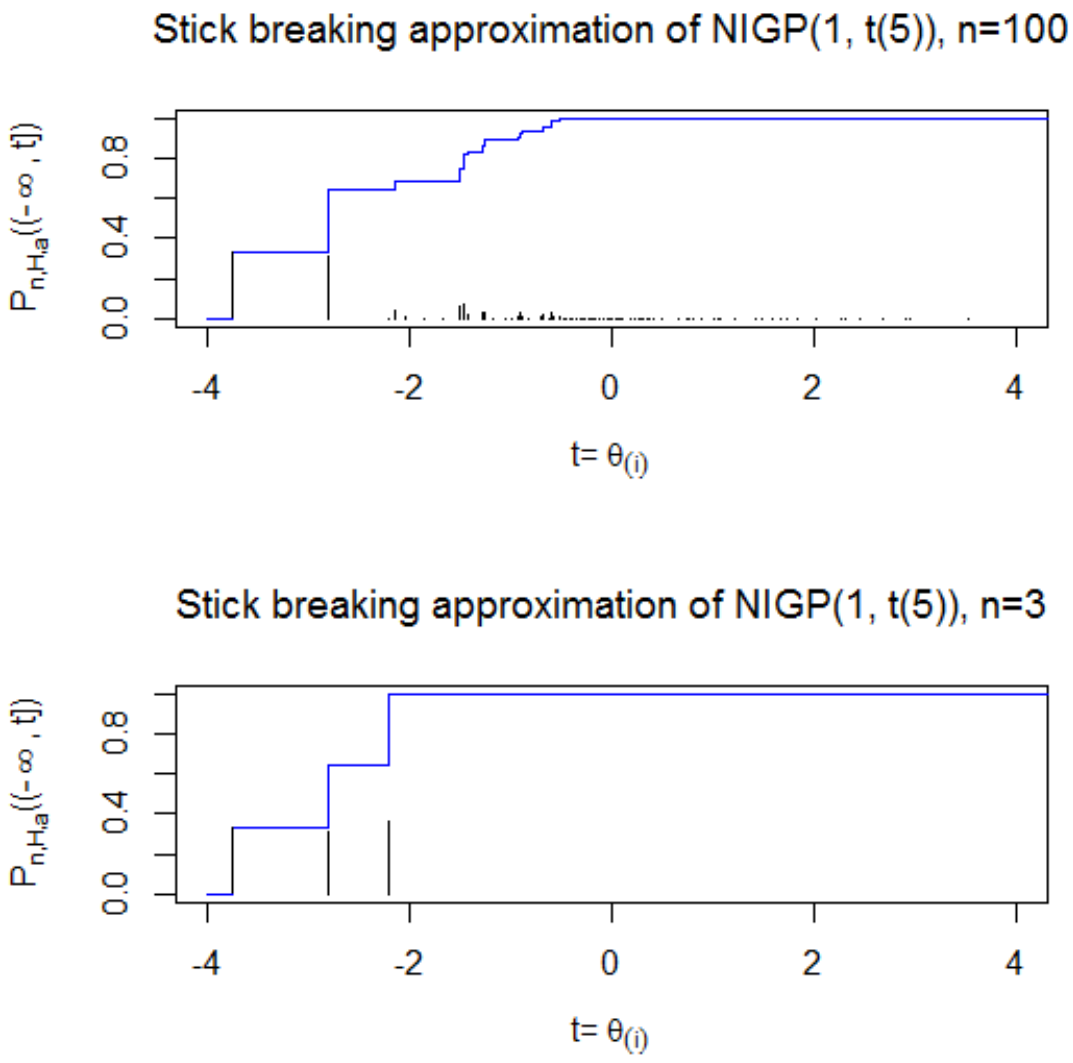

Figure 4.2: The plot at the top shows one sample path of the NIGP(1, t(5)) using the stick breaking approach with $n = 100$. The choice of n in this plot is chosen to be relatively large. The vertical lines show the weights in (4.2.4). The plot at the bottom depicts one sample path of NIGP(1, t(5)). Choosing $\epsilon = 0.0001$, we get $n = 3$ based on the stopping rule in (4.2.5).

almost surely. Note that L_n^{-1} is a decreasing function, therefore $L_n^{-1}(\Gamma_i/\Gamma_{n+1}) > L_n^{-1}(\Gamma_{i+1}/\Gamma_{n+1})$ almost surely.

Figure 4.3 depicts one sample path of the normalized inverse-Gaussian process using Al Labadi & Zarepour (2014) approximation. We take H to be a t-distribution with 5 degree of freedom, $a = 1$ and $n = 50$. The graph shows as well the weights in (4.2.4) in vertical lines at different locations $t = \theta_{(i)}$. Note that the weights are in decreasing order.

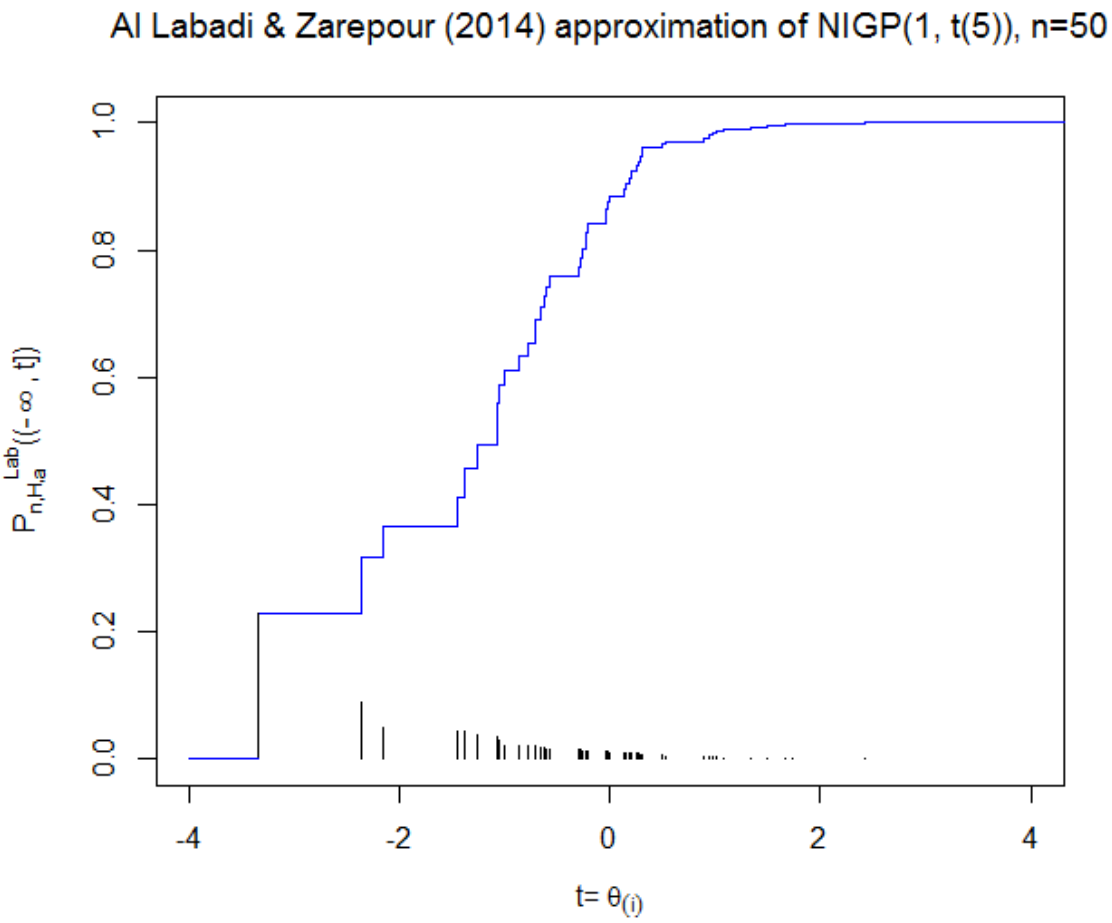

Figure 4.3: One sample path of the normalized inverse Gaussian process with $H \sim t(5)$, $a = 1$ and $n = 50$. The NIGP$(t(5), 1)$ is sampled using Al Labadi & Zarepour (2014) approximation

Chapter 5

Beta process

5.1 Beta process

5.1.1 Definition of the Beta process

In this section, we define the Beta process and some of its basic properties.

Definition 5.1.1 (Beta process) *[Thibaux & Jordan (2007)]*
A Beta process $B \sim BP(c, B_0)$ is a completely random measure whose Lévy measure depends on two parameters, c and B_0. The Lévy measure of the Beta process on $\Omega \times [0, 1]$ can be written as

$$\nu(d\theta, dp) = cp^{-1}(1-p)^{c-1}dpB_0(d\theta),$$

where B_0 is a diffuse finite measure on $(\Omega, \mathcal{F})$ and $c > 0$. The total mass of B_0 denoted by $\gamma = B_0(\Omega)$ is called the mass parameter.

Notice that B is only a finite measure but not necessarily a random probability measure. One of the basic properties of the Beta process is that for any set $S \in \Omega$

$$E[B(S)] = B_0(S) \quad \text{and} \quad Var[B(S)] = \frac{B_0(S)}{c+1}$$

See Hjort (1990) and Thibaux & Jordan (2007), for more details. Similar to the Dirichlet process, c is called the concentration parameter and B_0 is called the base measure. Note that as $c \to \infty$, $Var[B(S)] \to 0$, thus with higher value of c, the Beta process is more tightly close to the base measure B_0. In general the concentration parameter c is a function of θ but in this thesis we focus on the case where c is a constant.

5.1.2 Series representation of the Beta process

The Beta process with continuous base can be represented as a series representation following Ferguson (1972) such that

$$B^{Ferg}(\cdot) = \sum_{i=1}^{\infty} \nu^{-1}(\Gamma_i)\delta_{\theta_i}(\cdot), \tag{5.1.1}$$

where

$$\nu(x) = c\gamma \int_x^1 p^{-1}(1-p)^{c-1}dp,$$

and $(\theta_i)_{i\geq 1}$ is a sequence of i.i.d. random variables with common distribution B_0/γ independent of Γ_i. Note that for any set $S \in \Omega$, the infinite sum in (5.1.1) is finite only if B_0 is finite.

In the case when the base measure B_0 is discrete of the form $B_0 = \sum_i q_i\delta_{\theta_i}$, then B has atoms at the same locations θ_i and the Beta process is defined as follows

$$\begin{aligned} B &= \sum_{i=1}^{\infty} p_i\delta_{\theta_i} \\ p_i &\sim \text{Beta}(cq_i, c(1-q_i)), \end{aligned} \tag{5.1.2}$$

for $q_i \in (0,1)$. When the base measure B_0 is the combination of discrete and continuous, then the Beta process is the sum of two independent contributions. More details can be found in Thibaux & Jordan (2007).

5.1.3 Stick-breaking representation of the Beta process

Sampling B in (5.1.1) directly from the infinite Beta process is difficult because the inverse of the Lévy measure doesn't have a simple closed form. Wolpert and Ickstadt (1998b) introduce the inverse Lévy measure algorithm, a generic technique for pure-jump non negative Lévy processes that allows to sample from B. This technique generates the weights in (5.1.1) in decreasing order but requires inverting the incomplete beta function at each step which is computationally intensive. Paisley, John W and Zaas, Aimee K and Woods, Christopher W and Ginsburg, Geoffrey S and Carin, Lawrence (2010) and Broderick, Tamara and Jordan, Michael I and Pitman, Jim and others (2012) proposed the stick-breaking representation of the Beta process that provides a simple recursive procedure for obtaining the weights in equation (5.1.1). This approach provides an explicit representation of a draw B from the Beta process that doesn't require inverting the Lévy measure,

$$
\begin{aligned}
B &= \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)}) \delta_{\theta_{i,j}}, \qquad (5.1.3) \\
C_i &\sim Poisson(\gamma) \\
V_{i,j}^{(l)} &\overset{i.i.d.}{\sim} Beta(1, c) \\
\theta_{i,j} &\overset{i.i.d.}{\sim} B_0/\gamma.
\end{aligned}
$$

This is an analogue of Sethuraman's (1994) stick-breaking representation of the Dirichlet process. The only difference is that the weights resulting from the stick breaking representation of the Dirichlet process all come from one single stick, thus they add up to one. This is not the case in the stick-breaking representation of the Beta process where the weights come from different unit intervals. Therefore, they do not need to add to one. Nevertheless, the sum in (5.1.3) is finite almost surely. The shortcoming of the stick-breaking representation is discussed in Al Labadi and Zarepour (2015).

5.1.4 Finite Approximation of the Beta process

Another yet efficient way of generating the weights in (5.1.1) is by deriving the finite approximation of the Beta process defined by Paisley & Carin (2009) as follows

$$\begin{aligned} B_n &= \sum_{i=1}^{n} p_{i,n}\delta_{\theta_i} \\ p_{i,n} &\overset{i.i.d.}{\sim} \text{Beta}\Big(c\gamma_n, c\Big(1-\gamma_n\Big)\Big) \\ \theta_i &\overset{i.i.d.}{\sim} B_0/B_0(\Omega). \end{aligned}$$

Here $(p_{i,n})_{1\leq i\leq n}$ and $(\theta_i)_{1\leq i\leq n}$ are independent and $\gamma_n \to 0$ as $n \to \infty$.

Later Al Labadi & Zarepour (2015) prove that the finite approximation B_n converges weakly to Ferguson (1972) representation of the Beta process define in (5.1.1). In particular, they show, via proposition 3.21 of Resnick (1987), that as $n \to \infty$,

$$\begin{aligned} nP[p_{1,n} \in (x,1)] &= \frac{n\Gamma(c)}{\Gamma(c\gamma_n)\Gamma(c-c\gamma_n)}\int_x^1 p^{c\gamma_n-1}(1-p)^{c(1-\gamma_n)-1}dp \\ &\overset{v}{\to} \nu(x) = c\gamma\int_x^1 p^{-1}(1-p)^{c-1}dp, \end{aligned}$$

where, " $\overset{v}{\to}$ " denote vague convergence. One choice of γ_n is when $\gamma_n = \gamma/n$. Notice that for any $x > 0$, $\Gamma(x) = \Gamma(x+1)/x$, therefore $n/\Gamma(c\gamma/n) = c\gamma/\Gamma(c\gamma/n+1)$ when replacing x by $c\gamma/n$. Moreover, as $n \to \infty$

$$\frac{n}{\Gamma(c\gamma/n)} \to c\gamma$$

and

$$\frac{\Gamma(c)}{\Gamma(c-c\gamma/n)} \to 1.$$

By defining

$$\nu_n = \frac{\Gamma(c)}{\Gamma(c\gamma/n)\Gamma(c-c\gamma/n)}\int_x^1 p^{c\gamma/n-1}(1-p)^{c(1-\gamma/n)-1}dp,$$

we have as $n \to \infty$,

$$n\nu_n(x) \xrightarrow{v} \nu(x).$$

Moreover, they prove that as $n \to \infty$,

$$B_n = \sum_{i}^{n} \nu_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)\delta_{\theta_i} \xrightarrow{a.s.} B = \sum_{i=1}^{\infty} \nu^{-1}(\Gamma_i)\delta_{\theta_i}. \tag{5.1.4}$$

Interested readers can refer to Theorem 4 of Al Labadi & Zarepour (2015).

Based on the above results, Al Labadi & Zarepour (2015) describe an efficient algorithm to generate sample from an approximation of the Beta process $B \sim BP(c, B_0)$. Algorithm B gives details of the set of rules to sample from Beta process with continuous base.

Algorithm B: An approximation of the Beta process with continuous base

1. Choose a large positive integer n.
2. Generate $\theta_i \overset{i.i.d.}{\sim} B_{0/\gamma}$, for $i = 1, \cdots, n$.
3. Generate $E_i \overset{i.i.d.}{\sim} \text{Exp}(1)$ with p.d.f $f(x) = e^{-x}I(x > 0)$. Let $\Gamma_i = \sum_{k=1}^{i} E_k$.
4. Compute $(\nu_n^{-1}(\Gamma_i/\Gamma_{n+1}))_{1\leq i\leq n}$. This can be done by evaluating the quantile function of the beta distribution, $\text{Beta}(c\gamma/n, c(1-\gamma/n))$ at $1 - \Gamma_i/\Gamma_{n+1}$.

Figure 5.1 shows one sample path of the Beta process with $c = 0.8$, $B_0 \sim \text{Uniform}(0, 1)$ and $n = 15$ using Al Labadi & Zarepour (2014) algorithm. Similar to the Dirichlet process, vertical lines show the intensity of the weights in (5.1.4) at location $\theta_{(i)}$. Note that the weights are monotonically decreasing, therefore we are almost surely confident that there will not be an intense weight after the last weight observed.

Figure 5.2 depicts ten sample paths of the Beta process approximated by Al Labadi & Zarepour (2014) algorithm (Algorithm B) where we choose n equal to 100. The dashed line connected by dots represent the cumulative distribution of

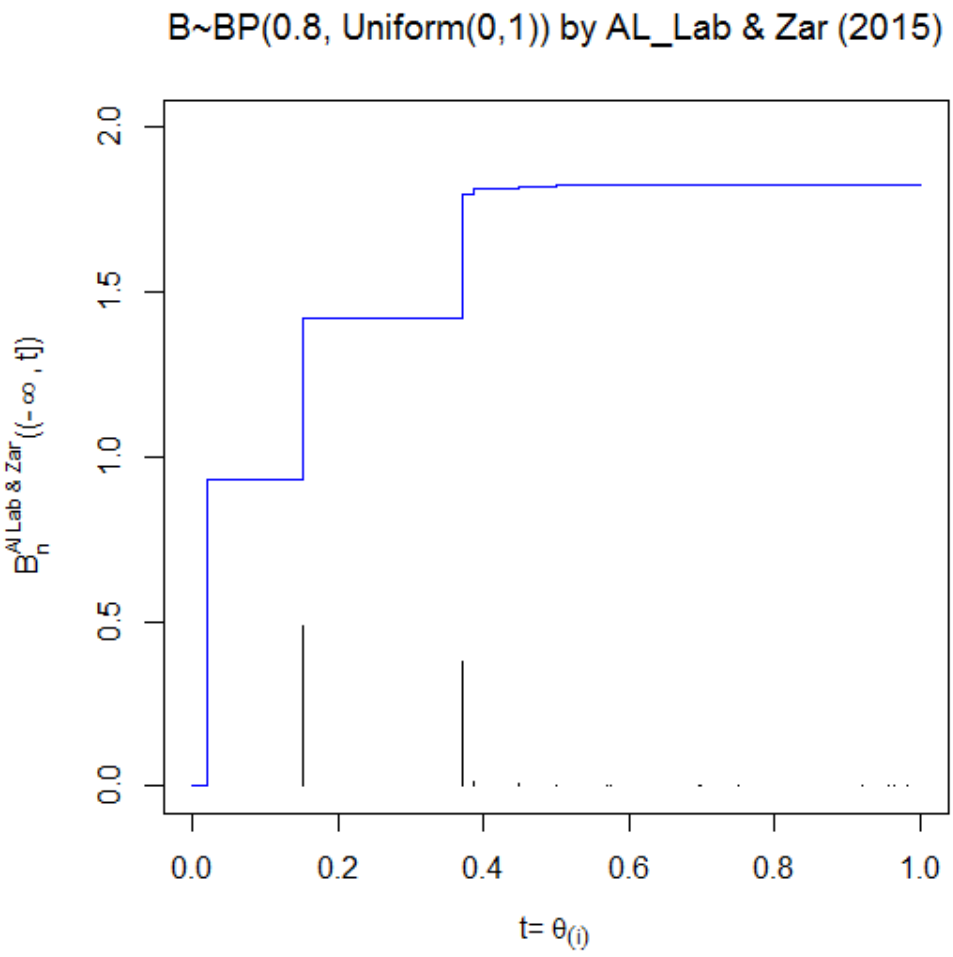

Figure 5.1: One sample path of the Beta process $B \sim BP(0.8,\text{Uniform}(0,1))$. The Beta process is approximated using Al Labadi & Zarepour (2014) algorithm (Algorithm B) with $n = 15$. The plot shows as well the weights in (5.1.4) by vertical lines at the associated atoms θ_i.

$B_0 \sim$ Uniform$(0, 1)$. The plot at the top represent the Beta process with $c = 1$ and $B_0 =$Uniform$(0, 1)$ where as the plot at the bottom represent the Beta process with same base measure B_0 but with $c = 20$. The concentration parameter c express the strength of belief in B_0, this is shown in Figure 5.2 where with increase value of c the Beta process approaches to $B_0 \sim$Uniform$(0, 1)$. This behaviour support our previous discussion on the basic properties of the Beta process.

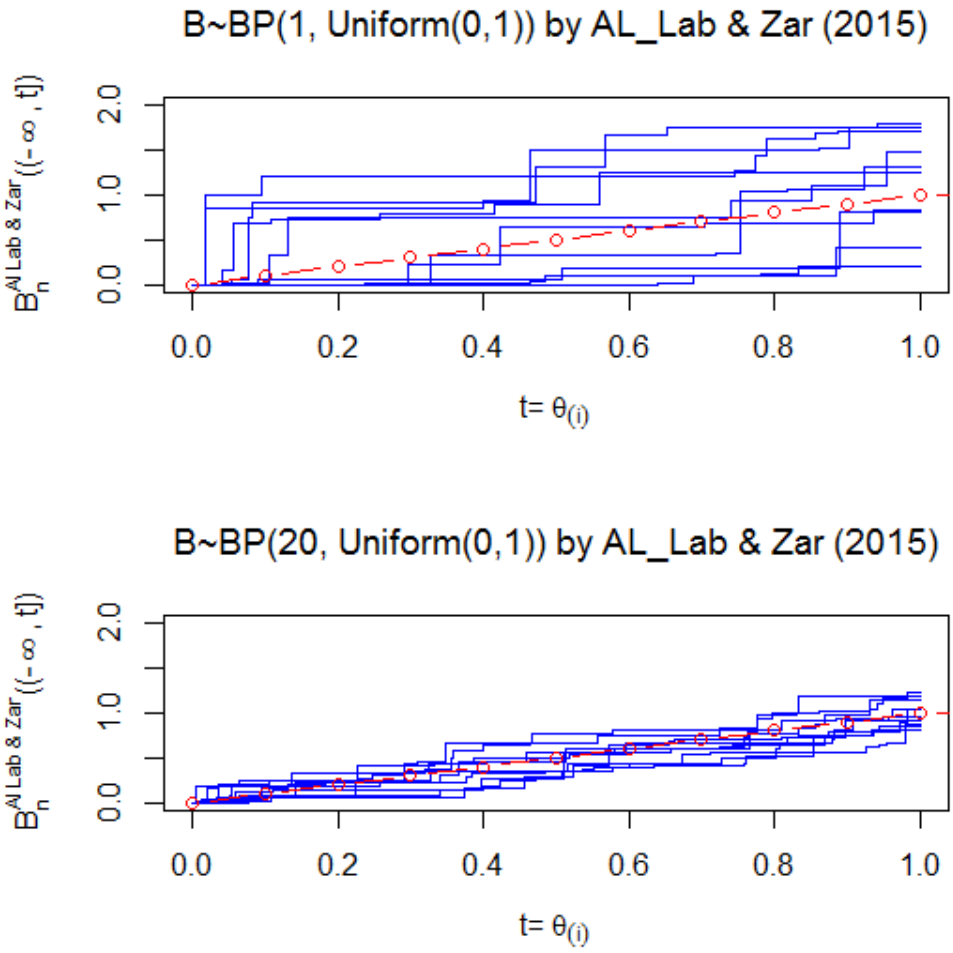

Figure 5.2: The plot at the top shows ten sample paths of the Beta process with $c = 1$ and $B_0 \sim$Uniform$(0, 1)$. The plot at the bottom shows ten sample paths of the Beta process with same base measure B_0 but with $c = 20$. We use Algorithm B to approximate the Beta process in both plots with $n = 100$. The dashed line connected by dots in both plots represents the cumulative distribution of the base measure $B_0 \sim$ Uniform$(0, 1)$.

5.2 Beta-Bernoulli process

Thibaux & Jordan (2007) show that the Beta process is the conjugate prior for the Bernoulli process. This conjugacy extends the conjugacy between the Beta and Bernoulli distribution. In this section, we define the Bernoulli process and we introduce an extension to the Beta process known by the Beta-Bernoulli process.

5.2.1 Definition of the Beta-Bernoulli process

Definition 5.2.1 (Bernoulli Process) *Let H be a measure on Ω. The* Bernoulli process *Y with base measure H, written $Y \sim BeP(H)$, is a completely random variable with Lévy measure*

$$\pi(d\theta, dp) = \delta_1(dp)H(d\theta),$$

where δ_1 is a measure concentrate at 1.

When H is a diffuse measure (or has no points of discontinuity), Y has an underlying Poisson process with intensity H. It can be proven that Y has the following representation

$$\begin{aligned} Y &= \sum_{k=1}^{K} \delta_{\theta_k} \\ K &\sim \text{Poisson}(H(\Omega)) \\ \theta_k &\overset{i.i.d.}{\sim} H/H(\Omega). \end{aligned} \tag{5.2.1}$$

When H is discrete (or consist of fixed points of discontinuity) of the form $H = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}$, then the Bernoulli process is defined at the same locations θ_k as follows

$$\begin{aligned} Y &= \sum_{i=k}^{\infty} b_k \delta_{\theta_k} \\ b_k &\sim \text{Bernoulli}(p_k). \end{aligned} \tag{5.2.2}$$

When the base measure is the mixture of discrete and continuous, the Beta-Bernoulli process is the superposition of two independent contributions.
The Beta process is useful as parameter for the Bernoulli process. When combining both processes together such that the base measure of the Bernoulli process is chosen to be the Beta process, the resulting process is known as the Beta-Bernoulli process.

Definition 5.2.2 (Thibaux and Jordan (2007)) *The Beta-Bernoulli process, denoted by $X \sim BeP(B)$, is a completely random measure with base measure the Beta process. The model is defined as follows*

$$X|B \overset{i.i.d.}{\sim} BeP(B)$$
$$B \sim BP(c, B_0),$$

where B_0 and c are define in Definition 5.1.1.

5.2.2 Series representation of the Beta-Bernoulli process

A draw from the Beta process $B \sim BP(c, B_0)$ generates a set of atoms $(p_i, \theta_i)_{i\geq 1}$, where p_i is the weight in (5.1.1) calculated at location θ_i. The Beta-Bernoulli has a series representation as follows

$$X = \sum_{i=1}^{\infty} b_i \delta_{\theta_i}, \tag{5.2.3}$$

where $b_i \sim \text{Bernoulli}(p_i)$ at location θ_i.
Note that from the strong law of large numbers, we get

$$E[X] = E[E[X|B]] = E[B] = B_0.$$

This property of the Beta-Bernoulli process ensure that the series representation in (5.2.3) converges even though there is an infinite terms.

5.2.3 Beta process conjugate prior for the Bernoulli process

The Beta process is the conjugate prior for the Bernoulli process. Following Thibaux & Jordan (2007) notation, suppose we observe N data points such that

$$\begin{aligned} X_1, \cdots, X_N | B &\overset{i.i.d.}{\sim} BeP(B) \\ B &\sim BP(c, B_0), \end{aligned}$$

then applying Theorem 3.3 of Kim (1999) the posterior distribution of B given the observed data $X_1, \cdots, X_N$ is

$$\begin{aligned} B|X_1, \ldots, X_N &\sim BP(c+N, B_n) \\ B_n &= \frac{c}{c+N} B_0 + \frac{1}{c+N} \sum_{i=1}^{N} X_i \\ B_n &= \frac{c}{c+N} B_0 + \frac{N}{c+N} \frac{\sum_{i=1}^{N} X_i}{N}. \end{aligned} \tag{5.2.4}$$

As $N \to \infty$, the base measure approaches the empirical distribution of X_i. And as $c \to \infty$ the base measure approaches to the prior guess B_0. We recall that the $X_i's$ are Beta-Bernoulli random measures not a random number.

The plot at the top of Figure 5.3 shows one sample path of the Beta process with $c = 1$, $B_0 \sim \text{Uniform}(0, 1)$. We use Paisley & Carin (2009) approximation and we follow Al Labadi & Zarepour (2015) algorithm with $n = 15$ to approximate the Beta process. Vertical lines show the intensity of the weights in (5.1.4) at location $\theta_{(i)}$. The plot at the bottom of Figure 5.3 shows 10 draws from the Beta-Bernoulli process, one per line. We use the Beta process displayed at the top of Figure 5.3 as the base measure to the Bernoulli process. A draw is a set of points $(b_i, \theta_i)_{1 \leq i \leq 15}$, such that $b_i \sim \text{Bernoulli}\ (B\{\theta_i\})$. Thus, each line has a black dot at position θ_i with probability $(B\{\theta_i\})$, $B \sim BP(1, \text{Uniform}(0, 1))$.

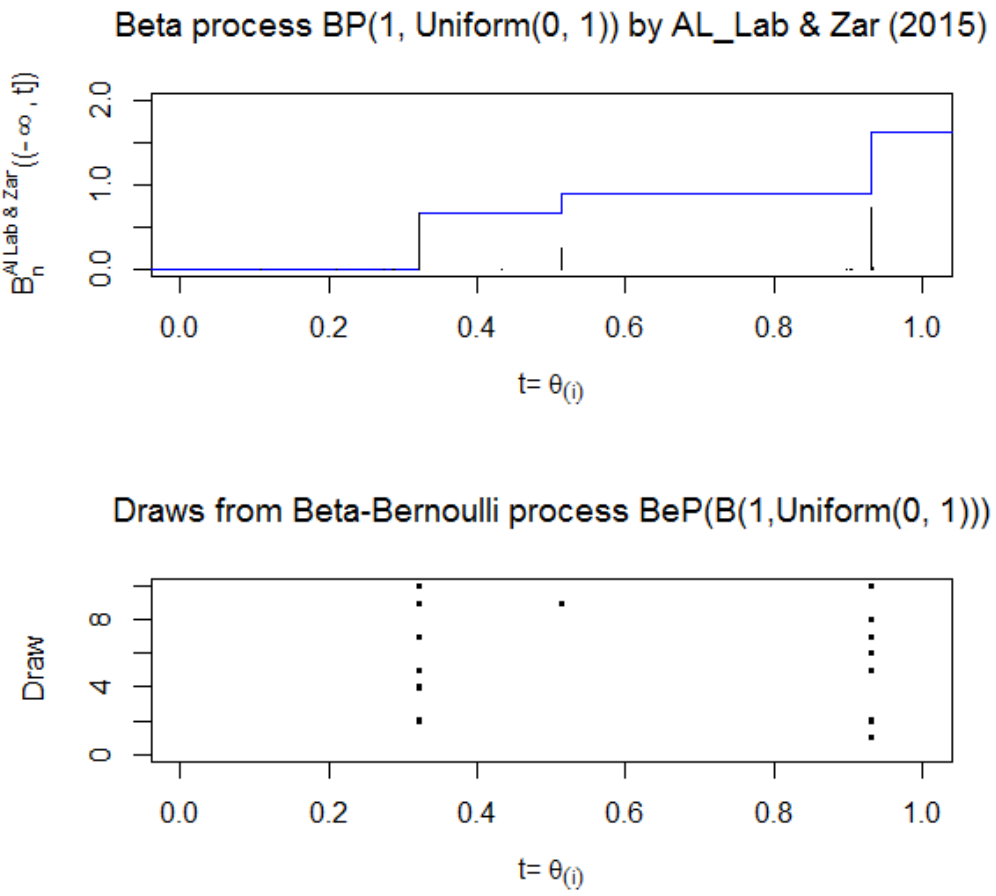

Figure 5.3: The plot at the top depicts one sample path of the Beta process with $c = 1$, $B_0 \sim$Uniform$(0, 1)$. The Beta process is approximated using Algorithm B with $n = 15$. The vertical lines shows the intensity of the weights in (5.1.4). The plot at the bottom shows 10 draws of the Bernoulli processes, one per line, with base measure the Beta process (displayed at the top of the figure).

5.3 Applications in Latent Feature Model

The Latent feature model has been widely used in many applications in different disciplines, particularly in machine learning. It has been used to decompose the data into a small number of components. Example of applications in nonparametric latent feature model are well explained in Miller & Jordan (2009) and in Paisley & Carin (2009).

In this section we describe two methodologies to sample from the posterior of a latent feature model. The first method is known as the Indian Buffet Process in the Computer Science community. The second method is a new technique we propose to sample from the posterior. This new technique focuses on efficiency. We describe in details the sampling techniques in both methodologies in the next sections.

5.3.1 Matrix Z

The Beta-Bernoulli process has been widely used as a prior in applications on latent feature models. For instance, suppose we observe N data points $Z_1, Z_2, \ldots, Z_N$ such that N is large. In the Bayesian framework, the goal in most applications on latent feature models is to infer a binary matrix $\mathbf{Z}$, often called factor loadings, where $\mathbf{Z}$ is an $N \times K$ matrix. The number of row N represents the number of observation and K represents the number of latent features or attributes. The dimension of K is infinite, thus the rows of $\mathbf{Z}$ consist of an infinite collection of factor loadings. The matrix $\mathbf{Z}$ is constructed in a way such that if the i^{th} observation possesses the k^{th} feature then Z_{ik} is equal to 1 otherwise it is equal to 0 where $1 \leq i \leq N$ and $1 \leq k \leq K$. Notice that each observation possesses one or more features. Therefore, each row of $\mathbf{Z}$ ideally has multiple $1's$. When a data set is modelled with a latent feature, there is a certain belief that the possession of any number of these features has an effect on the observed data. Moreover, there is a strong belief that when two entities have a great number of features in common, those two entities will most probably have the

same structure or behaviour. On most applications involving latent feature models, the Beta-Bernoulli process X_i is mapped to Z_i. The matrix $\mathbf{Z}$ with rows, $Z_1, \ldots, Z_N$, is constructed in a way such that θ_k labels the k^{th} column in $\mathbf{Z}$ and p_i represents the weight that observation i possesses feature k. Thus, entry at position $Z_{i,k}$ is 1 with probability p_i. Note that by Campbell's theorem, the base measure B of the Beta-Bernoulli process has finite mass. Therefore, the columns of $\mathbf{Z}$ have a finite number of non-zero entries. Storing only non-zero columns make $\mathbf{Z}$ a finite matrix which is computationally manageable. It is important to note that columns of the matrix $\mathbf{Z}$ are exchangeable.

To make the connection between the Beta-Bernoulli process and the matrix $\mathbf{Z}$, let us consider an example to describe this mapping. We extract the set of pairs $(p_k, \theta_k)_{1 \leq k \leq n}$ from the Beta process $B \sim BP(1, \text{Uniform}(0, 1))$ displayed in Figure 5.3 (the plot at the top). The first 9 estimated values of (p_k, θ_k) are reported in Table 5.1.

p_k	0.73	0.66	0.24	0.0087	0.0024	0.0011	0.00027	0.00019	0.000064
θ_k	0.93	0.32	0.51	0.93	0.27	0.32	0.89	0.11	0.9

Table 5.1: The table shows the first nine set of pairs $(p_k, \theta_k)_{1 \leq k \leq 9}$ extracted from a draw of the Beta process, $B \sim BP(1, \text{Uniform}(0, 1))$. The Beta process is approximated using Algorithm B with $n = 15$.

We extract from the first draw of the Beta-Bernoulli process, $X_1 \sim BeP(B)$, displayed in Figure 5.3 the set of pairs $(b_k, \theta_k)_{1 \leq k \leq 9}$ such that $b_k \sim \text{Bernoulli}(p_k)$. We report the first nine estimated values in Table 5.2.

For each set of pairs of the form $(b_k = 1, \theta_k)_{1 \leq k \leq n}$, θ_k is mapped in $\mathbf{Z}$ to represent a new latent feature. In particular, we choose to map θ_1 to label the first column of $\mathbf{Z}$, and θ_2 to label the second column of $\mathbf{Z}$ and so on. As we discussed earlier in this section, columns of the matrix $\mathbf{Z}$ are exchangeable therefore labelling columns of $\mathbf{Z}$

b_k	1	0	0	0	0	0	0	0	0
θ_k	0.93	0.32	0.51	0.93	0.27	0.32	0.89	0.11	0.9

Table 5.2: The table depicts the first nine values $(b_k, \theta_k)_{1\leq k\leq 9}$ extracted from a draw of a Bernoulli process with base measure $B \sim BP(1, \text{Uniform}(0, 1))$. Recall that $b_k \sim \text{Binomial}(p_k)$, where p_k is the probability displayed in Table 5.1.

by $\theta's$ can be done in a different order. Note that from Table 5.1 there is three non negligible weights. Then at most there will be three visible atoms $\theta's$ and consequently at most three labelled columns in $\mathbf{Z}$ representing a latent feature . Indeed, Table 5.2 shows only one pair of the form $(b_k = 1, \theta)$, then we map θ in $\mathbf{Z}$ to represent a new label.

Therefore, mapping X_1 to Z_1 gives

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Note that the value of K (the number of column of $\mathbf{Z}$) should be relatively large but for illustration purposes we choose $K = 10$. We use the same mapping technique to map $X_2, \ldots, X_{10}$ in $Z_2, \ldots, Z_{10}$. We get

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.3.1}$$

Note that zero's column means that there is no label (or latent feature) associated to that column yet.

5.3.2 Updating the matrix Z using Methodology I

Updating the matrix $\mathbf{Z}$ with a new observation, Z_{N+1} based on what have been observed $Z_1, \ldots, Z_N$ is the key in most nonparametric Bayesian applications. Recall that

$$B|X_1, \ldots, X_N \sim BP\left(c + N, \frac{c}{c+N}B_0 + \frac{1}{c+N}\sum_{k=1}^{N} X_k\right).$$

In the context of latent feature modelled by the matrix $\mathbf{Z}$, the posterior of the Beta process has another meaning for its discrete base measure as follows

$$B|X_1, \ldots, X_N \sim BP\left(c + N, \frac{c}{c+N}B_0 + \sum_{k=1}^{K} \frac{m_{N,k}}{c+N}\delta_{\theta_k}\right), \tag{5.3.2}$$

where, $m_{N,k} = \sum_{i=1}^{N} I(Z_{i,k} = 1)$. In words, $m_{N,k}$ is the number of time among N, feature k has been observed. Notice that the posterior base measure is constructed of

two components, one continuous and another discrete, it is reasonable to sample from each component independently since they do not overlap. Therefore, we can write the Bernoulli process as the sum of two independent Bernoulli processes as follows

$$X_{N+1}|X_1,\cdots,X_N \sim BeP\left(\frac{c}{c+N}B_0\right) + BeP\left(\sum_{k=1}^{K}\frac{m_{N,k}}{c+N}\delta_{\theta_k}\right)$$
$$X_{N+1}|X_1,\cdots,X_N \sim \qquad F \qquad + \qquad R,$$

where

$$F \sim BeP\left(\frac{c}{c+N}B_0\right)$$
$$R \sim BeP\left(\sum_{k=1}^{K}\frac{m_{N,k}}{c+N}\delta_{\theta_k}\right).$$

The algorithm to updated the matrix $\mathbf{Z}$ is define in Methodology I. It is worth mentioning that this technique is known by the two-parameter (c,γ) generalization of the Indian buffet process (Ghahramani & Griffiths, 2005), a well know process in the Computer Science community.

Methodology I

1. Sample a draw from $F \sim BeP\left(\frac{c}{c+N}B_0\right)$.
2. Map every pair $(1,\theta_k')$ extracted from F in step 1 to the matrix $\mathbf{Z}$.
3. Sample independently a draw from $R \sim BeP\left(\sum_{k=1}^{K}\frac{m_{N,k}}{c+N}\delta_{\theta_k}\right)$.
4. Map every pair of the form $(b_k = 1,\theta_k)$ extracted from R in step 4 to the matrix $\mathbf{Z}$.

See next discussion for the details of Methodology I, especially for the mapping in step 2 and 4.

Updating the matrix Z from the Bernoulli process with continuous base

Updating the matrix $\mathbf{Z}$ with a new observation Z_{N+1} can be done in two stages. First, it involves sampling from the two Bernoulli processes F and R independently, then mapping each process independently to the row Z_{N+1}. Sampling the Bernoulli process F with continuous base measure can be done as examined in (5.2.1)

$$F \sim BeP\left(\frac{c}{c+N}B_0\right)$$
$$F = \sum_{k=1}^{K_1} \delta_{\theta_k^{'}}$$

where

$$K_1 \sim \text{Poisson}\left(\frac{c\gamma}{c+N}\right)$$
$$\theta_k^{'} \sim B/B(\Omega) = B_0/\gamma. \qquad 1 \leq k \leq K_1$$

Note that for each pair of the form $(1, \theta_k^{'})_{1\leq k\leq K_1}$ extracted from F, $(\theta_k^{'})_{1\leq k\leq K_1}$ are the new label of features added to the matrix $\mathbf{Z}$. Let $k_1', \ldots, k_{K_1}'$ be K_1 indices of zero columns of $\mathbf{Z}$, then we let $(Z_{N+1,k'} = 1)_{k_1'\leq k'\leq k_{K_1}'}$.

Updating matrix Z from the Bernoulli process with discrete base

Sampling the Bernoulli process R with discrete base measure can be done as discussed in (5.2.2)

$$R \sim BeP\left(\sum_{k=1}^{K} \frac{m_{N,k}}{c+N}\delta_{\theta_k}\right),$$
$$R = \sum_{k=1}^{\infty} b_k \delta_{\theta_k}$$
$$b_k \sim \text{Bernoulli}\left(\frac{m_{N,k}}{c+N}\right)$$

Note that θ_k in the discrete base measure represent an existing latent feature in $\mathbf{Z}$. Therefore at each non zero column k of $\mathbf{Z}$, $Z_{N+1,k} = 1$ with probability $m_{N,k}/(c+N)$.

5.3.3 Updating the matrix Z using Methodology II

In this section we describe an alternative way of updating the matrix $\mathbf{Z}$. This can be done by sampling first from the posterior of the Beta process then from the Beta-Bernoulli process.

$$
\begin{aligned}
B|X_1,\ldots,X_N &\sim BP\left(c+N, \frac{c}{c+N}B_0 + \sum_{k=1}^{K}\frac{m_{N,k}}{c+N}\delta_{\theta_k}\right),\\
B|X_1,\ldots,X_N &\sim BP\left(c+N, \frac{c}{c+N}B_0\right) + BP\left(c+N, \sum_{k=1}^{K}\frac{m_{N,k}}{c+N}\delta_{\theta_k}\right)\\
B|X_1,\ldots,X_N &\sim \qquad B^{Cont} \qquad + \qquad B^{Disc},
\end{aligned}
$$

where

$$
\begin{aligned}
B^{Cont} &\sim BP\left(c+N, \frac{c}{c+N}B_0\right)\\
B^{Disc} &\sim BP\left(c+N, \sum_{k=1}^{K}\frac{m_{N,k}}{c+N}\delta_{\theta_k}\right).
\end{aligned}
$$

Now sampling a new observation can be done as follows

$$
\begin{aligned}
X_{N+1}|X_1,\cdots,X_N &\sim BeP(B^{Cont}) + BeP(B^{Disc})\\
X_{N+1}|X_1,\cdots,X_N &\sim \quad S \quad + \quad T \quad ,
\end{aligned}
$$

where,

$$
\begin{aligned}
S &\sim BeP(B^{cont})\\
T &\sim BeP(B^{Disc}).
\end{aligned}
$$

The following gives further details on the methodology.

Methodology II

1. (a) Sample a draw from B^{Cont}. Extract the set of pairs $(p_k^{Cont}, \theta_k')_{1\leq k\leq n}$ from B^{Cont}.

   (b) Sample a draw from $S \sim BeP(B^{Cont})$. Extract the set of pairs $(b_k^{Cont}, \theta_k')_{1\leq k\leq n}$.

   (c) Map every pair of the form $(b_k^{Cont} = 1, \theta_k')$ into the Matrix $\mathbf{Z}$.

2. (a) Sample a draw from B^{Disc}. Extract the set of pairs $(p_k^{Disc}, \theta_k)_{1\leq k\leq n}$.

   (b) Sample a draw from $T \sim BeP(B^{Disc})$. Extract the set of pairs $(b_k^{Disc}, \theta_k)_{1\leq k\leq n}$.

   (c) Map every pair of the form $(b_k^{Disc} = 1, \theta_k)$ in Matrix $\mathbf{Z}$.

Refer to the next discussion for details on updating the matrix $\mathbf{Z}$ using Methodology II.

Updating the matrix Z from the Beta-Bernoulli with continuous base

Updating $\mathbf{Z}$ would first involve sampling B^{Cont} and B^{Disc} independently then sampling from the two Beta-Bernoulli processes S and T independently. Following our discussion in (5.1.1), the Beta process B^{Cont} with continuous base has a series representation

$$B^{Cont} = \sum_{k=1}^{\infty} p_k^{Cont} \delta_{\theta_k'}.$$

Using Paisley & Carin (2009) finite approximation of the Beta process in (5.1.4), we have

$$B_n^{Cont} = \sum_{k=1}^{n} p_{k,n}^{Cont} \delta_{\theta_k'}$$

$$p_{k,n}^{Cont} \overset{i.i.d.}{\sim} Beta\left(\frac{c * \gamma*}{n}, c * \left(1 - \frac{\gamma*}{n}\right)\right) \tag{5.3.3}$$

$$\theta_k' \overset{i.i.d.}{\sim} B_0^*/\gamma^*, \tag{5.3.4}$$

where

$$\begin{aligned} c^* &= c + N \\ \gamma^* &= \frac{c}{c+N}\gamma \\ B_0^* &= \frac{c}{c+N}B_0. \end{aligned} \tag{5.3.5}$$

Replacing c^*, γ^* and B_0^* in (5.3.3) and (5.3.4) we get,

$$\begin{aligned} B_n^{Cont} &= \sum_{k=1}^{n} p_{k,n}^{Cont}\delta_{\theta_k'} \\ p_{k,n}^{Cont} &\overset{i.i.d.}{\sim} Beta\left(\frac{c\gamma}{n}, N + c - \frac{c\gamma}{n}\right) \\ \theta_k' &\overset{i.i.d.}{\sim} B_0/\gamma, \end{aligned}$$

The Beta-Bernoulli process $S \sim BeP(B_n^{Cont})$ has series representation

$$\begin{aligned} S &\sim BeP(B_n^{Cont}) \\ S &= \sum_{k=1}^{n} b_k^{Cont}\delta_{\theta_k'} \\ b_k^{Cont} &\sim \text{Bernoulli}(p_{k,n}^{Cont}). \end{aligned}$$

Note that when the Bernoulli process has a continuous base, the resulting process generates new locations θ''s. We map every pair of the form $(1, \theta_k')$ in Z_{N+1} such that every θ_k' label a zero column of $\mathbf{Z}$. Let k' represent the indices of a zero column of $\mathbf{Z}$, then we let $Z_{N+1,k'} = 1$.

Going back to our simulated example shown in Figure 5.3, we approximate the posterior of the Beta process with continuous base, B^{Cont} using Al Labadi & Zarepour (2015) algorithm for $n = 15$ and with the updated parameters $c^* = 11$ and $B_0^* \sim \frac{1}{11}\text{Uniform}(0,1)$. We extract from B^{Cont} the set of pairs $(p_k^{Cont}, \theta_k')_{1\leq k\leq 15}$. Table 5.3 shows some preliminary values of those pairs. Recall that the weights generated by Al Labadi & Zarepour (2015) are in decreasing order almost surely, therefore we omitted

p_k^{Cont}	0.068	0.00088	0.0000065	0.00011	0.000057
θ_k'	0.12	0.87	0.47	0.34	0.29

Table 5.3: The table shows some preliminary values of the pairs (p_k^{Cont}, θ_k') extracted from $B^{Cont} \sim BP(11, \frac{1}{11}\text{Uniform}(0,1))$.

the last 10 weights in the table because we are almost sure that there will not be a non negligible weight beyond that point.

We sample a draw from S, the set of pairs (b_k^{Cont}, θ_k') are reported in Table 5.4.

b_k^{Cont}	1	0	0	0	0
θ_k'	0.12	0.87	0.47	0.34	0.29

Table 5.4: The table depicts some preliminary values of the pairs (b_k^{Cont}, θ_k') extracted from $S \sim BeP(B^{Cont})$, the Beta-Bernoulli process with base measure B^{Cont}.

Note from Table 5.4 that there is only one pair such that $(b_k^{Cont} = 1, \theta_k' = 0.12)$. To map this pair in $\mathbf{Z}$, we choose the fourth column (zero column) of $\mathbf{Z}$ to label this new feature and we let $Z_{11,4} = 1$. Thus, we have

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Updating the matrix Z from the Beta-Bernoulli with discrete base

To finish updating the matrix $\mathbf{Z}$ we are left with sampling from the Beta process $B^{Disc} \sim BP\left(c+N, \sum_{k=1}^{K} \frac{m_{N,k}}{c+N}\delta_{\theta_k}\right)$ with discrete base B_0^{Disc}, then sampling from the Beta-Bernoulli process $T \sim BeB(B^{Disc})$. Following our discussion in (5.1.2), the Beta process B^{Disc} has series representation

$$\begin{aligned} B_n^{Disc} &= \sum_{k=1}^{n} p_k^{Disc}\delta_{\theta_k} \\ B_0^{Disc} &= \sum_{k=1}^{n} \frac{m_{N,k}}{c+N}\delta_{\theta_k} \\ p_k^{Disc} &\sim \text{Beta}(m_{N,k}, N - m_{N,k} + c). \end{aligned} \tag{5.3.6}$$

The Beta-Bernoulli process T has series representation

$$\begin{aligned} T &= \sum_{k=1}^{\infty} b_k^{Disc}\delta_{\theta_k} \\ b_k^{Disc} &\sim \text{Bernoulli}(p_k^{Disc}). \end{aligned} \tag{5.3.7}$$

We extract from the Beta-Bernoulli process T the set of pairs $(b_k^{Disc}, \theta_k)_{1\leq k\leq 15}$. Let $k_1', \ldots, k_n'$ be the indices of column of $\mathbf{Z}$ such that there is at least on row containing an atom 1, then we let $(Z_{N+1,k'} = b_k^{Disc})_{k_1'\leq k'\leq k_n'}$.

Going back to our simulated example, the first three column of $\mathbf{Z}$ in (5.3.1) have at least one row with an atom 1. Those columns are labelled previously by θ_1, θ_2 and θ_3 respectively. Table 5.5 shows the estimated values of p_k^{Disc} calculated based on (5.3.6), where Table 5.6 shows the estimated values of b_k^{Disc} calculated based on (5.3.7).

p_k^{Disc}	0.90	0.62	0.0055
θ_k	0.93	0.32	0.51

Table 5.5: The table shows the set of pairs $(p_k^{Disc}, \theta_k)_{1\leq k\leq 3}$ such that p_k^{Disc} is calculated based on (5.3.6).

b_k^{Disc}	1	1	0
θ_k	0.93	0.32	0.51

Table 5.6: The table shows the set of pairs $(b_k^{Disc}, \theta_k)_{1\leq k\leq 3}$, where $b_k^{Disc} \sim$ Bernoulli(p_k^{Disc}).

The matrix $\mathbf{Z}$ is updated with the Beta-Bernoulli process T as follows:

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Figure 5.4 shows a draw from the posterior of the Beta-Bernoulli process X_{11} given $X_1, \ldots, X_{10}$. The draw are the set of pairs of the form $(b_k^{Cont} = 1, \theta_k')$ and $(b_k^{Disc} = 1, \theta_k)$. The triangle pointed down represent a sampling of the Beta-Bernoulli process, T with discrete base measure and the triangle pointed up represent a sampling of the Beta-Bernoulli process, S with continuous base measure.

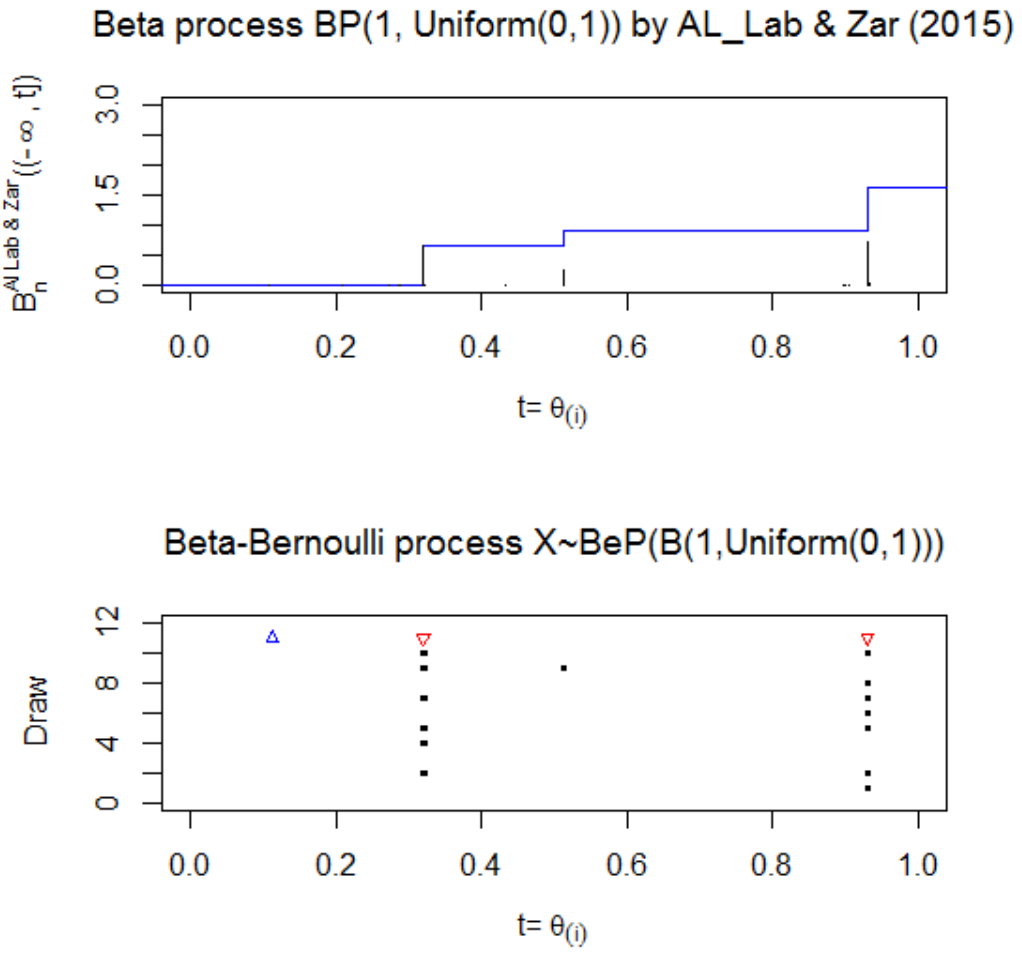

Figure 5.4: The plot at the top shows one draw of the Beta process $B \sim BP(1, \text{Uniform}(0,1))$ approximated by Algorithm B with $n = 15$. The plot at the bottom shows 10 draws of the Beta-Bernoulli process with base measure B. Draws are represented in the plot by dots at each pair of the form $(b_k = 1, \theta_i)$ generated from the Beta-Bernoulli process. The plot at the bottom shows as well one updated draw of the Beta-Bernoulli process given the 10 other observations. Triangle pointed down represent the update contributed by the discrete base, and triangle pointed up represent the update contributed by the continuous part of the updated Beta process.

5.3.4 Nonparametric Latent Feature Models for Link Prediction

Miller & Jordan (2009) introduce the nonparametric latent feature relational model used for social network data. This model seeks to extract latent structure representing the properties of individual entities from the observed data. In particular, we observe the relationships (or links) between a set of entities in a network and we try to predict unobserved links. For example, consider Facebook as our social network. In such network, we only know some subset of people who are friends with and some other who are not. The goal would be to predict which other people are likely to become friend.

5.3.5 Basic model

Assume we observe the directed links between a set of N entities. Let Y be an $N \times N$ binary matrix that contains these links. That is $Y_{ij} = 1$ if entity i is linked to entity j ($i \to j$), $Y_{ij} = 0$ if entity i is not linked to entity j, and Y_{ij} is left empty if we don't observe any link. The link can stand for different meanings such as "send a friend request or not", "friend or not", "colleague or not" or any other relationship. Depending of the relation, the matrix Y can be symmetric or asymmetric. The model decompose the binary matrix Y in two matrices Z and W. Where Z is a $N \times K$ matrix; each row of Z corresponds to an entity and each column corresponds to a feature such that $Z_{ik} = 1$ if entity i has feature k, otherwise take $Z_{ik} = 0$. For instance, we can have a separate feature for "statistician", "female", "athlete", and "painter" and the presence or absence of each of these features is what defines each person and determines their relationships. And W is a $K \times K$ real valued matrix where the (k, k') entry of W is $w_{k,k'}$. The value $w_{kk'}$ is the weight that affects the probability of having a link from entity i to entity j if both entity i has feature k and entity j has feature k'. If we are looking at the relation "send a friend request", then

the weight at the (statistician, athlete) entry of W would correspond to the weight that a statistician would send a friend request to an athlete. A positive weights would correspond to an increased probability, a negative weight would correspond to a decreased probability, and a zero weight would indicate that there is no correlation between these two features and the observed relation. Thus, the observed relations depend on binary valued latent features that influences its relations, weighted with a set of known covariates.

Following the notation of Miller & Jordan (2009), the model is defined as follows:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(\sigma(Z_i W Z_j^T)) \\ B &\sim \text{BP}(1, B_0) \\ Z &\sim \text{BeP}(B) \\ w_{kk'} &\sim \text{Normal}(0, \sigma_w^2), \end{aligned}$$

where, $\sigma(\cdot)$ is a function that transforms values on $(-\infty, \infty)$ to $(0, 1)$ such as the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$. Note that the Beta process B can be approximated using Algorithm B. Thus $B = \sum_{i=1}^{n} p_k \delta_{\theta_k}$ and $Z_{ik} \sim \text{Bernoulli}(p_k)$. Another contribution in this thesis is to modify (improve) the mathematical notation of what most Computer Scientist has adopted in machine learning community. The matrix Z is the map of the Beta-Bernoulli process $X \sim \text{BeP}(B)$ to Z we discussed earlier in Section 5.3.1. It is worth mentioning that Miller & Jordan (2009) have put an Indian Buffet Process IBP(α) prior on the matrix Z. Jordan (2007) proves that the Beta process with concentration parameter $c = 1$ and base measure B_0 is the IBP(α), with $\alpha = B_0(\Omega)$.

Given the full set of observation Y, we wish to infer the posterior distribution of the feature matrix Z and the weights W. This can be done by using Bayes' theorem, $p(Z, W|Y) \propto P(Y|Z, W)P(Z)P(W)$ with an independent prior on Z and W. For details on inference, interested reader can refer to Miller & Jordan (2009).

Appendix A

Definitions of background knowledge

In this appendix we discuss some properties of random measure which are mentioned throughout this thesis.

Definition A.0.1 (Convergence of Random Measures) *(Kallenberg, 1983) Let E be a Polish space and $\mathcal{B}(E)$ be a Borel σ-algebra generated by the open sets in E. A measure μ is called Radon if $\mu(K) < \infty$ for any compact set $K \in E$. Let $M_+(E)$ be the space of Radon measures in E. Let $\mathcal{M}_+(E)$ be the smallest σ-algebra of subsets of $M_+(E)$ making the maps $\mu \to \mu(f) = \int f(x)d\mu(x)$ from $M_+(E)$ to $\mathbb{R}$ measurable for all functions $f \in C_K^+(E)$, where $C_K^+(E)$ denotes the set of continuous functions $f : E \to [0, \infty)$ with compact support. Note that, $\mathcal{M}_+(E)$ is the Borel σ-algebra generated by the topology of vague convergence. If μ_n, $\mu \in M_+(E)$, we say that $(\mu_n)_n$ converges vaguely to μ, if $\mu_n(f) \xrightarrow{v} \mu(f)$ for any $f \in C_K^+(E)$.*

A random measure on E is any measurable map ξ defined on a probability space $(\Omega, \mathcal{F}, P)$ with values in $(M_+(E), \mathcal{M}_+(E))$. If ξ_n and ξ are random measures on E, we say that $(\xi_n)_n$ converges in distribution to ξ (we write $\xi_n \xrightarrow{d} \xi$) if $\{P \circ \xi_n^{-1}\}_n$ converges weakly to $P \circ \xi^{-1}$. By Theorem 4.2 of Kallenberg (1983), $\xi_n \xrightarrow{d} \xi$ if and

only if $\xi_n(f) \to \xi(f)$, *i.e.*

$$\int_E f(x)\xi_n(dx) \to \int_E f(x)\xi(dx), \quad \forall f \in C_K^+(E).$$

We say that $(\xi_n)_n$ *converges vaguely almost surely to* ξ *(and we write* $\xi_n \overset{a.s.}{\to} \xi$*) if there exist a set* $\widetilde{\Omega} \in \mathcal{F}$ *with* $P(\widetilde{\Omega}) = 1$ *such that* $\forall w' in \widetilde{\Omega}, \quad \xi_n(w,\cdot) \overset{v}{\to} \xi(w,\cdot)$, *i.e.*

$$\int_E f(x)\xi_n(w,dx) \to \int_E f(x)\xi(w,dx), \quad \forall f \in C_K^+(E).$$

The space $M_+(E)$ *endowed with the vague topology is a complete separable metric space Resnick (1987). For more details about random measures refer to Kallenberg (1983).*

Bibliography

[1] Al Labadi, Luai, and Mahmoud Zarepour (2014). On simulations from the two-parameter Poisson-Dirichlet process and the normalized inverse-Gaussian process. *Sankhya A 76.1*, 158-176.

[2] Al Labadi, Luai (2012). *On New Constructive Tools in Bayesian Nonparametric Inference*. PhD thesis, University of Ottawa .

[3] Labadi, Luai Al and Zarepour, Mahmoud (2015). *On Approximations of the Beta Process in Latent Feature Models. arXiv preprint arXiv:1411.3434.*

[4] Al Labadi, Luai and Zarepour, Mahmoud and others (2013). On asymptotic properties and almost sure approximation of the normalized inverse-Gaussian process. *Bayesian Analysis 8*, 553–568.

[5] Abramowitz, Milton and Stegun, Irene A (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing.* Dover Publications, Mineola, New York.

[6] Banjevic, Dragan and Ishwaran, Hemant and Zarepour, Mahmoud and others (2002). A recursive method for functionals of Poisson processes. *Bernoulli Society for Mathematical Statistics and Probability, Volume 8* 295–311.

[7] Bert Fristedt and Lawrence Gray (1996). A Modern Approach to Probability Theory. *Birkhauser Boston*.

[8] Bondesson, Lennart (1982). On simulation from infinitely divisible distributions. *Advances in Applied Probability* 855–869.

[9] Broderick, Tamara and Jordan, Michael I and Pitman, Jim and others (2012). Beta processes, stick-breaking and power laws. *Bayesian analysis. International Society for Bayesian Analysis. Vol 7* 439–476.

[10] Carlton, Matthew Aaron (1999) *Applications of the two-parameter Poisson-Dirichlet distribution.* PhD thesis, University of California, Los Angeles

[11] Caron, François and Fox, Emily B (2014). Bayesian nonparametric models of sparse and exchangeable random graphs. *arXiv preprint arXiv:1401.1137*.

[12] Caron, François and Teh, Yee Whye and Murphy, Thomas Brendan (2013). Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data. *CoRR*.

[13] Favaro, Stefano and Lijoi, Antonio and Prünster, Igor (2009). On the stick-breaking representation of normalized inverse Gaussian priors. *Biometrika, 99* 663774.

[14] Favaro, Stefano and Lijoi, Antonio and Mena, Ramsés H and Prünster, Igor (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71* 993–1008.

[15] Ferguson, Thomas S (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1* 209–230.

[16] Ferguson, Thomas S and Klass, Michael J (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics 1* 209–230.

[17] Ghahramani, Zoubin and Griffiths, Thomas L (2005). Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems* 475–482.

[18] Hjort, Nils Lid (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics, 18(3)* 1259–1294.

[19] Ishwaran, Hemant and James, Lancelot F (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 161–173.

[20] Ishwaran, Hemant and Zarepour, Mahmoud (2002). Exact and approximate sum representations for the Dirichlet process. *The Canadian Journal of Statistics 30* 269–283.

[21] Kallenberg, O (1983). Random Measures. *Third edition. Akademie-Verlag, Berlin.*

[22] Kim, Yongdai (1999). Nonparametric Bayesian estimators for counting processes. *Annals of Statistics, JSTOR* 562–588.

[23] Kingman, John Frank Charles (1992). Poisson processes. *Oxford University Press, Volume 3.*

[24] Kingman, John (1967). Completely random measures. *Pacific Journal of Mathematics 21* 59–78.

[25] Lijoi, Antonio and Mena, Ramsés H and Prünster, Igor (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association, 100* 1278–1291.

[26] Miller, Kurt (2011) *A Bayesian Nonparametric Latent Feature Models.* PhD thesis, University of California, Berkeley.

[27] Miller, Kurt and Jordan, Michael I and Griffiths, Thomas L (2009) *Nonparametric latent feature models for link prediction.* Advances in Neural Information Processing Systems, 1276–1284.

[28] Muliere, Pietro and Tardella, Luca (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *The Canadian Journal of Statistics 26*, 283–297.

[29] Nieto-Barajas, Luis E and Prünster, Igor (2009). A sensitivity analysis for Bayesian nonparametric density estimators. *Statistica Sinica, 19* 685–705.

[30] Paisley, John and Carin, Lawrence (2009). Nonparametric factor analysis with beta process priors. *Proceedings of the 26th Annual International Conference on Machine Learning* 777–784.

[31] Paisley, John W and Zaas, Aimee K and Woods, Christopher W and Ginsburg, Geoffrey S and Carin, Lawrence (2010). A stick-breaking construction of the beta process. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* 847–854.

[32] Pitman, Jim and Yor, Marc (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* 855–900.

[33] Resnick, Sidney I (1987). *Extreme values, regular variation, and point processes*. Springer-Verlag, New York.

[34] Robert L. Wolpert and Katja Ickstadt. (1998b). Simulation of Lévy random fiels. *Practical Nonparametric and Semiparametric Bayesian Statistics Lecture Notes in Statistics, 133* 227–242.

[35] Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica 4*, 639–650.

[36] Teh, Yee Whye and Jordan, Michael I (2010). Hierarchical Bayesian nonparametric models with applications. *Bayesian nonparametrics, Camb. Ser. Stat. Probab. Math* .

[37] Teh, Yee Whye (2010). Dirichlet process. *Encyclopedia of machine learning, Springer* 280–287.

[38] Thibaux, Romain Jean (2008). Nonparametric Bayesian models for machine learning. *Thesis*

[39] Thibaux, Romain and Jordan, Michael I (2007). Hierarchical beta processes and the Indian buffet process. *International conference on artificial intelligence and statistics* 564–571.

[40] Titsias, Michalis K (2008). The infinite gamma-Poisson feature model. *Advances in Neural Information Processing Systems*, 1513–1520.

[41] Wolpert, Robert L and Ickstadt, Katja (1998). Simulation of Lévy random fields. *Practical Nonparametric and Semiparametric Bayesian Statistics, Springer* 227–242.

[42] Zarepour, Mahmoud and Al Labadi, Luai (2012). On a rapid simulation of the Dirichlet process. *Statistics & Probability Letters 82.5*, 916–924.