On New Constructive Tools in Bayesian Nonparametric Inference

Luai Al Labadi

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

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

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

Abstract

The Bayesian nonparametric inference requires the construction of priors on infinite dimensional spaces such as the space of cumulative distribution functions and the space of cumulative hazard functions. Well-known priors on the space of cumulative distribution functions are the Dirichlet process, the two-parameter Poisson-Dirichlet process and the beta-Stacy process. On the other hand, the beta process is a popular prior on the space of cumulative hazard functions. This thesis is divided into three parts. In the first part, we tackle the problem of sampling from the above mentioned processes. Sampling from these processes plays a crucial role in many applications in Bayesian nonparametric inference. However, having exact samples from these processes is impossible. The existing algorithms are either slow or very complex and may be difficult to apply for many users. We derive new approximation techniques for simulating the above processes. These new approximations provide simple, yet efficient, procedures for simulating these important processes. We compare the efficiency of the new approximations to several other well-known approximations and demonstrate a significant improvement.

In the second part, we develop explicit expressions for calculating the Kolmogorov, Lévy and Cramér-von Mises distances between the Dirichlet process and its base measure. The derived expressions of each distance are used to select the concentration parameter of a Dirichlet process. We also propose a Bayesain goodness of fit test for simple and composite hypotheses for non-censored and censored observations.

Illustrative examples and simulation results are included.

Finally, we describe the relationship between the frequentist and Bayesian non-parametric statistics. We show that, when the concentration parameter is large, the two-parameter Poisson-Dirichlet process and its corresponding quantile process share many asymptotic pr operties with the frequentist empirical process and the frequentist quantile process. Some of these properties are the functional central limit theorem, the strong law of large numbers and the Glivenko-Cantelli theorem.

Acknowledgements

Special and sincere respect, gratitude and deep appreciation are expressed to my supervisors, Professor Raluca Balan and Professor Mahmoud Zarepour, for their invaluable supervision, skilful guidance, incredible support, continuous encouragement and being there for me every step of the way. This work could not be possible without both of you. Thank you deeply.

Appreciation is also extended to Professor Pierre-Jérôme Bergeron, Professor Michael Evans, Professor Rafal Kulik and Professor Mohamedou Ould Haye for kindly taking time to serve on my committee, and to provide thoughtful comments.

Financial support form the Natural Sciences and Engineering Research Council of Canada (NSERC), Ontario Graduate Scholarship (OGS), University of Ottawa Excellence Scholarship, University of Ottawa Admission Scholarship and Faculty of Graduate and Postdoctoral Studies Conference Travel Grants is acknowledged and greatly appreciated.

I also offer my regards to all of those who supported me in any respect throughout the process of writing this thesis.

Finally, I want to convey my heartfelt gratitude and appreciation to my beloved family. To my mother for her endless love, sacrifices, continuous blessings and prayers. To my beloved wife, Dima Allabbadi, goes the most heartfelt thanks. I will be forever grateful for your support, patience, sacrifice and understanding. To my daughters,

Bana and Ghena, who kept asking “When are you going to be done, Dad?”. You have been a constant source of inspiration, motivation and joy to me. To my parents-in-law for consideration, help and support. To my uncle, Mr. Husni Ayesh, and his wife for their encouragement and support.

Dedication

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

Contents

List of Figures

Chapter 1

Introduction

In Bayesian parametric inference, it is assumed that the data are generated from a probability measure $\{F_\theta, \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$. Typically, θ has low dimension. Assuming that θ is random, a prior distribution on θ is constructed. Then Bayes theorem combines both the likelihood and prior beliefs into the posterior distribution for θ given the data. Based on this posterior, we can obtain various inferences about the parameters such as posterior means, standard deviations and confidence intervals. However, restricting the inference to a specific parametric form may give unsatisfactory inferences. Therefore, to relax the parametric assumption, one may use a nonparametric approach. In the Bayesian context, this consists of constructing a prior on an infinite dimensional space such as the space of all possible distribution functions. In the literature, a nonparametric approach in Bayesian framework is often called a *Bayesian nonparametric inference* (Ghosh and Ramamoorthi, 2003).

Typically, there are two requirements for constructing a prior in Bayesian nonparametric inference: a mathematically tractable form for the posterior distribution given the observations, and a simulation algorithm for drawing samples from the prior and the posterior distribution. As for the first point, *conjugacy* is a desirable property. It is worth pointing out that, in a Bayesian nonparametric setup, the term "conju-

gacy" is used with slightly different meanings. Lijoi and Pruïster (2010) introduced two types of conjugacy: *parametric* conjugacy and *structural* conjugacy. The former occurs when the posterior process is of the same form of the prior process. The latter, namely structural conjugacy, identifies a model where the posterior process has the same structure as the prior process, in the sense that they both belong to the same general class of random probability measures. Clearly, structural conjugacy does not necessarily imply parametric conjugacy. On the other hand, parametric conjugacy implies structural conjugacy. It is more convenient to have parametric conjugacy because this implies that if we know how to sample from the prior, then the same type of algorithm can be used for the posterior.

A well-known prior, which plays a central role in Bayesian nonparametric inference, is the Dirichlet process introduced by Ferguson (1973). However, the Dirichlet process is not a suitable prior for all data. For example, the Dirichlet process is not a (parametric) conjugate prior for the survival time distribution when the sample contains right censored observations; see Theorem 3 of Ferguson and Phadia (1979), or Susarla and Ryzin (1976). Hence, an alternative nonparametric prior needs to be placed in the presence of right censored data. For instance, Walker and Muliere (1997) defined the beta-Stacy process as a prior on the space of cumulative distribution functions. Hjort (1990) constructed the beta process as a prior on the space of cumulative hazard functions. Since these two processes are conjugate priors given possibly right censored data, they are suitable for applications in survival analysis, reliability theory and other fields.

The outline of this thesis is as follows: In Chapter 2, we review some properties of Bayesian nonparametric priors, with special emphasis on the Dirichlet process, the two-parameter Poisson-Dirichlet process, the beta process and the beta-Stacy process. The notations and the results introduced in this chapter will be employed in the following chapters.

In Chapter 3, we recall the representation of Ferguson and Klass (1972) and the

representation of Wolpert and Ickstadt (1998). We compare these two representations from the computational point of view and clarify why the representation of Wolpert and Ickstadt is more appropriate for nonhomogeneous processes. As an example, we derive the Ferguson and Klass representation (1972) for the beta process.

In Chapter 4, we derive new approximations for the Dirichlet process, the two-parameter Poisson-Dirichlet process, the beta process and the beta-Stacy process. These new approximations provide simple, yet efficient, procedures for simulating these processes. We compare the efficiency of the new approximations to several other well-known approximations and demonstrate a significant improvement.

In Chapter 5, we use the sum representations of the Dirichlet process to derive explicit expressions to calculate the Kolmogorov, Lévy and Cramér-von Mises distances between the Dirichlet process and its base measure.

In Chapter 6, we use the distance formulas derived in Chapter 5 in two applications. In the first application, they are used to select a suitable concentration parameter for a Dirichlet process. In the second application, we propose a Bayesian goodness of fit test to examine a simple and a composite hypothesis for non-censored and censored observations.

In Chapter 7, we describe the relationship between the frequentist and Bayesian nonparametric statistics. We show that, when the concentration parameter is large, the two-parameter Poisson-Dirichlet process and its corresponding quantile process share many asymptotic properties with the frequentist empirical and quantile processes. Some of these properties are the functional central limit theorem, the strong law of large numbers and the Glivenko-Cantelli theorem.

Finally, in Chapter 8, we present a summary of the thesis and outline plans for future research problems.

Chapter 2

Bayesian Nonparametric Priors

The main objective of this chapter is to review some properties of Bayesian nonparametric priors, with special emphasis on the Dirichlet process, the two-parameter Poisson-Dirichlet process, the beta process and the beta-Stacy process. The notations and the results introduced in this chapter will be employed in the next chapters. Proofs of the results covered in this chapter are not included since they can be found in the original literature (exact references are included). However, we point out that the result obtained in Lemma 2.1.10 is new and it reveals an interesting property of the weights in existing representations of the Dirichlet process.

2.1 The Dirichlet Process

2.1.1 Definition and Basic Properties

The Dirichlet process, formally introduced in Ferguson (1973), is considered the cornerstone of Bayesian nonparametric inference. It is a prior law over the space of probability distribution functions, whose finite-dimensional marginals have a Dirichlet distribution. We begin by recalling the definition of the Dirichlet distribution.

Definition 2.1.1 *We say that the random vector* $(Z_1, \ldots, Z_k)$ *has the Dirichlet distribution with parameters* $(a_1, \ldots, a_k)$*, where* $a_i > 0$ *for all* i*, if it has the joint density*

$$f(z_1, \ldots, z_k) = \frac{\Gamma\left(\sum_{i=1}^k a_i\right)}{\prod_{i=1}^k \Gamma(a_i)} \prod_{i=1}^k z_i^{a_i - 1} \; I_{\mathbb{S}}(z_1, \ldots, z_k),$$

where $\mathbb{S}$ *is the simplex*

$$\mathbb{S} = \left\{ (z_1, \ldots, z_k) : z_i \geq 0, \sum_{i=1}^k z_i = 1 \right\}.$$

In the above definition,

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

is the gamma function. We denote by $D(a_1, \ldots, a_k)$ the Dirichlet distribution with parameters $a_1, \ldots, a_k$.

Following Ferguson (1973), the Dirichlet process is defined as follows:

Definition 2.1.2 *Let* $(\mathfrak{X}, \mathcal{A})$ *be an arbitrary measurable space and* H *be a probability measure on* $(\mathfrak{X}, \mathcal{A})$*. Let* $a > 0$ *be arbitrary. A random probability measure* $P = \{P(A)\}_{A \in \mathcal{A}}$ *is called a Dirichlet process on* $(\mathfrak{X}, \mathcal{A})$ *with parameters* a *and* H*, if for any finite measurable partition* $\{A_1, \ldots, A_k\}$ *of* $\mathfrak{X}$*, the joint distribution of the vector* $(P(A_1), \ldots P(A_k))$ *has the Dirichlet distribution with parameters* $(aH(A_1), \ldots, aH(A_k))$*, where* $k \geq 2$*. We assume that if* $H(A_j) = 0$*, then* $P(A_j) = 0$ *with probability one.*

Throughout this thesis, we use the same notation for the probability measure and its corresponding cumulative distribution function, i.e. $P(t) = P((-\infty, t])$ and $H(t) = H((-\infty, t])$ when $\mathfrak{X} = \mathbb{R}$. If P is a Dirichlet process with parameters a and H, we write $P \sim DP(a, H)$. For any $A \in \mathcal{A}$, $P(A)$ has a beta distribution with parameters $aH(A)$ and $a(1 - H(A))$. Thus,

$$E(P(A)) = H(A) \quad \text{and} \quad Var(P(A)) = \frac{H(A)(1 - H(A))}{1 + a}. \tag{2.1.1}$$

The probability measure H is called the base measure of P. Clearly, from (2.1.1), H plays the role of the center of the process, while a can be viewed as the concentration parameter. The larger a is, the more likely it is that the realization of P is close to H. Specifically, for any fixed set $A \in \mathcal{A}$ and $\epsilon > 0$, we have $P(A) \xrightarrow{p} H(A)$ as $a \to \infty$ since

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

In this thesis, "$\xrightarrow{d}$", "$\xrightarrow{p}$" and "$\xrightarrow{a.s.}$" denote convergence in distribution, convergence in probability and almost sure convergence, respectively.

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

Definition 2.1.3 *Let P be a random probability measure on $(\mathfrak{X}, \mathcal{A})$. We say that $X_1, \ldots, X_m$ is a sample of size m from P if for any $n = 1, 2, \ldots$ and sets $B_1, \ldots B_n$, $C_1, \ldots C_m \in \mathcal{A}$;*

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

almost surely. See Ferguson (1973).

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

The next theorem describes the posterior distribution given the data for the Dirichlet process. It shows that the Dirichlet process has the (parametric) conjugacy property. That is, the posterior distribution given the data is again a Dirichlet process. For the proof, see Ferguson (1973, Theorem 1). In the theorem and through out the thesis, we use a "$*$" as a superscript to denote posterior quantities.

Theorem 2.1.4 *If $X_1, \ldots, X_m$ is a sample from $P \sim DP(a, H)$, then the posterior distribution of P given $X_1, \ldots, X_m$ coincides with the distribution of the Dirichlet process with parameters a^* and H^*, where $a^* = a + m$ and*

$$H^* = \frac{a}{a+m} H + \frac{m}{a+m} \frac{\sum_{i=1}^{m} \delta_{X_i}}{m}. \tag{2.1.2}$$

Here and throughout this thesis, δ_X denotes the Dirac measure at X, i.e. $\delta_X(A) = 1$ if $X \in A$ and 0 otherwise.

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

2.1.2 Exact and Approximate Sum Representations for the Dirichlet Process

Ferguson (1973) proposed a series representation as an alternative definition for the Dirichlet process. This representation is based on the earlier work of Ferguson and Klass (1972), in which they provided a sum representation for processes with independent increments (and no Gaussian part) based on the arrival times of a homogeneous Poisson process. The representation is described in the following theorem. For the proof see Ferguson and Klass (1972), Ferguson (1973), and Banjevic, Ishwaran and Zarepour (2002). Also see Section 1 of Chapter 3 for a general review of the Ferguson and Klass representation (1972).

Theorem 2.1.5 *Let $(E_k)_{k\geq 1}$ be a sequence of independent and identically distributed (i.i.d.) random variables with an exponential distribution of mean 1. Define*

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

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

$$P^{Ferg.}(\cdot) = \sum_{i=1}^{\infty} \frac{L^{-1}(\Gamma_i)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)} \delta_{\theta_i}(\cdot), \tag{2.1.4}$$

where

$$L(x) = a \int_x^{\infty} t^{-1} e^{-t} dt, \quad x > 0, \tag{2.1.5}$$

and

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

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

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

$$L(dx) = ax^{-1}e^{-x}dx$$

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

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

From Theorem 2.1.5, it follows clearly that a realization of the Dirichlet process must necessarily be a discrete probability measure, even when the base measure is absolutely continuous. This fact was noted by Ferguson (1973), and Blackwell and McQueen (1973). It is worth mentioning that this discreteness property is no more troublesome than the discreteness of the empirical process. In spite of the discreteness property, the support of the Dirichlet process is very large. For more details, see Proposition 3 of Ferguson (1973).

Remark 2.1.6 *The series* $\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)\delta_{\theta_i}(\cdot)$ *is the Ferguson and Klass representation of the gamma process. Thus, the Dirichlet process can be viewed as a normalized gamma process.*

Remark 2.1.7 *Brix (1999) introduced the so-called G-measure which is an interesting extension of the gamma process with a more generalized Lévy measure*

$$M_{a,\kappa,\theta}(x) = \frac{a}{\Gamma(1-\kappa)} \int_x^\infty e^{-\theta t} t^{-\delta-1}\, dt,$$

where $(a, \kappa, \theta) \in (0,\infty) \times (0,1] \times [0,\infty)$ *or* $(0,\infty) \times (-\infty, 0] \times (0,\infty)$. *The case when* $\kappa = 0$ *and* $\theta = 1$ *coincides with the Lévy measure of the gamma process given in (2.1.5). Also refer to Ishwaran and Zarepour (2009) and Argiento, Guglielmi and Pievatolo (2010) for more discussion regarding the G-measures.*

Working with (2.1.4) is difficult in practice because there is no closed form for the inverse of the Lévy measure (2.1.5). Moreover, to determine the random weights in (2.1.4) an infinite sum must be computed. Wolpert and Ickstadt (1998) described an approximate evaluation of $L(x)$ and used it for the simulation of the gamma process. Brix (1999) used a similar approach to sample from G-measures. To use this technique to sample from the Dirichlet process, a further normalization of the gamma process (a nonnegative finite process) is necessary in order to convert it to a probability measure. Bondesson (1982) provided a series representation of the gamma process. Bondesson's representation for the Dirichlet process with parameters a and H has the following form (Rosiński, 2001; Ishwaran and Zarepour, 2002):

Theorem 2.1.8 *Let* Γ_i *and* θ_i *be as defined in Theorem 2.1.5. Let*

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

where $(E_i')_{i\geq 1}$ *is a sequence of i.i.d. random variables with an exponential distribution of mean 1, independent of both* $(\Gamma)_{i\geq 1}$ *and* $(\theta)_{i\geq 1}$. *Then* $P^{Bond.}$ *is a Dirichlet process with parameters a and H.*

As in the case of the representation (2.1.4), it is impossible to sample directly the Dirichlet process using Bondesson's construction, due to the presence of an infinite sum. As an alternative, one can approximate the Dirichlet process using the

truncation

$$P_n^{\text{Bond.}}(\cdot) = \sum_{i=1}^{n} \frac{e^{-\Gamma_i/a} E_i'}{\sum_{i=1}^{n} e^{-\Gamma_i/a} E_i'} \delta_{\theta_i}(\cdot). \tag{2.1.6}$$

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

$$n = \inf \left\{ j : \frac{e^{-\Gamma_j/a} E_j'}{\sum_{i=1}^{j} e^{-\Gamma_i/a} E_i'} < \epsilon \right\}. \tag{2.1.7}$$

The random stopping rule (2.1.7) is similar to the one suggested by Muliere and Tardella (1998). It is important to note that the weights in Bondesson's representation are not monotonically decreasing, in an almost sure sense (they are only eventually stochastically decreasing). More precisely, it can be proved that, for any $i \geq 1$,

$$\Pr \left\{ \frac{e^{-\Gamma_{i+1}/a} E_{i+1}'}{\sum_{i=1}^{\infty} e^{-\Gamma_i/a} E_i'} < \frac{e^{-\Gamma_i/a} E_i'}{\sum_{i=1}^{\infty} e^{-\Gamma_i/a} E_i'} \right\} < 1. \tag{2.1.8}$$

To see this, note that the probability on the left-hand side of (2.1.8) is equal to

$$\Pr \left\{ e^{-\Gamma_{i+1}/a} E_{i+1}' < e^{-\Gamma_i/a} E_i' \right\} \;=\; \Pr \left\{ e^{-E_{i+1}/a} < \frac{E_i'}{E_{i+1}'} \right\} \tag{2.1.9}$$

Note that $e^{-E_{i+1}/a} \stackrel{d}{=} beta(a,1)$ and $E_i'/E_{i+1}' \stackrel{d}{=} F(2,2)$, where $F(2,2)$ is the F distribution with parameters 2, 2 and "$\stackrel{d}{=}$" denotes equality in the distribution. Hence, the probability in (2.1.9) is equal to

$$\int_0^1 \int_x^\infty a x^{a-1} \times \frac{1}{(1+y)^2} dy dx = \int_0^1 \frac{a x^{a-1}}{1+x} dx \;=\; \sum_{k=0}^{\infty} (-1)^k \frac{a}{k+a}.$$

For $a = 1$ and $a = 10$, the probability on left-hand side of (2.1.8) is equal to 0.6931 and 0.5249, respectively. It is worth noting that the probability in (2.1.8) is independent of i. Thus, it typically happens in (2.1.7) that

$$\frac{e^{-\Gamma_k/a} E_k'}{\sum_{i=1}^{k} e^{-\Gamma_i/a} E_i'} > \epsilon,$$

for some $k > n$. Therefore, Bondesson's representation (2.1.6) is inefficient for simulation purposes.

A radically different (constructive) definition of the Dirichlet process is given by Sethuraman (1994) using a "stick-breaking" approach. This representation is stated in the next theorem (Sethuraman, 1994).

Theorem 2.1.9 *Let $(\beta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with a $Beta(1,a)$ distribution. Define*

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

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

$$P^{Seth.}(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(\cdot). \tag{2.1.10}$$

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

Notice that, unlike the constructions of Wolpert and Ickstadt (1998) and Bondesson (1982), the stick-breaking construction does not involve a normalization. Sethurman's stick-breaking representation can be used to approximately simulate the Dirichlet process using a truncation argument. By truncating the higher order terms in the sum we can approximate the Sethuraman stick-breaking representation by

$$P_n^{\text{Seth.}}(\cdot) = \sum_{k=1}^{n} p_i \delta_{\theta_i}(\cdot), \tag{2.1.11}$$

where $(\beta_i)_{i\geq 1}$, $(p_i)_{i\geq 1}$, and $(\theta_i)_{i\geq 1}$ are as defined in (2.1.10) with $\beta_n = 1$ (hence β_n does not have a beta distribution). The assumption that $\beta_n = 1$ is necessary to make the weights add to 1, almost surely (Ishwaran and James, 2001). A random stopping rule for choosing $n = n(\epsilon)$ was proposed by Muliere and Tardella (1998) where, for $\epsilon \in (0,1)$,

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

The following lemma shows that the weights $(p_i)_{i\geq 1}$ in the stick-breaking representation are not strictly decreasing, almost surely (they are only stochastically decreasing). This makes the truncated stick-breaking representation inefficient for simulations.

Lemma 2.1.10 *For $i \geq 1$, we have* $\Pr\{p_{i+1} < p_i\} = \sum_{k=0}^{\infty}(-1)^k a/(k+a)$.

Proof: Since $p_i \stackrel{d}{=} \exp(-\Gamma_{i-1}/a) - \exp(-\Gamma_i/a)$, where $\Gamma_0 = 0$ (Ishwaran and Zarepour, 2002),

$$\begin{aligned}
\Pr\{p_{i+1} < p_i\} &= \Pr\left\{e^{-\Gamma_i/a} - e^{-\Gamma_{i+1}/a} < e^{-\Gamma_{i-1}/a} - e^{-\Gamma_i/a}\right\} \\
&= \Pr\left\{2 < e^{E_i/a} + e^{-E_{i+1}/a}\right\} \\
&= \int_0^1 \int_{2-x}^{\infty} ax^{a-1} \times ay^{-a-1} dy dx \\
&= \int_0^1 \frac{ax^{a-1}}{(2-x)^a} dx.
\end{aligned}$$

Now, a substitution of $u = 1 - x$ in the last integral gives

$$\int_0^1 a(1+u)^{-a}(1-u)^{a-1} dx. \tag{2.1.12}$$

A Taylor expansion of $(1+u)^{-a}$ concludes that the expression displayed in (2.1.12) is equal to

$$\begin{aligned}
\sum_{k=0}^{\infty} a\binom{-a}{k}\int_0^1 u^k(1-u)^{a-1}du &= \sum_{k=0}^{\infty} a(-1)^k\binom{a+k-1}{k}\int_0^1 u^k(1-u)^{a-1}du \\
&= \sum_{k=0}^{\infty} a(-1)^k \frac{(a+k-1)\cdots a}{k!}\frac{\Gamma(k+1)\Gamma(a)}{\Gamma(a+k+1)} \\
&= \sum_{k=0}^{\infty}(-1)^k\frac{a}{k+a}.
\end{aligned}$$

∎

It follows from Lemma 2.1.10 that the probability $\Pr\{p_{i+1} < p_i\}$ coincides with the probability given in the left-hand side of (2.1.8), which is surprising. Furthermore, the probability $\Pr\{p_{i+1} < p_i\}$ does not depend on i.

Finally, an alternative and very useful approximation of the Dirichlet process can be obtained from certain finite mixture models. In particular, Ishwaran and Zarepour (2002) proved the next result.

Theorem 2.1.11 *For any $n \geq 1$, consider the finite-dimensional Dirichlet prior*

$$P_n^{F.D.}(\cdot) = \sum_{i=1}^{n} p_{i,n}\delta_{\theta_i}(\cdot), \tag{2.1.13}$$

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

$$P_n^{F.D.}(g) := \int g(x)P_n^{F.D.}(dx) \xrightarrow{d} P(g) := \int g(x)P(dx)$$

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

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

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

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

In Chapter 4, we propose a sum representation which involves monotonically decreasing weights. This new representation is a simple, yet highly accurate, almost sure approximation of the Dirichlet process.

2.2 The Two-Parameter Poisson-Dirichlet Process

An important extension of the Dirichlet process is the two-parameter Poisson-Dirichlet process developed by Pitman and Yor (1997). We start by recalling the Pitman and Yor (1997) stick-breaking definition of the two-parameter Poisson-Dirichlet process on an arbitrary measurable space $(\mathfrak{X}, \mathcal{A})$.

Definition 2.2.1 *For $0 \leq \theta < 1$, $a > -\theta$, let $(\beta_i)_{i\geq 1}$ be a sequence of independent random variables with a $Beta(1-\theta, a+i\theta)$ distribution. Define*

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

Let $p_1 \geq p_2 \geq \ldots$ be the ranked values of $(\widetilde{p_i})_{i\geq 1}$. Moreover, let $(\theta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with common distribution H, independent of $(\beta_i)_{i\geq 1}$. Then the random probability measure

$$P_{H,\theta,a}(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(\cdot) \tag{2.2.1}$$

is called a two-parameter Poisson-Dirichlet process *on $(\mathfrak{X}, \mathcal{A})$ with parameters θ, a and H.*

It is worth mentioning that Ishwaran and James (2001) referred to the process $\widetilde{P}_{H,\theta,a}(\cdot) = \sum_{i=1}^{\infty} \widetilde{p_i}\delta_{\theta_i}(\cdot)$ as the Pitman-Yor process, where $(\widetilde{p_i})_{i\geq 1}$ and $(\theta_i)_{i\geq 1}$ are as defined in Definition 2.2.1. The two-parameter Poisson-Dirichlet process with parameters θ, a and H is denoted by $PDP(H;\theta,a)$, and we write $P_{H,\theta,a} \sim PDP(H;\theta,a)$. In the literature, the probability measure H is called the *base measure* of $P_{H,\theta,a}$, while the parameters θ and a are called the *discount parameter* and the *concentration parameter*, respectively (Buntine and Hutter, 2010; Teh, 2006). The representation (2.2.1) clearly shows that any realization of the two-parameter Poisson-Dirichlet process must be a discrete probability measure. Note that, the special case $PDP(H;0,a)$ represents the Dirichlet process, i.e. $PDP(H;0,a) = DP(a,H)$. The law of the weights $(p_1, p_2, \ldots)$ is called the two-parameter Poisson-Dirichlet distribution, denoted by $PD(\theta, a)$. The two-parameter Poisson-Dirichlet distribution has many applications in different fields such as population genetics, ecology, statistical physics and number theory. See Feng (2010) for more details. On the other hand, the two-parameter Poisson-Dirichlet process has been recently used in applications in Bayesian nonparametric statistics such as computer science (Teh, 2006), species sampling (Jang, Lee and Lee, 2010; Navar-

rete, Quintana and Müller, 2008) and genomics (Favaro, Lijoi, Mena and Prünster, 2009).

The calculations of the moments for the two-parameter Poisson-Dirichlet process are carried out in Carleton (1999). Let A be a measurable subset of $\mathfrak{X}$. Then

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

Furthermore, for any two disjoint sets A_i and $A_j \in \mathcal{A}$,

$$E(P_{H,\theta,a}(A_i)P_{H,\theta,a}(A_j)) = H(A_i)H(A_j)\frac{\theta+a}{1+a}. \tag{2.2.3}$$

It follows from (2.2.2) that the base measure H plays the role of the center of the process, while both θ and a control the variability of $P_{H,\theta,a}$ around H. Observe that, for any fixed set $A \in \mathcal{A}$ and $\epsilon > 0$, we have

$$\Pr\{|P_{H,\theta,a}(A) - H(A)| > \epsilon\} \leq H(A)(1 - H(A))\frac{1-\theta}{(1+a)\epsilon^2}. \tag{2.2.4}$$

Thus, $P_{H,\theta,a}(A) \xrightarrow{p} H(A)$ as $a \to \infty$ (for a fixed θ) or as $\theta \to 1$, $\theta < 1$ (for fixed a).

As given in Section 2.1, Ferguson (1973) defined the Dirichlet process by prescribing the finite-dimensional distribution of this process corresponding to an arbitrary measurable partition of the measure space. Then a series representation of the Dirichlet process was given. On the other hand, the two-parameter Poisson-Dirichlet process is only defined through the stick-breaking representation (2.2.1). The finite-dimensional distributions of the two-parameter Poisson-Dirichlet process were studied by Carleton (1999), which he called the *RS distributions*. Unless $\theta \in \{0, 1/2\}$, the RS distributions cannot be written in a tractable form. For the case $\theta = 0$, the RS distribution is the Dirichlet distribution given by Definition 2.1.1. On the other hand, when $\theta = 1/2$, the RS distribution was first obtained by Carleton (1999, Theorem 3.1). See also James, Lijoi, and Prünster (2005, Proposition 4.7) for a different technique of deriving the RS distribution in this case. James, Lijoi, and Prünster (2005) pointed out that the two-parameter Poisson-Dirichlet process when $\theta = 1/2$ is viewed

as a mixture of inverse-Gaussian normalized process. See Lijoi, Mena and Prünster (2005) for details.

Definition 2.2.2 *We say that the random vector $(Z_1, \ldots, Z_k)$ has the RS distribution with parameters $(\gamma_1, \ldots, \gamma_k; 1/2, a)$, where $\gamma_i > 0$ for all i and $a > -1/2$, if it has the joint probability density function*

$$f(z_1, \ldots, z_k) = \frac{\gamma_1 \cdots \gamma_k \Gamma(a + \frac{k}{2})}{\pi^{\frac{k-1}{2}} \Gamma(a + \frac{1}{2})} \frac{z_1^{-3/2} \ldots z_k^{-3/2}}{\left(\frac{\gamma_1^2}{z_1} + \cdots + \frac{\gamma_k^2}{z_k}\right)^{a+\frac{k}{2}}} I_{\mathbb{S}}(z_1, \ldots, z_k),$$

where $\mathbb{S}$ is the simplex

$$\mathbb{S} = \left\{ (z_1, \ldots, z_k) : z_i \geq 0, \sum_{i=1}^{k} z_i = 1 \right\}.$$

In general, the RS distribution with parameters $(\gamma_1, \ldots, \gamma_k; \theta, a)$ is denoted by RS $(\gamma_1, \ldots, \gamma_k; \theta, a)$.

The next theorem describes the posterior distribution given the data for the two-parameter Poisson-Dirichlet process. It shows that the two-parameter Poisson-Dirichlet process does not have, in general, the conjugacy property. For a detailed proof of the theorem, see Carleton (1999, page 48).

Theorem 2.2.3 *Let $X_1, \ldots, X_m$ be a sample from $P_{H,\theta,a}$. Let K be the number of distinct X_i's, X_j' be the jth distinct X_i, and m_j be the number of X_i equal to X_j'. Then*

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

where $P_{H,\theta,a+K\theta} \sim PDP(H; \theta, a + K\theta)$, $(W_1, \ldots, W_{K+1}) \sim D(n_1 - \theta, \ldots, n_K - \theta, a + K\theta)$, and $P_{H,\theta,a+K\theta}$ and $(W_1, \ldots, W_{K+1})$ are conditionally independent given $X_1, \ldots, X_m$.

2.3 Lévy Processes

Definition 2.3.1 *(Applebaum, 2004) A cadlag* [1] *(i.e., right-continuous with left limits a.s.) stochastic process* $\{Z(t)\}_{t\geq 0}$ *defined on a probability space* $(\Omega, \mathscr{F}, \mathbb{P})$ *is called a Lévy process if it satisfies the following conditions:*

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

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

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

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

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

or equivalently $Z(t) \xrightarrow{p} Z(s)$ *as* $t \to s$.

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

The Lévy-Khintchine representation theorem states that a Lévy process is the sum of a deterministic function, a Brownian motion, and a pure jump part (Applebaum, 2004). Its characteristic function is given by

$$E\left(e^{iuZ(t)}\right) = \exp\left\{i\gamma_t u - \frac{\sigma_t^2}{2}u^2 + \int_{\mathbb{R}-\{0\}} \left(e^{ius} - 1 - ius I_{\{|s|\leq 1\}}\right) L_t(ds)\right\}, \; u \in \mathbb{R},$$

[1] Some authors do not impose the cadlag property in the definition of a Lévy process. In fact, every Lévy process (defined without the cadlag property) has a unique modification that is cadlag (Sato, 1999, Theorem 30). Therefore, the cadlag property can be assumed without loss of generality (Cont and Tankov, 2004, Definition 3.1)

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

Lévy processes have been used in Bayesian nonparametric inference by Doksum (1974), Ferguson (1973, 1974), Ferugson and Phadia (1979), Hjort (1990), and Walker and Muliere (1997), with the following additional properties:

(5) $Z(t)$ is nonnegative: $Z(t) \geq 0$ a.s.

(6) $Z(t)$ is nondecreasing a.s.

Any Lévy process which satisfies conditions (1)-(6) has no diffusion component (Cont and Tankov, 2004, Proposition 3.10; Lee, 2007, page 658). For such a process $\{Z(t)\}_{t\geq 0}$ there exist at most countably many fixed points of discontinuity. A point t_i such that $\Pr\{Z(\{t_i\}) > 0\} > 0$ corresponds to an atom of $Z(t)$ whose location is constant, and is called a *fixed point of discontinuity* (Daley and Vere-Jones 2008, Definition 9.3.I, page 39). By contrast, if $Z(\{t_i\}) > 0$ but $\Pr\{Z(\{t_i\}) > 0\} = 0$ then such an atom is termed a *random atom*. Let $M = \{t_1, t_2, \ldots\}$ be the set of fixed points of discontinuity of $\{Z(t)\}_{t\geq 0}$. Assume that the corresponding jumps $S_1, S_2, \ldots$ are independent (nonnegative) random variables with corresponding densities $f_{t_1}, f_{t_2}, \ldots$ (with respect to some convenient measures). The difference

$$Z_c(t) = Z(t) - \sum_{i\geq 1} S_i I\{t_i \leq t\},$$

is a nondecreasing process with independent increments, without fixed points of continuity. Therefore, $\{Z_c(t)\}_{t\geq 0}$ has a Lévy representation with log-Laplace transform given by: (Ferguson 1974)

$$\log E e^{uZ_c(t)} = ub(t) + \int_0^\infty (e^{us} - 1) L_t(ds), \quad t \geq 0, u \in \mathbb{R},$$

where b is a nondecreasing continuous function with $\lim_{t\to-\infty} b(t) = 0$, and $(L_t)_{t\geq 0}$ is a family of Lévy measures on the Borel subsets of $(0, \infty)$ satisfying the following regularity conditions:

(i) $L_0 = 0$.

(ii) For every Borel set B, $L_t(B)$ is nondecreasing and continuous in t. Note that, L_t is absolutely continuous with respect to L_∞. We denote $n_t(s) = L_t(ds)/L_\infty(ds)$, for $s > 0$.

(iii) For all $t \geq 0$, $L_t(\cdot)$ is a measure on the Borel sets of $(0, \infty)$.

(iv) $\int_0^\infty s/(1+s)L_\infty(ds) < \infty$, or equivalently $\int_0^\infty (1 \wedge s)L_\infty(ds) < \infty$, where the notation $1 \wedge s$ denotes the minimum of 1 and s. (Since L_∞ is a Lévy measure, $\int_1^\infty L_\infty(ds) < \infty$ and it suffices to verify that $\int_0^1 sL_\infty(ds) < \infty$.)

In summary, a Lévy process $\{Z(t)\}_{t\geq 0}$ with conditions (1)-(6) is specified by the four quantities: $M = \{t_1, t_2, \ldots\}$, $\{f_{t_1}, f_{t_2}, \ldots\}$, b, and L_t. The function b represents the nonrandom component of the process $\{Z(t)\}_{t\geq 0}$. In what follows, we assume that b is identically zero. Thus, $\{Z(t)\}_{t\geq 0}$ increases only in jumps a.s., so that it is discrete with probability one (Ferguson, 1974). In particular, $\{Z(t)\}_{t\geq 0}$ can be represented as a superposition of two components:

$$Z(t) = Z_c(t) + Z_f(t),$$

where

$$Z_c(t) = \sum_{i=1}^{\infty} J_i I\{\theta_i \leq t\} \tag{2.3.1}$$

with random positive jumps (weights) J_i and random locations (atoms) θ_i. On the other hand,

$$Z_f(t) = \sum_{i\geq 1} S_i I\{t_i \leq t\}. \tag{2.3.2}$$

where $S_1, S_2 \ldots$ are independent and nonnegative random variables with corresponding density functions $f_{t_1}, f_{t_2} \ldots$. Here the process $\{Z_f(t)\}_{t\geq 0}$ represents the part due to the existence of fixed points of discontinuity. Thus, drawing samples from $\{Z_f(t)\}_{t\geq 0}$ is straightforward. To characterize the process $\{Z(t)\}_{t\geq 0}$ entirely, it remains to specify the pure jump process $\{Z_c(t)\}_{t\geq 0}$. This issue will be considered in Chapter 3.

2.4 Neutral to the Right Processes

Neutral to the right processes were introduced in Doksum in (1974) and constitute a large class in Bayesian nonparametric inference. Well-known examples of neutral to the right process are the Dirichlet process (Ferguson, 1973), the beta process (Hjort, 1990), and the beta-Stacy process (Walker and Muliere, 1997). Historically, the concept of "neutrality" goes back to Conner and Mossiman (1969) who studied various concepts of independence for random proportions. According to them, a random vector $(p_1 \ldots, p_{m+1})$ of proportions with $\sum_{i=1}^{m+1} p_i = 1$ is a neutral to the right process if p_1 is independent of $p_2/(1-p_1)$, (p_1, p_2) is independent of $p_3/(1-p_1-p_2)$, and so on. Doksum (1974) formally developed this concept into a class of nonparametric priors, obtained its basic properties and showed that if the prior is a neutral to the right process, then so is the posterior distribution (even if some of the observations are right censored). It is worth recalling here that the Dirichlet process has the (parametric) conjugacy property only in the case when all observations are uncensored. In the presence of right censored observations, the posterior distribution of the Dirichlet process is only a neutral to the right process but is no longer a Dirichlet process. Indeed, the posterior distribution in this case is beta-Stacy (Walker and Muliere, 1997).

In general, the neutral to the right processes are defined over all real numbers $(-\infty, \infty) = \mathbb{R}$. In this thesis, unless otherwise is stated, they are defined over $[0, \infty) = \mathbb{R}^+$.

Definition 2.4.1 *A random distribution function* $\{F(t)\}_{t\in\mathbb{R}^+}$ *is called a neutral to the right process if for every* $k > 1$ *and* $0 < t_1 < t_2 < \ldots < t_k$, *there exists nonnegative independent random variables* $V_1, V_2, \ldots, V_k$ *such that*

$$(F(t_1), F(t_2), \ldots, F(t_k)) \stackrel{d}{=} \left(V_1, 1-(1-V_1)(1-V_2), \ldots, 1-\prod_{i=1}^{k}(1-V_i) \right).$$

Note that, from the previous equations

$$F(t_i) \stackrel{d}{=} 1 - \prod_{j=1}^{i}(1 - V_j), \quad i = 1, \ldots, k,$$

we get

$$F(t_i) - F(t_{i-1}) \stackrel{d}{=} V_i \prod_{j=1}^{i-1}(1 - V_j).$$

Therefore,

$$\frac{F(t_i) - F(t_{i-1})}{1 - F(t_{i-1})} \stackrel{d}{=} V_i.$$

Essentially, Definition 2.4.1 says that $\{F(t)\}_{t\geq 0}$ is neutral to the right if the normalized increments

$$F(t_1), \left(F(t_2) - F(t_1)\right) / \left(1 - F(t_1)\right), \ldots, \left(F(t_k) - F(t_{k-1})\right) / \left(1 - F(t_{k-1})\right) \quad (2.4.1)$$

are independent for all $0 < t_1 < t_2 < \ldots < t_k$. Note that, if we let $p_i = F(t_i) - F(t_{i-1})$ for any $i = 1, \ldots, k$, and $F(t_0) = 0$, then (2.4.1) coincides with the Conner and Mossiman (1969) definition of a neutral to the right random proportion.

The mathematical tractability of neutral to the right processes stems from their connection with Lévy processes. The following theorem establishes this connection. For the proof, see Doksum (1994, Theorem 3.1).

Theorem 2.4.2 *$\{F(t)\}_{t\geq 0}$ is a neutral to the right process if and only if the process*

$$Z(t) = -\log(1 - F(t))$$

has independent increments, is nondecreasing a.s., right-continuous a.s., $Z(0) = 0$ a.s., and $\lim_{t\to\infty} Z(t) = \infty$ a.s.

Observe that, since the Lévy process $\{Z(t)\}_{t\geq 0}$ is discrete with probability one, the neutral to the right process $F(t) = 1 - \exp(-Z(t))$ must be discrete with probability one.

The main property of neutral to the right processes which makes them attractive in applications is their conjugacy property.

Theorem 2.4.3 *(Doksum, 1974) Let $F = \{F(t)\}_{t\in\mathbb{R}^+}$ be neutral to the right process. If $X_1, \ldots, X_m$ is a sample from F, then the posterior distribution of F given $X_1, \ldots, X_m$ is also a neutral to the right process.*

Ferguson and Phadia (1974) extended this theorem and showed that the neutral to the right priors are conjugate for right censored observations. See Doksum (1974, Theorem 4.2) and Ferguson and Fadia (1974, Theorem 3) for the description of the posterior distributions. See also Dey, Erickson, and Ramamoorthy (2003) for other properties of neutral to the right processes.

2.5 The Beta Process

The Dirichlet process, the two-parameters Poisson-Dirichlet process and the beta-Stacy process (next section) are priors on the space of cumulative distribution functions. However, in many applications in survival analysis, the data is known in terms of the hazard rate or the cumulative hazard rate. Hjort (1990) introduced the beta process as a prior on the space of cumulative hazard functions. The beta process possesses two important features: it is a generalization of the Dirichlet process and it is a conjugate prior, given possibly right censored data. In this chapter, we consider only the continuous-time version of the beta process. For the discrete-time version, see Hjort (1990) and Sinha (1997).

Let T be a random variable with cumulative distribution function $F(t)$ on $\mathbb{R}^+$ and $F(0) = 0$. Let A be the cumulative hazard function of F, defined as: (see also Appendix A)

$$A(t) = \int_0^t \frac{dF(s)}{F[s,\infty)}. \tag{2.5.1}$$

As in Hjort (1990), the symbol A is used in different ways. $A[a,b)$, $A(t)$, and $A\{t\}$ denote respectively the increment of the function A in the interval $[a,b)$, the value of the function A at a point t, and the increment of the function A at t. Note

that

$$A[s, s+ds) = dA(s) = \frac{dF(s)}{F[s,\infty)} = \Pr\left\{T \in [s, s+ds) | T \geq s\right\}$$

and

$$A[a,b) = \int_{[a,b)} \frac{dF(s)}{F(s,\infty)}, \quad 0 \leq a \leq b < \infty.$$

F can be recovered from A by using the product integral formula (A.0.12), given in Appendix A, such that

$$F(t) = 1 - \prod_{[0,t]} \left(1 - dA(s)\right), t \geq 0. \tag{2.5.2}$$

The quantity $\prod_{[0,t]} (1 - dA(s)) = \exp(-A[0,t])$ if and only if A is continuous. Thus $-\log(1-F) = A$ only if A is continuous (see Appendix A for the details).

Hjort (1990) constructed prior distributions for A as follows. Let $\mathscr{B}$ be the set of all nondecreasing, right continuous functions B on $\mathbb{R}^+$ having $B(0) = 0$. Let $\mathscr{F}$ be the set of all cumulative distribution functions F on $\mathbb{R}^+$ having $F(0) = 0$. Define the space of cumulative hazard rates as:

$$\mathscr{A} = \left\{\text{those } A \text{ in } \mathscr{B} \text{ for which (2.5.2) leads to an } F \text{ in } \mathscr{F}\right\}.$$

The goal is to place a probability measure on $(\mathscr{A}, \Sigma_{\mathscr{A}})$, where $\Sigma_{\mathscr{A}}$ is the $\sigma-$algebra generated by the Borel cylinder sets. For this prior, Hjort (1990) considered Lévy processes, which are nonnegative, nondecreasing processes on $\mathbb{R}^+$ that start at zero and have independent increments. However, not every Lévy process can be used as a prior for A. For example, the gamma process does not have paths that almost surely produce proper cumulative distribution functions in (2.5.2). This is because its sample paths could have jumps whose size exceed 1, causing the cumulative distribution function in (2.5.2) to exceed 1. Recognizing this limitation, Hjort (1990) constructed a rich class of priors, called the beta processes, whose sample paths lie in $\mathscr{A}$ almost surely.

Definition 2.5.1 *Let A_0 be a cumulative hazard function with a finite (countable) number of jumps taking place at $M = \{t_1, t_2 \ldots\}$ and $c(\cdot)$ be a piecewise continuous, nonnegative function on $\mathbb{R}^+$. A process A is called a beta process with parameters $c(\cdot)$ and $A_0(\cdot)$, denoted by*

$$A \sim BP\left(c(\cdot), A_0(\cdot)\right),$$

if it is a Lévy process with the set $M = \{t_1, t_2 \ldots\}$ of fixed points of discontinuity and jump sizes $\{S_1, S_2, \ldots\}$, i.e.

$$A(t) = A_c(t) + \sum_{i \geq 1} S_i I\left\{t_i \leq t\right\},$$

for which:

- $E\left(e^{-uA(t)}\right) = \left[\prod_{i:t_i \leq t} E\left(e^{-uS_i}\right)\right] \exp\left\{\int_0^1 \left(e^{-us} - 1\right) L_t(ds)\right\}.$

- $L_t(ds) = \left[\int_0^t c(z) s^{-1} (1-s)^{c(z)-1} dA_{0,c}(z)\right] ds, \quad \text{for } t \geq 0, \quad 0 < s < 1, \tag{2.5.3}$

where

$$A_{0,c}(t) = A_0(t) - \sum_{t_i \leq t} A_0\left\{t_i\right\}. \tag{2.5.4}$$

- $S_i = A\left\{t_i\right\} \sim \text{Beta}\left(\text{c(t}_\text{i})\text{A}_0\left\{\text{t}_\text{i}\right\}, \text{c(t}_\text{i})\left(1 - \text{A}_0\left\{\text{t}_\text{i}\right\}\right)\right)$

A basic property of the beta process is: (see Hjort, 1990)

$$E(A(t)) = A_0(t), \quad \text{Var}(A(t)) = \int_0^t \frac{dA_0(s)}{c(s)+1}. \tag{2.5.5}$$

Thus, A_0 can be viewed as the prior guess about the cumulative hazard and can be chosen from a parametric model such as the exponential model (see Example A.0.1 in the Appendix). On the other hand, $c(\cdot)$ can be interpreted as concentration parameter or the strength of belief in the prior guess. It is interesting to observe the similarity

between the parameters of the beta process and those of the Dirichlet process, which also has two parameters: the base measure and the concentration parameter.

Let $X_1, \ldots, X_m$ be i.i.d random variables given A, i.e. with common cumulative distribution function $F = F_A$ defined in (2.5.2). Suppose that $(T_1, \delta_1), \ldots, (T_m, \delta_m)$ are observed, where $T_i = \min(X_i, C_i)$, $\delta_i = I\{X_i \leq C_i\}$ and $C_1, \ldots, C_m$ are censoring times. Clearly, $\delta_i = 1$ if X_i is observed, and $\delta_i = 0$ if X_i is right censored. The censoring times can be assumed to be either fixed (i.e. nonrandom) or i.i.d. and independent of the survival times X_i's. Define the *counting process* N by

$$N(t) = \sum_{i=1}^{m} I\{T_i \leq t \text{ and } \delta_i = 1\} \tag{2.5.6}$$

and the (left-continuous) *at-risk process* Y by

$$Y(t) = \sum_{i=1}^{m} I\{T_i \geq t\}, \tag{2.5.7}$$

where I is the indicator function. In particular, $N\{t\}$ is the number of observed X_i's at the exact spot t.

The next theorem, due to Hjort (1990), gives the posterior distribution of the beta process given right censored data.

Theorem 2.5.2 *Let $A \sim BP(c(\cdot), A_0(\cdot))$ as in Definition 2.5.1. Define $(T_i, \delta_i)_{1 \leq i \leq m}$ as described before. Then the posterior distribution of A given $(T_1, \delta_1), \ldots, (T_m, \delta_m)$ coincides with the distribution of the beta process with*

$$c^*(t) = c(\cdot) + Y(\cdot) \quad \text{and} \quad \mathrm{A_0^*(t)} = \int_0^{\mathrm{t}} \frac{\mathrm{c(s)dA_0(s) + dN(s)}}{\mathrm{c(s) + Y(s)}}. \tag{2.5.8}$$

If t_i is a fixed point of discontinuity of A, then the posterior distribution of the jump at t_i is given by

$$S_i^* | \text{data} \sim Beta\left(c(t_i)A_0\{t_i\} + N\{t_i\}, c(t_i)\left(1 - A_0\{t_i\}\right) + Y(t_i) - N\{t_i\}\right) \tag{2.5.9}$$

It is common to start with a continuous prior guess A_0, i.e. $M = \emptyset$ ($A_0 = A_{0,c}$ and $A = A_c$ in the previous definition). However, the posterior process contains fixed points of discontinuity even if the prior distribution does not. In this case, $A_0\{t_i\} = 0$, and hence (2.5.9) simplifies to:

$$S_i^*|\text{data} \sim Beta\left(N\{t_i\}, c(t_i) + Y(t_i) - N\{t_i\}\right). \tag{2.5.10}$$

The Lévy measure L_t^* of $A_c^*(t)$ is given by:

$$L_t^*(ds) = \left[\int_0^t \left(c(z) + Y(z)\right) s^{-1}(1-s)^{c(z)+Y(z)-1} dA_{0,c}^*(z)\right] ds, \tag{2.5.11}$$

where $dA_{0,c}^*(z) = c(z)dA_{0,c}(z)/\left(c(z) + Y(z)\right)$.

We end this section with the following remarks:

Remark 2.5.3 *From Theorem 2.5.2 and (2.5.5), the Bayes estimate of $A(t)$, with respect to the quadratic loss, is given by:*

$$\widehat{A}(t) = \mathrm{E}\left[\mathrm{A(t)|data}\right] = \mathrm{A_0^*},$$

where A_0^* is defined in (2.5.8). From (2.5.2), a nonparametric Bayes estimator of F is:

$$\widehat{F}(t) = 1 - \prod_{[0,t]} \left[1 - \frac{cdA_0 + dN}{c + Y}\right];$$

see also Theorem 4.3 of Hjort. As $c(\cdot)$ decreases to zero, $\widehat{A}$ and $\widehat{F}$ tend to the nonparametric Nelson-Aalen and Kaplan-Meier estimators of A and F, respectively. On the other hand, $\widehat{A}$ and $\widehat{F}$ become the prior guess $\widehat{A}_0$ and $\widehat{F}_0$ as $c(\cdot)$ grows large. This gives one more reason to call $c(\cdot)$ the concentration parameter.

Remark 2.5.4 *Hjort (1990) showed that the Dirichlet process is a particular beta process. Let P be a Dirichlet process on $\mathfrak{X} = \mathbb{R}$, with parameters a and H. Let $A(t) = P\left([0,t]\right)$. Define $c(t) = aH[t,\infty)$ and $A_0(t) = \int_0^t dH(s)/H[s,\infty)$. Hjort proved that $\{A(t)\}_{t\in\geq 0}$ is a beta process with parameter $c(\cdot)$ and $A_0(\cdot)$.*

2.6 The Beta-Stacy Process

The Dirichlet process is not a (parametric) conjugate prior when the sample contains right censored observations. Walker and Muliere (1997) showed that, in this case, the posterior distribution coincides with the distribution of a beta-Stacy process. As in the beta process, the use of the beta-Stacy process in the literature is justified as follows: it generalizes the Dirichlet process and it is conjugate to both exact and right censored observations. The later property makes the beta-Stacy process suitable for applications in survival analysis. In this thesis, we only consider the continuous-time version of the beta-Stacy process. For the discrete-time version, see Walker and Muilere (1997).

A beta-Stacy process is a neutral to the right process which is defined through another Lévy process called the log-beta process. The two definitions below are taken from Walker and Muliere (1997).

Definition 2.6.1 *Let $\beta(\cdot)$ be a positive function and α be a measure concentrated on $\mathbb{R}^+$ which is absolutely continuous with respect to the Lebesgue measure such that $\int_0^\infty d\alpha(z)/\beta(z) = \infty$. A Lévy process $\{Z(t)\}_{t\geq 0}$ with Lévy measure*

$$L_t(ds) = \left[\frac{1}{1-e^{-s}}\int_0^t e^{-s\beta(z)}\alpha(dz)\right]ds, \quad s>0 \tag{2.6.1}$$

is called a log-beta process with parameters α and β.

Note that, the assumption of absolute continuity of α is equivalent to requiring no fixed point of discontinuity be present. However, such discontinuities appear when dealing with the posterior. Next we define the beta-Stacy process.

Definition 2.6.2 *Let $k(\cdot)$ be a positive function defined $\mathbb{R}^+$ and F_0 be a cumulative distribution function on $\mathbb{R}^+$ which is absolutely continuous. Let $\{Z(t)\}_{t\geq 0}$ be the log-beta process with parameters*

$$\alpha(dz) = k(z)F_0(dz) \quad \textit{and} \quad \beta(z) = k(z)F_0([z,\infty)).$$

Then the process $F(t) = 1 - e^{-Z(t)}$ *is called the* beta-Stacy process *on* $\mathbb{R}+$ *with parameters* $k(\cdot)$ *and* F_0.

As shown in Walker and Muliere (1997), $E[F(t)] = F_0(t)$, making F_0 the prior guess. In the original definition of Walker and Muliere (1997), the measures $\alpha(\cdot)$ and $F_0(\cdot)$ are not necessarily absolutely continuous. Since in practice α and G are absolutely continuous, we consider only this case. Thus, the prior distribution of $\{Z(t)\}_{t\geq 0}$ and $\{F(t)\}_{t\geq 0}$ are completely specified by the Lévy measure given by (2.6.1).

Next we describe the posterior distributions of $\{Z(t)\}_{t\geq 0}$ and $\{F(t)\}_{t\geq 0}$. Let $X_1, \ldots, X_m$ be an i.i.d. sample from F. As in the previous section, suppose that $(T_1, \delta_1), \ldots, (T_m, \delta_m)$ is observed, where $T_i = \min(X_i, C_i)$, $\delta_i = I\{X_i \leq C_i\}$ and $C_1, \ldots, C_m$ are censoring times. Define the counting process N of uncensored data and the (left-continuous) at-risk process Y as in (2.5.6) and (2.5.7), respectively.

The following theorem gives the parameters of the posterior of the log-beta process, given possibly right censored observations. For the proof, see Walker and Muliere (1997).

Theorem 2.6.3 *Let* $\{Z(t)\}_{t\geq 0}$ *be a log-beta process with parameters* α *and* β, *and let* $F(t) = 1 - \exp(-Z(t))$. *Given the data* $(T_1, \delta_1), \ldots, (T_m, \delta_m)$ *from* F, $\{Z(t)\}_{t\geq 0}$ *is a log-beta process with parameters*

$$\alpha^*(t) = \alpha(t) + N(t) \tag{2.6.2}$$

and

$$\beta^*(t) = \beta(t) + Y(t) - N\{t\}. \tag{2.6.3}$$

The posterior Lévy measure for $Z(t)$ is given by

$$L_t^*(ds) = \left[\frac{1}{1-e^{-s}} \int_0^t e^{-s(\beta(z)+Y(z))} d\alpha(z)\right] ds. \tag{2.6.4}$$

The posterior process contains fixed points of discontinuity. These extra points occur at the exact (non-censored) observations. If t_i is an exact data with corresponding

jump S_i^*, then

$$1 - \exp(-S_i^*) \sim Beta\left(N\{t_i\}, \beta(t_i) + Y(t_i) - N\{t_i\}\right). \tag{2.6.5}$$

Observe that, if $N\{t_i\} = 1$, then the random jump S_i^* has an exponential density with mean value $[\beta(t_i) + Y(t_i) - N\{t_i\}]^{-1}$.

Corollary 2.6.4 *Let $\{F(t)\}_{t\geq 0}$ be a beta-Stacy process with parameters $k(\cdot)$ and F_0. Let $(T_1, \delta_1), \ldots, (T_m, \delta_m)$ be a sample, possibly with right censoring, from F. Then, given the data, F is a beta-Stacy process with parameters $k^*(\cdot)$ and F_0^*, where*

$$k^*(t) = \frac{\beta^*(t)}{F_0^*[t,\infty)} = \frac{\beta(t) + Y(t) - N\{t\}}{F_0^*[t,\infty)}$$

and

$$F_0^*(t) = 1 - \prod_{[0,t]} \left(1 - \frac{d\alpha^*(z)}{\beta^*(z) + \alpha^*\{z\}}\right) = 1 - \prod_{[0,t]} \left(1 - \frac{k(z)dF_0(z) + dN(z)}{k(z)F_0[z,\infty) + Y(z)}\right),$$

where $\alpha^*(s)$ and $\beta^*(s)$ are defined in (2.6.2) and (2.6.3), and $\prod$ stands for the product integral.

Sampling from a beta-Stacy process relies on sampling from a log-beta process. More details are given in Chapter 4. We end this section with the following remarks:

Remark 2.6.5 *The Bayes estimate of $F(t)$, with respect to the quadratic loss, is given by:*

$$\widehat{F}(t) = E\left[F(t)|data\right] = F_0^*(t), \tag{2.6.6}$$

where $F_0^*(t)$ is defined as in Corollary 2.6.4. As $k(\cdot)$ tends to zero, the nonparametric Kaplan-Meier estimator is obtained. On the other hand, F^* becomes the prior guess F_0 as $k(\cdot)$ grows large. The parameter $k(\cdot)$ can be viewed as the concentration parameter.

Remark 2.6.6 *Walker and Muliere (1997) showed that the beta-Stacy process includes various neutral to the right processes proposed in the literature. For instance, the Dirichlet process is obtained if* F_0 *is continuous and* $k(s) = k > 0$ *for all* $s \geq 0$. *The simple homogenous process (Ferguson and Phadia, 1979) is obtained when* $\beta(\cdot)$ *is constant.*

Chapter 3

Series Representations of Pure Jump Processes

As mentioned in the previous chapter, any Lévy process $\{Z(t)\}_{t\geq 0}$ can be decomposed into two parts: $\{Z_f(t)\}_{t\geq 0}$ and $\{Z_c(t)\}_{t\geq 0}$, where the general forms of these two components are given in (2.3.1) and (2.3.2), respectively. The process $\{Z_f(t)\}_{t\geq 0}$, which is the part due to the existence of fixed points of discontinuity, is completely specified. In this chapter, we will derive the exact form of $\{Z_c(t)\}_{t\geq 0}$.

In the literature, there are two general representations that can be used to specify pure jump processes (Lévy processes with no Gaussian components and no fixed points of discontinuity). The first representation is due to Ferguson and Klass (1972), the other one is due to Wolpert and Ickstadt (1998). These two representations are used in different places in the literature. For example, Ferguson (1973) used the Ferguson and Klass representation (1972) to define and construct the Dirichlet process. Recently, De Blasi, Favaro, and Muliere (2010) adapted the Ferguson and Klass representation to sample from the beta-Stacy process. On the other hand, Wolpert and Ickstadt (1998) used their algorithm to sample from the gamma process, the beta process, the stable process and the simple homogeneous process. Epifani,

Lijoi, and Pruñster (2003) used Wolpert and Ickstadt representation to sample the mean of a random distribution chosen according to the beta-Stacy process.

In this chapter, we recall the representation of Ferguson and Klass (1972) and the representation of Wolpert and Ickstadt (1998). We compare these two representations from the computational point of view and clarify why the representation of Wolpert and Ickstadt is more appropriate for nonhomogeneous processes. Our comparison is different than the one in Walker and Damien (2000), where it is shown that the Wolpert and Ickstadt representation (1998) can be derived from the Ferguson and Klass representation (1972). Our comparison is relatively new and is not emphasized in the literature of Bayesian nonparametric inference. Our last objective is to derive the Ferguson and Klass representation (1972) for the beta process and to use this representation to sample from the prior and the posterior process.

3.1 The Ferguson-Klass Representation

A pure jump process can be written as a sum of a countable number of jumps of random heights at a countable number of random points. Ferguson and Klass (1972) proposed a general representation that can be applied to construct such processes. Let $\{Z_c(t)\}_{t\geq 0}$ be a pure jump process with corresponding Lévy measure $L_t(\cdot)$. We restrict the domain of t to be the interval $[0,T]$, where $T>0$ is fixed. In some cases, we allow that $T=\infty$. We assume that the Lévy measure $L_t(\cdot)$ satisfies the regularity conditions (i)-(iv) stated in Section 2.3 applied to $t\in[0,T]$. Let $J_1, J_2, \ldots$ be nonnegative random variables defined by

$$\Pr\{J_1 \leq x_1\} = \exp\{-L_T([x_1,\infty))\},$$

and for $i=2,3,\ldots$

$$\Pr\{J_i \leq x_i | J_{i-1}=x_{i-1},\ldots,J_1=x_1\} = \exp\{-(L_T([x_i,\infty)) - L_T([x_{i-1},\infty)))\}$$

for $0 < x_i < x_{i-1}$. The nonnegative random variables $J_1, J_2, \ldots$ are called the jumps and can be obtained via

$$\Gamma_i = L_T(J_i), \tag{3.1.1}$$

where Γ_i is given by (2.1.3) and $L_T(x) = L_T([x, \infty))$. Therefore, $J_i = L_T^{-1}(\Gamma_i)$, where $L_T^{-1}(y) = \inf\{x : L_T(x) \geq y\}$. We would like to emphasize that in the Ferguson and Klass representation, the jumps are generated in a decreasing order according to their size, i.e. $J_1 \geq J_2 \geq \cdots$.

Note that by condition (iv) (Section 2.3), L_T satisfies the following condition:

$$\int_0^1 sL_T(ds) < \infty. \tag{3.1.2}$$

If (3.1.2) holds, then the Ferguson and Klass representation of $\{Z_c(t)\}_{t\in[0,T]}$ is given by

$$Z_c(t) = \sum_{i=1}^{\infty} J_i I\{U_i \leq n_t(J_i)\}, \tag{3.1.3}$$

where $U_1, U_2, \ldots$ are i.i.d. random variables with a uniform distribution on $[0, 1]$, independent of $J_1, J_2, \ldots$ and

$$n_t(s) = \frac{L_t(ds)}{L_T(ds)}.$$

For a fixed s, $n_t(s)$ is a nondecreasing function on t, $n_0(s) = 0$ and $n_T(s) = 1$. Thus, $n_t(s)$ behaves like a distribution function on $[0, T]$.

As in Walker and Damien (2000), it is possible to rewrite $Z_c(t)$ in (3.1.3) as follows. Let θ_i be the solution of

$$U_i \;=\; n_{\theta_i}(J_i). \tag{3.1.4}$$

Given the jump J_i, let the cumulative distribution function of the random locations θ_i be such that

$$\Pr\{\theta_i \leq t | J_i\} = n_t(J_i).$$

Note that $\{U_i \leq n_t(J_i)\} = \{\theta_i \leq t\}$ and hence

$$\begin{aligned} Z_c(t) &= \sum_{i=1}^{\infty} J_i I\{\theta_i \leq t\} \\ &= \sum_{i=1}^{\infty} L_T^{-1}(\Gamma_i) I\{\theta_i \leq t\}. \end{aligned} \tag{3.1.5}$$

Observe that, from the simulation point of view, finding J_i (to be used in the Ferguson and Klass representation (3.1.3)) requires solving equation (3.1.1) for J_i. In some cases (e.g. when $\{Z_c(t)\}_{t\in[0,T]}$ is the beta process) solving this equation involves evaluating a double integral and making an inversion to find J_i. This is computationally quite difficult. On the other hand, finding the locations θ_i in the Ferguson and Klass representation requires solving equation (3.1.4) for θ_i. This involves solving a single integral and making an inversion to find θ_i. Therefore, simulating a pure jump Lévy process $\{Z_c(t)\}_{t\in[0,T]}$ using the Ferguson and Klass representation (3.1.3) is very complicated and cannot be implemented using the classical statistical softwares.

A simpler form of the Ferguson and Klass representation can be obtained when the process $\{Z_c(t)\}_{t\in[0,T]}$ is homogeneous, i.e. its Lévy measure $L_t(\cdot)$ is linear in t: $L_t(s) = tL_T(s)$. More generally, if

$$L_t(ds) = G(t)L_T(ds),$$

where $\{G(t)\}_{t\in[0,T]}$ is an absolutely continuous cumulative distribution function, then $n_t(s) = L_t(ds)/L_T(ds) = G(t)$ is independent of s. Hence, the points at which the jumps $J_1, J_2, \ldots$ occur are i.i.d., independent of $J_1, J_2, \ldots$ and having $G(t)$ as a cumulative distribution function (Ferguson and Klass, 1972).

In the next section, we discuss the Wolpert and Ickstadt representation of a pure jump process, which is computationally simpler than that of the Ferguson and Klass representation. However, it is still difficult to implement in practice, as it requires advanced numerical methods.

3.2 The Wolpert-Ickstadt Representation

Wolpert and Ickstadt (1998) introduced a general algorithm for simulating a Lévy process whose Lévy measure $(L_t)_{t\in[0,T]}$ is described below. Let $\{Z_c(t)\}_{t\in[0,T]}$ be a pure jump process as defined in the previous section. Let $(L_t)_{t\in[0,T]}$ be a family of Lévy measures which satisfy the regularity conditions (i)-(iv) mentioned in Section 2.3 applied to $t \in [0,T]$, where $T > 0$ is fixed. Assume that $(L_t)_{t\in[0,T]}$ takes the form

$$L_t(ds) = \left[\int_0^t K(s,z)\eta(dz)\right] ds = \left[\int_0^t \psi(s,z)\Pi(dz)\right] ds, \quad \text{for } t \geq 0, \quad s > 0, \quad (3.2.1)$$

where $\psi(s,z) = \eta([0,T])K(s,z)$ and $\Pi(dz) = \eta(dz)/\eta([0,T])$ is a probability measure on $[0,T]$. In addition, assume that

$$\int_0^T \int_0^\infty (1 \wedge s)\psi(s,z)ds\Pi(dz) < \infty. \quad (3.2.2)$$

Note that, the Lévy measures discussed so far in the thesis satisfy this condition: for the beta process $\psi(s,z) = A_0(T)c(z)s^{-1}(1-s)^{c(z)-1}$ and $\Pi(dz) = [A_0(T)]^{-1}\eta(dz)$, where $\eta([0,t]) = A_0(t)$ (see (2.5.3)), while for the log-beta process $\psi(s,z) = \alpha([0,T])(1-e^{-s})^{-1}e^{-s\beta(z)}$ and $\Pi(dz) = [\alpha([0,T])]^{-1}\alpha(dz)$ (see (2.6.1)). Observe that, condition (3.2.2) in the Wolpert and Ickstadt representation is equivalent to condition (3.1.2) in the Ferguson and Klass representation.

The steps of the Wolpert-Ickstadt algorithm are:

(1) Generate $\theta_i \overset{i.i.d.}{\sim} \Pi$, for $i = 1, 2, \ldots$.

(2) Let $\Gamma_i = E_1 + \cdots + E_i$, where $(E_i)_{i\geq 1}$ are i.i.d. random variables with exponential distribution of mean 1, independent of $(\theta_i)_{i\geq 1}$.

(3) Define

$$M_z(x) = \int_x^\infty \psi(s,z)ds. \quad (3.2.3)$$

(4) For each $i \geq 1$, solve the equation

$$M_{\theta_i}(J_i) \;=\; \Gamma_i \tag{3.2.4}$$

for J_i, where $\Gamma_i = E_1 + \cdots + E_i$, $(E_i)_{i\geq 1}$ are i.i.d. random variables with exponential distribution of mean 1 and independent of $(\theta_i)_{i\geq 1}$.

(5) Set

$$Z_c(t) = \sum_{i=1}^{\infty} M_{\theta_i}^{-1}(\Gamma_i) I\{\theta_i \leq t\}. \tag{3.2.5}$$

The process $\{Z_c(t)\}_{t\in[0,T]}$ given by (3.2.5) is a Lévy process whose Lévy measure is given by (3.2.1). (See Wolpert and Ickstadt (1998) for the proof.)

We end this section by carrying out a brief comparison between the Ferguson and Klass representation and the Wolpert and Ickstadt representation. This comparison is different from the one in Walker and Damien (2000), where it is shown that the Wolpert and Ickstadt representation can be derived from the Ferguson and Klass representation (1972). Our comparison is relatively new and is not emphasized in the literature of Bayesian nonparametric inference. A major difference between the two representations is whether we generate the jump sizes $(J_i)_{i\geq 1}$ first or the jump locations $(\theta_i)_{i\geq 1}$. In the Ferguson and Klass representation, $(J_i)_{i\geq 1}$ are drawn before $(\theta_i)_{i\geq 1}$ while in the Wolpert and Ickstadt representation this order is reversed. Finding J_i in Ferguson and Klass representation requires solving equation (3.1.1) for J_i. In some cases (e.g. when $\{Z_c(t)\}_{t\in[0,T]}$ is the log-beta process) solving this equation involves evaluating a double integral and making an inversion to find J_i. This is computationally intensive. On the other hand, finding J_i in Wolpert and Ickstadt representation requires solving equation (3.2.4), which relies on evaluating a single integral and making an inversion to find J_i. Having a single integral here offers a great advantage of the Wolpert and Ickstadt representation over the Ferguson and Klass representation. Indeed, this enables us in the next section to derive a very simple

approximation for the beta process and the log-beta process. On the other hand, finding the location θ_i in the Ferguson and Klass representation requires solving equation (3.1.4) for θ_i. This involves solving a single integral and making an inversion to find θ_i. Therefore, simulating a pure jump Lévy process $\{Z_c(t)\}_{t\geq 0}$ using the Ferguson and Klass representation is very complicated and cannot be implemented using the classical statistical softwares. Furthermore, finding θ_i in the Wolpert and Ickstadt representation is straightforward as θ_i are generated, in general, from well-known distributions. Thus, from the computational perspective, working with the Wolpert and Ickstadt representation necessitates less effort than working with the Ferguson and Klass representation. However, it is still difficult to work with the representation of Wolpert and Ickstadt since the inverse of M_{θ_i} in (3.2.3) has no closed form. Observe that, if the process $\{Z_c(t)\}_{t\in[0,T]}$ is homogeneous, for instance the gamma process, then the two representations are equivalent as it does not matter if we draw first J_i or θ_i.

3.3 Generating the Beta Process Based on Series Representations

In this section, we derive the Ferguson and Klass representation of the beta process. To the best of our knowledge, this representation has not been discussed in the literature. As for the Wolpert and Ickstadt representation of the beta process, see for example, Lee (2007).

Let $A \sim BP(c(\cdot), A_0(\cdot))$ on $[0,T]$, where $T > 0$ is fixed. As mentioned in Section 2.5, if A has fixed points $M = \{t_1, t_2 \ldots\}$ of discontinuity with random jumps $S_i = A\{t_i\}$, then the process

$$A_c(t) = A(t) - \sum_{i\geq 1} S_i I\{t_i \leq t\}$$

does not have any fixed points of discontinuity and its Lévy measure is given by (2.5.3). We assume that $A_{0,c}(T) < \infty$, where $A_{0,c}$ is defined by (2.5.4). We show first that this Levy measure satisfies the regularity conditions (i)-(iv) mentioned in Section 2.3. Conditions (i)-(iii) are clearly satisfied. It remains to verify condition (iv): by Fubini's theorem

$$\int_0^1 sL_T(ds) = \int_0^T \int_0^1 sc(z)s^{-1}(1-s)^{c(z)-1}dsdA_{0,c}(z) = \int_0^T dA_{0,c}(z) = A_{0,c}(T) < \infty.$$

The Ferguson and Klass representation for $A_c(t)$ is given by (3.1.5) with

$$n_t(s) = \frac{\int_0^t c(z)s^{-1}(1-s)^{c(z)}dA_{0,c}(z)}{\int_0^T c(z)s^{-1}(1-s)^{c(z)}dA_{0,c}(z)} = \frac{\int_0^t c(z)(1-s)^{c(z)}dA_{0,c}(z)}{\int_0^T c(z)(1-s)^{c(z)}dA_{0,c}(z)}. \tag{3.3.1}$$

Note that, when $c(z) = c$, i.e. $\{A(t)\}_{t\in[0,T]}$ is a homogeneous process, then (3.3.1) reduces to $n_t(s) = A_{0,c}(t)/A_{0,c}(T)$.

Next we discuss briefly how to sample the beta process for both the prior and the posterior case, based on the Ferguson and Klass representation. Applying this algorithm is computationally intensive and requires special numerical methods.

Algorithm A: Ferguson and Klass representation for simulating the prior beta process. Although it is possible to find formulas for the beta process in the presence of fixed points of discontinuity, we do not consider this for the prior process. Hence, we take $M = \emptyset$ (i.e. $A_0 = A_{0,c}, A = A_c$). Thus, as a prior, the beta process is characterized entirely by its Lévy measure. The steps for simulating the prior beta process with parameters $c(\cdot)$ and $A_0(\cdot)$ are the following:

(1) Let $\Gamma_i = E_1 + \cdots + E_i$, where $(E_i)_{i\geq 1}$ are i.i.d. random variables with exponential distribution of mean 1.

(2) For each $i \geq 1$, let J_i be the solution of $\Gamma_i = L_T(J_i)$, where $L_T(x) = L_T([x,1))$, $x > 0$ and the measure L_t is given by (2.5.3).

(3) Generate i.i.d. random variables $(U_i)_{i\geq 1}$ from the uniform distribution on $[0,1]$, independent of $(E_i)_{i\geq 1}$.

(4) For $i \geq 1$, let θ_i be the solution of $U_i = n_{\theta_i}(J_i)$ in $[0, T]$ where $n_t(s)$ is given by (3.3.1).

The process $A(t) = \sum_{i=1}^{\infty} J_i I(\theta_i \leq t) = \sum_{i=1}^{\infty} L_T^{-1}(\Gamma_i) I(\theta_i \leq t)$ is a beta process with parameters $c(\cdot)$ and $A_0(\cdot)$. This series is an infinite series. In practice, we truncate this series and use the approximation

$$A_n(t) = \sum_{i=1}^{n} J_i I(\theta_i \leq t) = \sum_{i=1}^{n} L_T^{-1}(\Gamma_i) I(\theta_i \leq t).$$

Note that, unlike the two-parameter Poisson-Dirichlet process, the weights $(J_i)_{i \geq 1}$ in the beta process do not add to one.

Algorithm B: Ferguson and Klass representation for simulating the posterior beta process. As mentioned in Section 2.5, sampling form the posterior beta process consists of two tasks:

(a) Simulating from the posterior process $A_c^*(t)$, which is a pure jump process with no fixed point of discontinuity whose Lévy measure is known. The updated Lévy measure is given by (2.5.11). Thus, the algorithm used for simulating the prior process can be used for simulating the posterior process with $c(z)$ replaced by $c(z) + Y(z)$ and $dA_0(z)$ replaced by $c(z)dA_0(z)/(c(z) + Y(z))$. Denote the approximated process by $A_{c,n}^*(t)$.

(b) Generating random jumps $\{S_i^* : \delta_i = 1, i = 1, \ldots, m\}$ from the distribution given in (2.5.10). These jumps are associated with the fixed jumps of discontinuity $\{T_i : \delta_i = 1, i = 1, \ldots, m\}$ and occur at points where the data is right censored.

Putting the simulation for the two components together results in a sample from the posterior beta process. Hence, the posterior process $A^*(t)$ is approximated by:

$$A_n^*(t) \quad = \quad A_{c,n}^*(t) + A_f^*(t)$$

$$= \sum_{i=1}^{n} J_i^* I(\theta_i^* \leq t) + \sum_{i=1}^{l} S_i^* I\left\{T_i \leq t\right\},$$

where $(J_i^*)_{1\leq i\leq n}$ and $(\theta_i^*)_{1\leq i\leq n}$ are the jumps and the locations of the approximated process $A_{c,n}^*(t)$, and $l \leq m$ is the number of distinct non-censored (exact) observations.

In summary, the weights in the Ferguson and Klass representation are always strictly decreasing. This feature of the jumps is very important for simulation purposes. On the other hand, excluding the homogeneous processes (e.g. the gamma process), the weights in the Wolpert and Ickstadt representation are not strictly decreasing. In fact, the weights in the Wolpert and Ickstadt representation are unordered. Except for this undesirable property of unordered weights, the Wolpert and Ickstadt representation has several advantages over the Ferguson and Klass representation. In fact, it is much easier to simulate than that of Ferguson and Klass. This justifies the frequent use of the Wolpert and Ickstadt representation in the literature of Bayesian nonparametric inference. However, applying the Wolpert and Ickstadt representation requires finding $M_{\theta_i}^{-1}(\Gamma_i)$ at each step (see Section 3.2). Since $M_{\theta_i}(x)$ is not a standard mathematical function, finding $M_{\theta_i}^{-1}(\Gamma_i)$ at each step is not straightforward. To bypass this major problem, we offer in the next section a finite sum approximation which converges a.s. to the Wolpert and Ickstadt representation. This new approach is very general and can be applied to any process whose Lévy measure is given by (3.2.1) and satisfies (3.2.2).

Chapter 4

Simulation of Bayesian Nonparametric Priors

In this chapter, new approximations of the Dirichlet process, the two-parameter Poisson-Dirichlet process, the beta process and the beta-Stacy process are developed. These new approximations provide simple, yet efficient, procedures for simulating these important processes. We compare the efficiency of the new approximations to several other well-known approximations and demonstrate a significant improvement. The material included in Section 4.1 is published in Zarepour and Al Labadi (2012).

4.1 Rapid Simulation of the Dirichlet Process

The definition of the Dirichlet process and several of its series representations with various approximations have been discussed in Section 2.1. In this section, we derive a finite sum representation involving monotonically decreasing weights and show that it converges almost surely to Ferguson's representation of the Dirichlet process. This is one of the main contributions of this chapter. An extensive simulation study evaluating the accuracy of our method and its comparison to several well-known

approaches is also presented. Our results show that the new approximation is faster, simpler, and more efficient.

4.1.1 Introduction

Simulating from the Dirichlet process plays a crucial role in Bayesian nonparametric inference. For example, evaluating the distribution of random variables of the form

$$T(P) = \int g(x)dP(x),$$

where $g(x)$ is an arbitrary a measurable function and $P \sim DP(a, H)$, is considered by many authors. Cifarelli and Regazzini (1990) found the probability density function for the mean, i.e. when $g(x) = x$. In some special cases, the density can be expressed in a closed form. For example, when $g(x) = x$ and H is the uniform distribution on $[0, 1]$, the density equals (Diaconis and Kemperman, 1996, Regazzini, Guglielmi and Di Nunno, 2002; Lijoi and Regazzini, 2004)

$$\frac{e}{\pi}(1-x)^{x-1}x^{-x}\sin(\pi x), \quad x \in [0,1]. \tag{4.1.1}$$

Sometimes the distribution of the random functional $T(P)$ is derived through integral transforms, but these are often difficult to utilize. Due to the complexity of this problem, the distribution of $T(P)$ is simply approximated by the empirical distribution of $T(P_n)$ where P_n is an approximation of P. Muliere and Tardella (1998) used an approximation procedure based on the stick-breaking representation given by Sethuraman (1994). They defined a random probability measure P_n that is the sum of n terms in the stick-breaking series representation, where $n = n(\epsilon)$ is chosen by a random stopping rule for some small $\epsilon > 0$ tolerance value (see also Subsection 2.1.2). By repeatedly selecting a random draw from P_n, we can approximate the distribution of $T(P)$ by $T(P_n)$.

The Dirichlet process is also employed as a prior in the Bayesian hierarchical

mixture model:

$$x_i|\theta_i \overset{\text{ind.}}{\sim} K(x_i|\theta_i), \quad \theta_i \overset{\text{i.i.d.}}{\sim} P, \quad P \sim DP(a, H), \quad i = 1, \ldots, n,$$

where $K(x_i|\theta_i)$ is a probability distribution with parameter θ_i (when K is a normal density, then the above becomes a normal mixture model). Dirichlet process mixtures models (DPM) were introduced by Ferguson (1983) and Lo (1984) for density estimation. Escobar and West (1995) considered normal DPM and provided MCMC algorithms for the computation of the posterior distribution. Construction of more flexible models using the Dirichlet process became possible with the development of methods for nonconjugate models (MacEachern and Müller, 1998; Walker and Damien, 1998; Neal, 2000). These methods use a representation where the Dirichlet process is integrated out. There are methods that consider approximations to the Dirichlet process instead of integrating it out. For example see Ishwaran and James (2001) and Kottas and Gelfand (2001). Often the simulated values obtained from MCMC iterations are used to approximate posterior functionals of the Dirichlet process. Therefore, a rapid method for simulating such values is valuable.

Other recent areas of application of sampling from the Dirichlet process include testing statistical hypotheses (Muliere and Tardella, 1998; Swartz, 1999), finance (Kacperczyk, Damien and Walker, 2003; Zarepour, Bedard and Dabrowski, 2008), econometrics (Chib and Hamilton, 2002), epidemiology (Dunson, 2005), genetics (Dunson, Herring and Mulheri-Engel, 2008), medicine (Kottas, Branco and Gelfand, 2002), and machine learning (Blei, Ng and Jordan , 2003). Since the Dirichlet process is a weighted average of an infinite number of atoms, having an exact sample from the Dirichlet process is crucial. Recent approaches include the sliced sampling technique of Griffin and Walker (2011) and retrospective sampling by Papaspiliopoulos and Roberts (2008). However, these are very complex algorithms and may be difficult to apply for many users. In the next subsection, we propose a simple yet highly accurate approximation to Ferguson's representation of the Dirichlet process.

4.1.2 Monotonically Decreasing Approximation of the Dirichlet Process

Various approximations of the Dirichlet process have been discussed in Section 2.1. The drawbacks associated with applying those approximations to sample from the Dirichlet process are explained in details in that section. In this subsection, we derive a finite sum representation which converges almost surely to Ferguson's sum representation of the Dirichlet process. This representation gives a new simple way to approximate the Dirichlet process. The new representation requires fewer number of terms for a more precise approximation of the Dirichlet process and yields far smaller variability among the weights.

Let X_n be a random variable with distribution $Gamma(a/n, 1)$. Define

$$G_n(x) = \Pr(X_n > x) = \int_x^\infty \frac{1}{\Gamma(a/n)} e^{-t} t^{a/n-1} dt. \tag{4.1.2}$$

and

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

The following proposition describes properties of $G_n(x)$.

Proposition 4.1.1 *For $x > 0$, the function $G_n(x)$ defined in (4.1.2) has the following properties as $n \to \infty$:*

(i) $nG_n(x) \to L(x)$, where $L(x)$ is defined by (2.1.5).

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

Proof: To prove (i), note that for any $x > 0$,

$$\Gamma(x) = \frac{\Gamma(x+1)}{x}. \tag{4.1.3}$$

With $x = a/n$ in (4.1.3) we obtain

$$\frac{n}{\Gamma(a/n)} = \frac{a}{\Gamma(a/n+1)}. \tag{4.1.4}$$

But for $x > 0$, $\Gamma(x)$ is a continuous function. Thus, from (4.1.4), we get

$$\frac{n}{\Gamma(a/n)} \to a. \tag{4.1.5}$$

Next we apply the dominated convergence theorem. First notice that, since any convergent sequence of real numbers is bounded, $n/\Gamma(a/n)$ is bounded by, say, c. There are two cases to consider. When $x > 1$, then the dominator is ce^{-t}. When $x < 1$, we write the integral as a sum of two integrals: from x to 1 and from 1 to ∞. For the integral from x to 1, the dominator is $cx^{-1}e^{-t}$. For the integral from 1 to ∞, the dominator is ce^{-t}. Hence, the dominated convergence theorem applies and we have

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

To prove (ii), notice that the left hand side of (i) is a sequence of monotone functions converging to a continuous monotone function for every $x > 0$ (de Haan and Ferreira, 2006, page 5). Thus, (i) is equivalent to $G_n^{-1}(x/n) \to L^{-1}(x)$. ∎

Proposition 4.1.1 gives a simple procedure for an approximate evaluation of both $L(x)$ and $L^{-1}(x)$ for any $x > 0$. For computational simplicity, a more convenient approximation is presented in the following Corollary. The proof follows straightforwardly by taking $x = \Gamma_i$ in Proposition 4.1.1 and using the fact that $\Gamma_{n+1}/n \xrightarrow{a.s.} 1$ as $n \to \infty$ (by the strong law of large numbers).

Corollary 4.1.2 *For any fixed $i \geq 1$, we have:*

$$G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right) \xrightarrow{a.s.} L^{-1}(\Gamma_i),$$

as $n \to \infty$.

The utility of Corollary 4.1.2 stems from the fact that all values of $G_n^{-1}(\Gamma_i/\Gamma_{n+1})$ are nonzero for $i \leq n$. This is not the case when working with $G_n^{-1}(\Gamma_i/n)$.

The following lemma provides a finite sum representation which converges almost surely to the Ferguson and Klass (1972) sum representation for the gamma process. Convergence of all random measures is taken with respect to the vague topology. See Appendix B for background on convergence of random measures. In the proof of the next lemma, let $\left(\left[G_n^{-1}\left(\Gamma_i/\Gamma_{n+1}\right)\right]'\right)_{1\leq i\leq n}$ be the corresponding weights associated with the ordered locations $\left(\theta_{(i)}\right)_{1\leq i\leq n}$ such that $\sum_{i=1}^{n}\left[G_n^{-1}\left(\Gamma_i/\Gamma_{n+1}\right)\right]'\delta_{\theta_{(i)}} = \sum_{i=1}^{n} G_n^{-1}\left(\Gamma_i/\Gamma_{n+1}\right)\delta_{\theta_i}$, where $\theta_{(1)} \leq \cdots \leq \theta_{(n)}$ represent the order statistics of $\theta_1, \ldots, \theta_n$.

Lemma 4.1.3 *If $(\theta_i)_{i\geq 1}$ is a sequence of i.i.d. random variables with common distribution H, independent of $(\Gamma_i)_{i\geq 1}$, then as $n \to \infty$*

$$\mathcal{G}_n = \sum_{i=1}^{n} G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)\delta_{\theta_i} \overset{a.s.}{\to} \mathcal{G} = \sum_{i=1}^{\infty} L^{-1}\left(\Gamma_i\right)\delta_{\theta_i}.$$

Here, $(\Gamma_i)_{i\geq 1}$, $L(x)$ and $G_n(x)$ are defined in (2.1.3), (2.1.5) and (4.1.2), respectively.

Proof: We will show that

$$\sum_{i=1}^{n}\left[G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)\right]'\delta_{\theta_{(i)}} \overset{a.s.}{\to} \sum_{i=1}^{\infty}\left[L^{-1}\left(\Gamma_i\right)\right]'\delta_{\theta_{(i)}},$$

as $n \to \infty$. By Lemma 2 of Grandell (1977) (also see the proof of Theorem 3), it is enough to show that, for all k fixed,

$$\mathcal{G}_n^{(k)} = \sum_{i=1}^{k}\left[G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)\right]'\delta_{\theta_{(i)}} \overset{a.s.}{\to} \mathcal{G}^{(k)} = \sum_{i=1}^{k}\left[L^{-1}\left(\Gamma_i\right)\right]'\delta_{\theta_{(i)}},$$

as $n \to \infty$. This is clear since $\left[G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)\right]' \overset{a.s.}{\to} \left[L^{-1}\left(\Gamma_i\right)\right]'$ as $n \to \infty$. ∎

From Lemma 4.1.3, by normalizing the finite sum $\mathcal{G}_n$, it is possible to obtain a sum representation that converges almost surely to Ferguson's representation of the Dirichlet process. This important result is stated formally in the next theorem.

Theorem 4.1.4 *Let $(\theta_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with values in $\mathfrak{X}$ and common distribution H, independent of $(\Gamma_i)_{i\geq 1}$. Then as $n \to \infty$,*

$$P_n^{new} = \sum_{i=1}^{n} \frac{G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)}{\sum_{i=1}^{n} G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)} \delta_{\theta_i} \overset{a.s.}{\to} P^{Ferg.} = \sum_{i=1}^{\infty} \frac{L^{-1}(\Gamma_i)}{\sum_{i=1}^{\infty} L^{-1}(\Gamma_i)} \delta_{\theta_i}. \tag{4.1.6}$$

Here $(\Gamma_i)_{i\geq 1}$, $L(x)$ and $G_n(x)$ are defined in (2.1.3), (2.1.5) and (4.1.2), respectively.

In fact, for any $1 \leq i \leq n$, $\Gamma_i/\Gamma_{n+1} < \Gamma_{i+1}/\Gamma_{n+1}$ almost surely. Since G_n^{-1} is a decreasing function, we have $G_n^{-1}(\Gamma_i/\Gamma_{n+1}) > G_n^{-1}(\Gamma_{i+1}/\Gamma_{n+1})$ almost surely. That is, the weights of the new representation given in Theorem 4.1.4 decrease monotonically for any fixed positive integer n. Thus, we anticipate that this new representation will yield highly accurate approximations of the Dirichlet process. The next subsection offers strong empirical evidence to support this claim.

The next algorithm uses Theorem 4.1.4 to generate a sample from the approximate Dirichlet process with parameters a and H.

Algorithm C: New algorithm for simulating the Dirichlet process.

(1) Fix a relatively large positive integer n (for example if $\epsilon = 10^{-10}$, use $\bar{n}$ in Table 4.1).

(2) Generate $\theta_i \overset{\text{i.i.d.}}{\sim} H$ for $i = 1, \ldots, n$.

(3) For $i = 1, \ldots, n+1$, generate E_i from an exponential distribution with mean 1, independent of $(\theta_i)_{1\leq i\leq n}$ and let $\Gamma_i = E_1 + \cdots + E_i$.

(4) For $i = 1, \ldots, n$, compute $G_n^{-1}(\Gamma_i/\Gamma_{n+1})$, which is simply the quantile function of the $Gamma(a/n, 1)$ distribution evaluated at $1 - \Gamma_i/\Gamma_{n+1}$.

(5) Set P_n^{new} as defined in Theorem 4.1.4.

4.1.3 Empirical Results: Comparison with Other Methods

In this section, we compare the new approximation of the Dirichlet process (Algorithm C) with the three approximation methods described in Subsection 2.1.2 (Wolpert and Ickstadt (WI), Bondesson and stick-breaking). In what follows, let $(w_{n,i})_{1\leq i\leq n}$ be the weights (i.e. probability multiples of the Dirac measure) in any of these four different approximations of the Dirichlet process. We denote by $w_{n,n}$ the nth weight. Although the new technique is not based on a truncation method, for comparison purposes a random stopping rule similar to that given by (2.1.7) can be developed for the new approximation. Specifically in the case of our new approximation, for a given tolerance value of $\epsilon \in (0,1)$, $n = n(\epsilon)$ can be selected by

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

For each of the four approximations, we computed the truncation value $n = n(\epsilon)$ that makes the weight $w_{n,n}$ less than $\epsilon = 10^{-10}$. For each approximation, we generated 100 independent values for the random stopping value n. That is, we obtained $(n^{(i)})_{1\leq i\leq 100}$, where each $n^{(i)} = n^{(i)}(\epsilon)$ denotes the truncation value for the ith simulation such that $w_{n^{(i)},n^{(i)}} < 10^{-10}$. We denote the averaged truncation values by $\bar{n} = \overline{n(\epsilon)} = \sum_{i=1}^{100} n^{(i)}/100$ and the sample variance $S_n^2 = \frac{1}{100}\sum_{i=1}^{100}\left(n^{(i)} - \bar{n}\right)^2$. From Table 4.1, one can immediately see that the new approximation has the smallest $\bar{n}$ and S_n^2 in each example.

The box plots given in Figure 4.1 illustrate the distribution of the truncation value $n = n(\epsilon)$ for $a = 1$ and $a = 10$, respectively.

For a given n and a, we calculated 100 values of the nth weight $w_{n,n}$ for each approximation. That is, we obtained $(w_{n,n}^{(i)})_{1\leq i\leq 100}$, where $w_{n,n}^{(i)}$ is the nth weight in the ith simulation method. Table 4.2 depicts the average weight $\overline{w}_n = \sum_{i=1}^{100} w_{n,n}^{(i)}/100$ for each approximation. For instance, for $n = 500$ and $a = 10$ the size of the 500th (average) weight in the new method is 1.92×10^{-117}, while they are larger than

Table 4.1: $\bar{n} = \overline{n(\epsilon)}$ is the average obtained from repeating the simulations 100 times so that the nth weight in each simulation is less than $\epsilon = 10^{-10}$. The sample variance S_n^2 is computed for the 100 values of the nth weight.

	New		WI		Bondesson		Stick-breaking	
a	$\bar{n}$	S_n^2	$\bar{n}$	S_n^2	$\bar{n}$	S_n^2	$\bar{n}$	S_n^2
1	9	2.2	21	8.0	20	9.1	21	8.5
5	23	13.3	95	31.1	90	53.9	90	45.1
10	38	23.6	185	56.63	173	180.3	171	167.3
20	61	66.5	362	105.2	327	519.3	322	843.5
50	123	213.4	869	5449.1	730	3875.4	731	3257.2
100	212	375.2	1701	11397.9	1297	23817.0	1321	16365.4

1.92×10^{-26} for the other methods. It is clear from the table that the nth weight from the new approximation is substantially smaller than that of other representations. This shows that the convergence rate of the new representation is empirically faster than other representations.

To study the variability of the weights in the Dirichlet process, for each approximation, we computed 50 sequential values of the weights with certain n and a. The corresponding plots for these weights are given in Figure 4.2, where the x-axis represents the numbers $i = \{1, 2, \ldots, n\}$ and the y-axis represents the corresponding weights $(w_{n,i}^{(j)})_{1 \leq i \leq n}$, for $j = 1, \ldots, 50$. The figure shows that the variability among the weights in the new representation is much less than that of other representations. It also confirms that the weights of the new approximation method vanish much earlier than others.

Figure 4.3 compares the graph of density of the random variable $T(P)$, discussed in Subsection 4.1.1, with its approximations $T(P_n)$. The closed form of the true

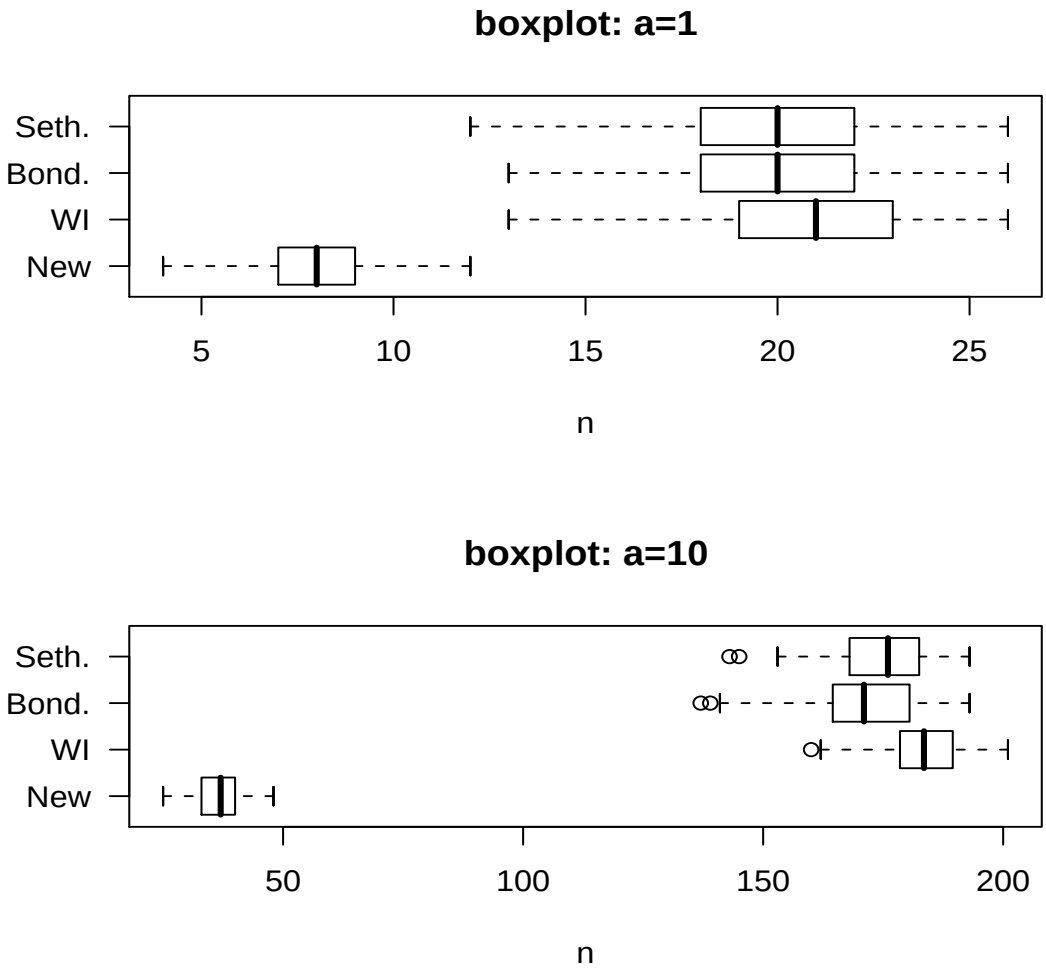

Figure 4.1: Box plot of 100 values of $n = n(\epsilon)$ in four representations when $a = 1$ and $a = 10$. The average and the variance of the 100 values of the nth weight are given in Table 1. Here WI, Bond. and Seth. stand for Wolpert and Ickstadt, Bondesson and Sethuraman (stick-breaking) representations, respectively.

density is given by (4.1.1). The true expectation of $T(P)$ is equal to 0.5. Each approximation is obtained by drawing 20000 independent samples by using one of the four approximations. The figure clearly shows that the new method performs

Table 4.2: For a given n and a, the (average) value of the nth weight in each representation is computed. Each simulation was repeated 100 times.

n	a	$\overline{w}_n$: New	$\overline{w}_n$: WI	$\overline{w}_n$: Bondesson	$\overline{w}_n$: Stick-breaking
200	1	0	4.76×10^{-83}	2.62×10^{-88}	9.30×10^{-92}
	10	1.36×10^{-66}	6.26×10^{-11}	4.90×10^{-13}	1.47×10^{-10}
	50	3.92×10^{-17}	3.01×10^{-06}	7.44×10^{-07}	3.90×10^{-04}
	100	5.56×10^{-11}	1.05×10^{-05}	5.76×10^{-06}	1.13×10^{-03}
500	1	0	7.30×10^{-246}	4.71×10^{-219}	4.90×10^{-211}
	10	1.92×10^{-117}	1.40×10^{-26}	3.73×10^{-26}	4.65×10^{-23}
	50	6.01×10^{-37}	2.78×10^{-09}	4.78×10^{-09}	2.96×10^{-07}
	100	6.42×10^{-20}	3.12×10^{-07}	1.52×10^{-07}	1.06×10^{-04}
1000	1	0	0	0	0
	10	0	2.87×10^{-47}	8.48×10^{-45}	2.21×10^{-46}
	50	3.16×10^{-70}	8.53×10^{-14}	1.16×10^{-14}	3.20×10^{-11}
	100	3.98×10^{-37}	3.50×10^{-09}	1.31×10^{-08}	2.63×10^{-07}
5000	1	0	0	0	0
	10	0	2.21×10^{-220}	4.01×10^{-219}	1.96×10^{-216}
	50	0	2.13×10^{-48}	2.98×10^{-49}	2.00×10^{-46}
	100	2.88×10^{-191}	6.97×10^{-27}	7.44×10^{-28}	2.48×10^{-24}

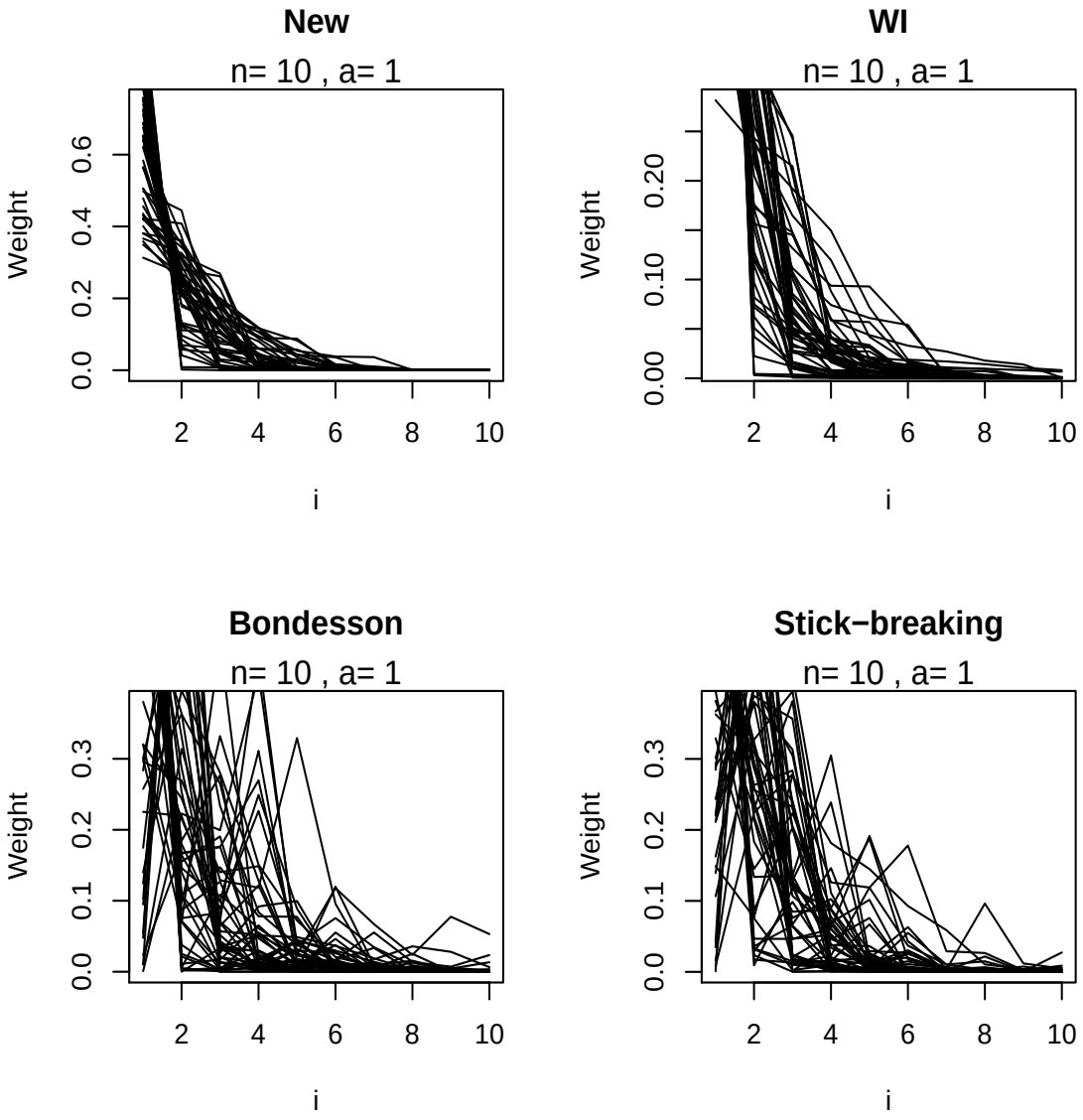

Figure 4.2: Plots of the sequence of weights (probability multiples of the Dirac measure) of the four representations. The x-axis represents the numbers $i = \{1, 2, \ldots, n\}$ and the y-axis represents the corresponding weights.

perfectly in this example.

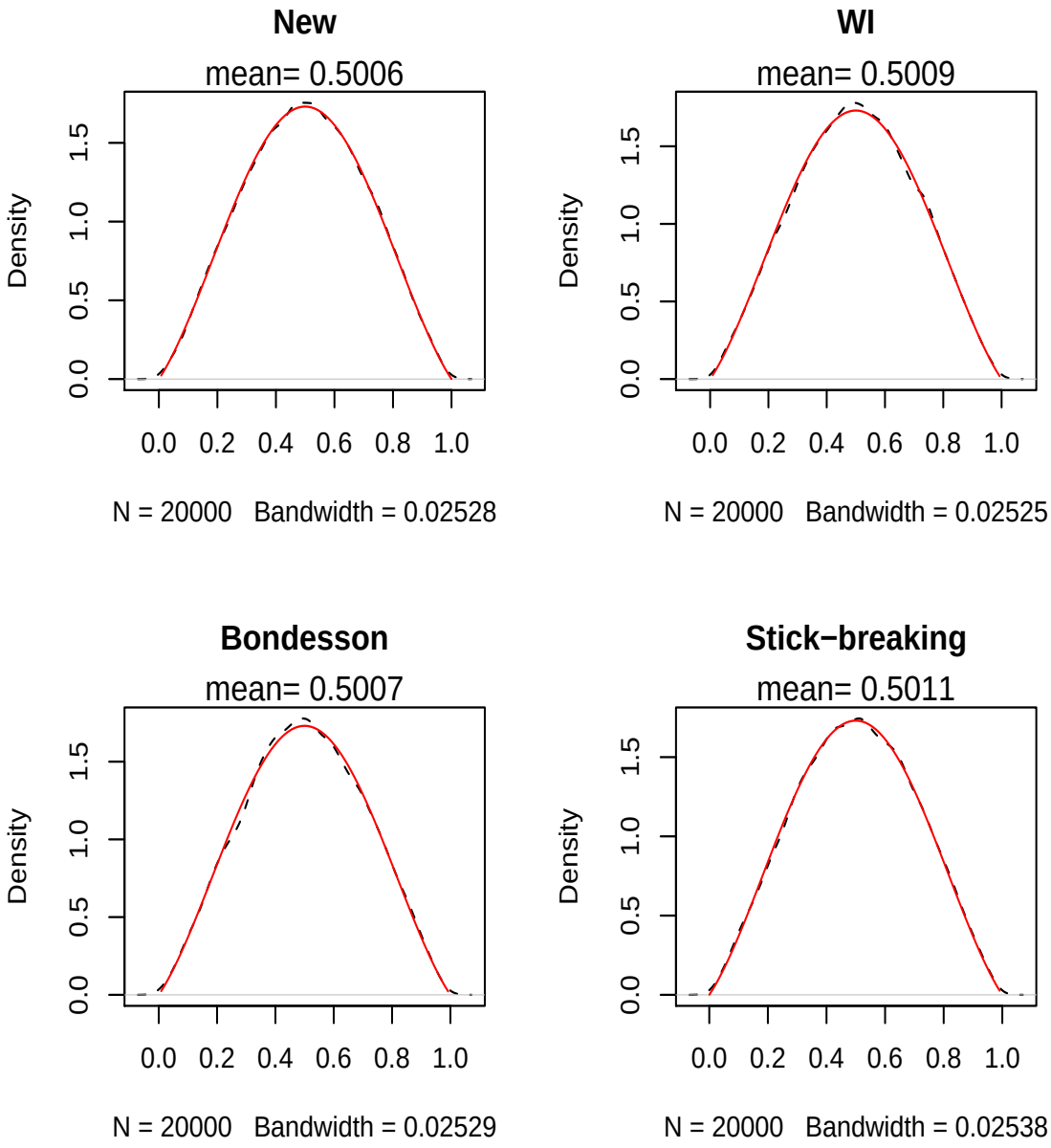

Figure 4.3: Comparisons of the true (solid line) and the approximated (dashed) density of the random variable $T(P)$ discussed in the Subsection 4.1.1.

Finally, Figure 4.4 provides approximate simulations of a Dirichlet process with $H = N(0, 1)$ and $a = 100$ using each of the methods. Clearly, the new approximation performs very well.

4.2 An Accurate Algorithm for Simulating the Two-Parameter Poisson-Dirichlet Process

As in the Dirichlet process, the stick-breaking representation of the two-parameter Poisson-Dirichlet process (see Definition 2.2.1) can be used to approximately simulate the two-parameter Poisson-Dirichlet process using a truncation argument. By truncating the higher order terms in the sum (2.2.1), we can approximate the stick breaking representation by:

$$P^n_{H,\theta,a}(\cdot) = \sum_{k=1}^{n} p_i \delta_{\theta_i}(\cdot). \tag{4.2.1}$$

Here, $(\beta_i)_{i\geq 1}$, $(p_i)_{i\geq 1}$, and $(\theta_i)_{i\geq 1}$ are as given by Definition 2.2.1 with $\beta_n = 1$ (hence β_n does not have a beta distribution). The assumption that $\beta_n = 1$ is necessary to make the weights add to 1, almost surely (Ishwaran and James, 2001). A random stopping rule for choosing $n = n(\epsilon)$, where $\epsilon \in (0, 1)$, is:

$$n = \inf \left\{ i : \widetilde{p}_i = (1 - \beta_1) \dots (1 - \beta_{i-1})\beta_i < \epsilon \right\}. \tag{4.2.2}$$

The random stoping rule in (4.2.2) is similar to the one in (4.2.2) proposed by Muliere and Tradella (1998) for the Dirichlet process. The following lemma shows that the weights $(p_i)_{\imath\geq 1}$ in the stick-breaking representation are not strictly decreasing, almost surely (they are only stochastically decreasing). This makes the truncated stick-breaking representation inefficient for simulation purposes.

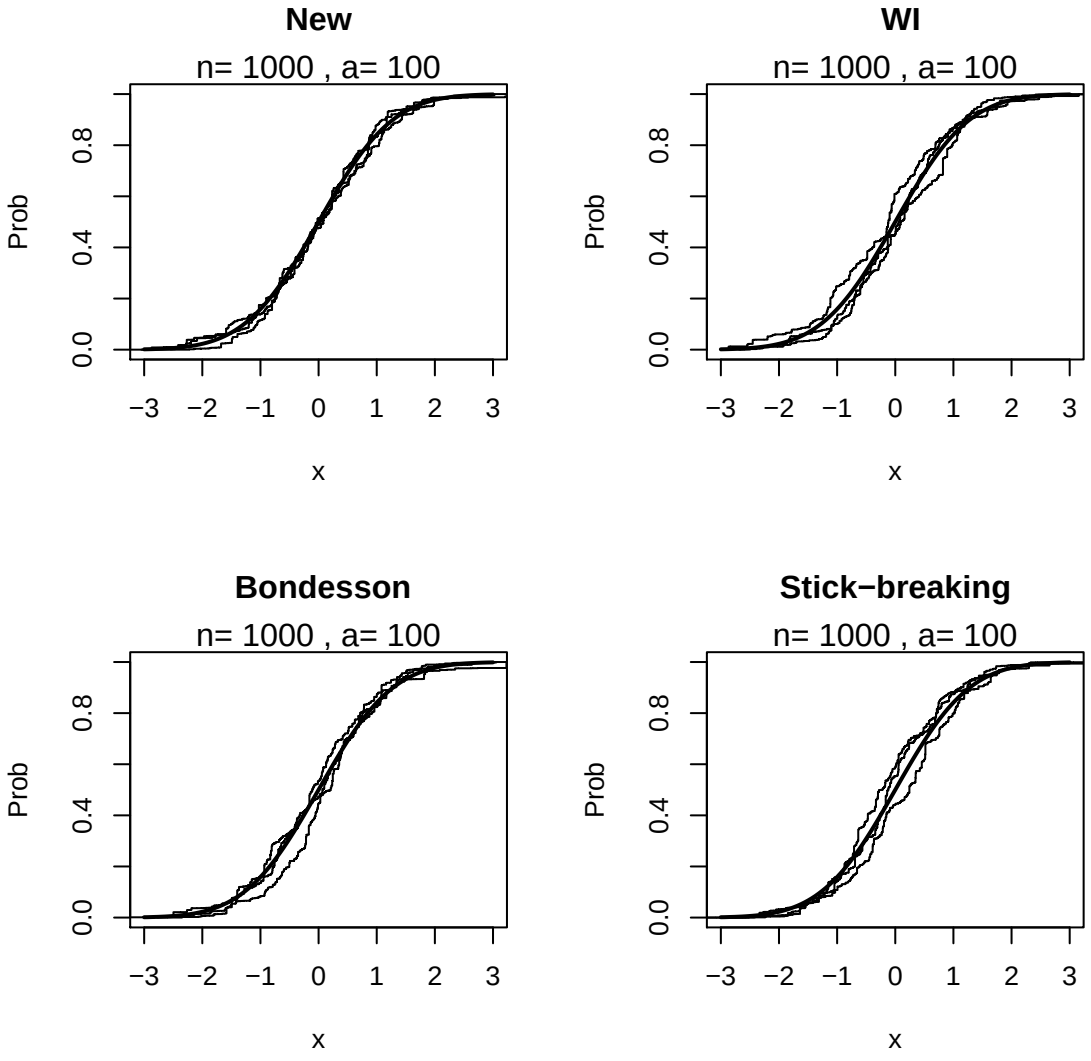

Figure 4.4: Sample paths of a Dirichlet process with $H = N(0, 1)$ and $a = 100$. The solid (thick) line denotes the cumulative distribution function of H.

Lemma 4.2.1 *Let* $(\widetilde{p}_i)_{i\geq 1}$ *be as in Definition 2.2.1. We have* $\Pr\{\widetilde{p}_{i+1} < \widetilde{p}_i\} = \int_0^1 \int_0^y f(x,y)dxdy$, *where*

$$f(x,y) = \frac{x^{\alpha_1-1}(1+x)^{-\alpha_1-\beta_1}}{B(\alpha_1,\beta_1)} \times \frac{y^{\alpha_1-1}(1-y)^{\beta_2-1}}{B(\alpha_1,\beta_2)} I\{x \geq 0\} I\{0 < y < 1\},$$

$B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, $\alpha_1 = 1-\theta$, $\beta_1 = a+i\theta$ *and* $\beta_2 = a+(1+i)\theta$.

Proof: Since $\widetilde{p}_i = \beta_i \prod_{k=1}^{i-1}(1-\beta_k)$, we have

$$\begin{aligned} \Pr\{\widetilde{p}_{i+1} < \widetilde{p}_i\} &= \Pr\{\beta_{i+1}(1-\beta_i) < \beta_i\} \\ &= \Pr\left\{\beta_{i+1}\frac{(1-\beta_i)}{\beta_i} < 1\right\}. \end{aligned}$$

Since β_i is a random variable with the $Beta(1-\theta, a+i\theta)$ distribution, it follows that $\beta_i/(1-\beta_i)$ has the beta distribution of the second kind with parameters $\alpha_1 = 1-\theta$ and $\beta_1 = a+i\theta$ (Balakrishnan and Lai, 2009, page 12). That is, $\beta_i/(1-\beta_i)$ has the density

$$f(x) = \frac{x^{\alpha_1-1}(1+x)^{-\alpha_1-\beta_1}}{B(\alpha_1,\beta_1)} I\{x \geq 0\}.$$

The lemma follows from the fact that $(\beta_i)_{i\geq 1}$ is a sequence of independent random variables with a $Beta(1-\theta, a+i\theta)$ distribution. ∎

It follows clearly from Lemma 4.2.1 that the probability $\Pr\{\widetilde{p}_{i+1} < \widetilde{p}_i\}$ depends on i, θ and a. Table 4.3 depicts some values for this probability.

The next proposition provides an interesting approach to construct a two-parameter Poisson-Dirichlet process. For the proof of the proposition, see Pitman and Yor (1997, Proposition 22).

Proposition 4.2.2 *For* $0 < \theta < 1$ *and* $a > 0$, *suppose* $(p_1(0,a), p_2(0,a), \ldots)$ *and* $(p_1(\theta,0), p_2(\theta,0), \ldots)$ *has respective distributions* $PD(0,a)$ *and* $PD(\theta,0)$. *Independent of* $(p_1(0,a), p_2(0,a), \ldots)$, *let* $(p_1^i(\theta,0), p_2^i(\theta,0), \ldots)$, $i = 1,2,\ldots$, *be a sequence of independent copies of* $(p_1(\theta,0), p_2(\theta,0), \ldots)$. *Let* $(p_i)_{i\geq 1}$ *be the descending order statistics of* $\{p_i(0,a)p_i^j(\theta,0), i,j = 1,2,\ldots\}$. *Then* $(p_1,p_2,\ldots)$ *has a* $PD(\theta,a)$ *distribution.*

Table 4.3: Some values of $\Pr\{\widetilde{p}_{i+1} < \widetilde{p}_i\}$.

i	θ	a	$\Pr\{\widetilde{p}_{i+1} < \widetilde{p}_i\}$
1	0.1	1	0.6724
10	0.1	1	0.6073
100	0.1	1	0.5212
1	0.5	1	0.5977
10	0.5	1	0.5263
100	0.5	1	0.5031
1	0.9	1	0.5230
10	0.9	1	0.5044
100	0.9	1	0.5005
1	0.1	10	0.5231
10	0.1	10	0.5212
100	0.1	10	0.5117
1	0.5	10	0.5151
10	0.5	10	0.5106
100	0.5	10	0.5027
1	0.9	10	0.5041
10	0.9	10	0.5023
100	0.9	10	0.5004

It follows form Proposition 4.2.2, that the weights in the two-parameter Poisson-Dirichlet process can be constructed based on two boundary selections of the parameters. The first selection is when $\theta = 0$. This choice of parameters corresponds to the Dirichlet process. The other selection of parameters is when $a = 0$, which yields a measure whose random weights are based on a stable law with index $0 < \theta < 1$. Therefore, the Dirichlet process $P_{H,0,a}$ and the stable law process $P_{H,\theta,0}$ are two essential processes in simulating the two-parameter Poisson-Dirichlet process. First we consider simulating these two key processes.

A simple, yet efficient, procedure for approximating the Dirichlet process was described in Algorithm C in Section 4.1. On the other hand, for the stable law process, Pitman and Yor (1997, Proposition 10) proved that

$$P_{H,\theta,0}(\cdot) = \sum_{i=1}^{\infty} \frac{\Gamma_i^{-1/\theta}}{\sum_{i=1}^{\infty} \Gamma_i^{-1/\theta}} \delta_{\theta_i}(\cdot), \tag{4.2.3}$$

where $\Gamma_i = E_1 + \cdots + E_i$ and $(E_i)_{i\geq 1}$ is a sequence of i.i.d. random variables with an exponential distribution with mean of 1.

The representation (4.2.3) can be used to simulate an approximation of the stable law process using a truncation argument. By truncating the higher order terms in the sum we can approximate the stick breaking representation by

$$P^n_{H,\theta,0}(\cdot) = \sum_{i=1}^{n} \frac{\Gamma_i^{-1/\theta}}{\sum_{i=1}^{n} \Gamma_i^{-1/\theta}} \delta_{\theta_i}(\cdot), \tag{4.2.4}$$

A random stopping rule for choosing $n = n(\epsilon)$, where $\epsilon \in (0, 1)$, is:

$$n = \inf \left\{ j : \frac{\Gamma_j^{-1/\theta}}{\sum_{i=1}^{j} \Gamma_i^{-1/\theta}} < \epsilon \right\}. \tag{4.2.5}$$

It is easy to see that the weights $\left(\Gamma_i^{-1/\theta} / \sum_{i=1}^{n} \Gamma_i^{-1/\theta}\right)_{1\leq i\leq n}$ are strictly decreasing. Thus, simulating the stable law process through the representation (4.2.4) is very efficient. The next algorithm can be used to sample from an approximation of the stable law process.

Algorithm D: Simulating an approximation of the stable law process.

(1) Fix a relatively large positive integer r. One can also apply the random stopping rule (4.2.5).

(2) Generate $\theta_i \overset{\text{i.i.d.}}{\sim} H$ for $i = 1, \ldots, r$.

(3) For $i = 1, \ldots, r+1$, generate E_i from an exponential distribution with mean 1, independent of $(\theta_i)_{1 \le i \le r}$ and let $\Gamma_i = E_1 + \cdots + E_i$.

(4) For each $i = 1, \ldots, r$, the corresponding weights are $\Gamma_i^{-1/\theta} / \sum_{i=1}^{r} \Gamma_i^{-1/\theta}$.

Now we present an efficient algorithm for simulating the two-parameter Poisson-Dirichlet process. This algorithm is based on Proposition 4.2.2, Algorithm C and Algorithm D.

Algorithm E: Simulating an approximation of the two-parameter Poisson-Dirichlet Process.

(1) Use Algorithm C in Section 4.1 to generate n weights of the Dirichlet process. Denote these weights by $(p_1(0,a), \ldots, p_2(0,a))$.

(2) Use Algorithm D to generate r weights for an approximation of the stable law process. Denote these weights by $(p_1(\theta,0), \ldots, p_r(\theta,0))$.

(3) Repeat step (2) to generate n i.i.d. copies of $(p_1(\theta,0), \ldots, p_r(\theta,0))$. Denote these copies by $(p_1^1(\theta,0), \ldots, p_r^1(\theta,0)), \ldots, (p_1^n(\theta,0), \ldots, p_r^n(\theta,0))$.

(4) Find the product $p_i(0,a)p_i^j(\theta,0)$ of the weights generated in step (1) and step (3), where $i = 1, \ldots, n$ and $j = 1, \ldots, r$. That is, find $(p_1(0,a)p_1^1(\theta,0), \ldots, p_1(0,a)p_r^1(\theta,0), \ldots, p_n(0,a)p_1^n(\theta,0), \ldots, p_n(0,a)p_r^n(\theta,0))$.

(5) The weights of the two-parameter Poisson-Dirichlet process are those weights obtained in step (4) written in descending order. Denote these weight by $(p_i)_{1 \le i \le nr}$.

(6) Generate $\theta_i \overset{\text{i.i.d.}}{\sim} H$ for $i = 1, \ldots, nr$.

(7) The approximated two-parameter Poisson-Dirichlet process is given by the representation (4.2.1) with n in the summation replaced by nr.

4.2.1 Empirical Results: A Comparison with the Stick-breaking Approximation

In this section, we compare the new approximation of the two-parameter Poisson-Dirichlet process (Algorithm E) with the stick-breaking approximation given in (4.2.1). In the simulation, we set $n = 100$, $r = 500$ and H to be a uniform distribution on $[0, 1]$. To compare these algorithms we generate 1000 sample paths from the two-parameter Poisson-Dirichlet $P_{H,\theta,a}$ for different values of θ and a by using the two approximations. We compute the sample mean and the standard deviation of the generated process at $x = 0.1, 0.2, \ldots, 0.9, 1.0$. In Table 4.4, we report the absolute maximum of the differences between the sample (approximated) mean and the sample (approximated) standard deviation with the exact values, where the exact values are given by (2.2.2). We refer to these values as the maximum mean error and the maximum standard deviation error. For instance, for $\theta = 0.9$ and $a = 10$, the maximum mean error is 0.0025 and the maximum standard deviation error is 0.0953 in the new approach, while they are 0.0115 and 0.1386 in the stick-breaking approximation. It is clear form the table that both the maximum mean error and the maximum standard deviation error in the new approach are smaller than those obtained by the stick-breaking approximation. Thus, empirically, simulating the two-parameter Poisson-Dirichlet process by using the the new approximation (Algorithm E) gives very accurate results.

Sample paths for the approximate two-parameter Poisson-Dirichlet process with a uniform distribution on [0,1] as a base measure with different concentration and discount parameters are given in Figures 4.5 and 4.6. Distinctly, the new approx-

imation performs very well in all cases. On the other hand, a clear disadvantage of the stick-breaking approximation appears when θ is close to 1 ($\theta < 1$). In this case, as seen in Figure 4.6, contrary to our anticipation (see inequality (2.2.4)), the two-parameter Poisson-Dirichlet process is not in the proximity of the base measure. Thus, the stick-breaking representation performs very poorly when θ is close to 1 ($\theta < 1$).

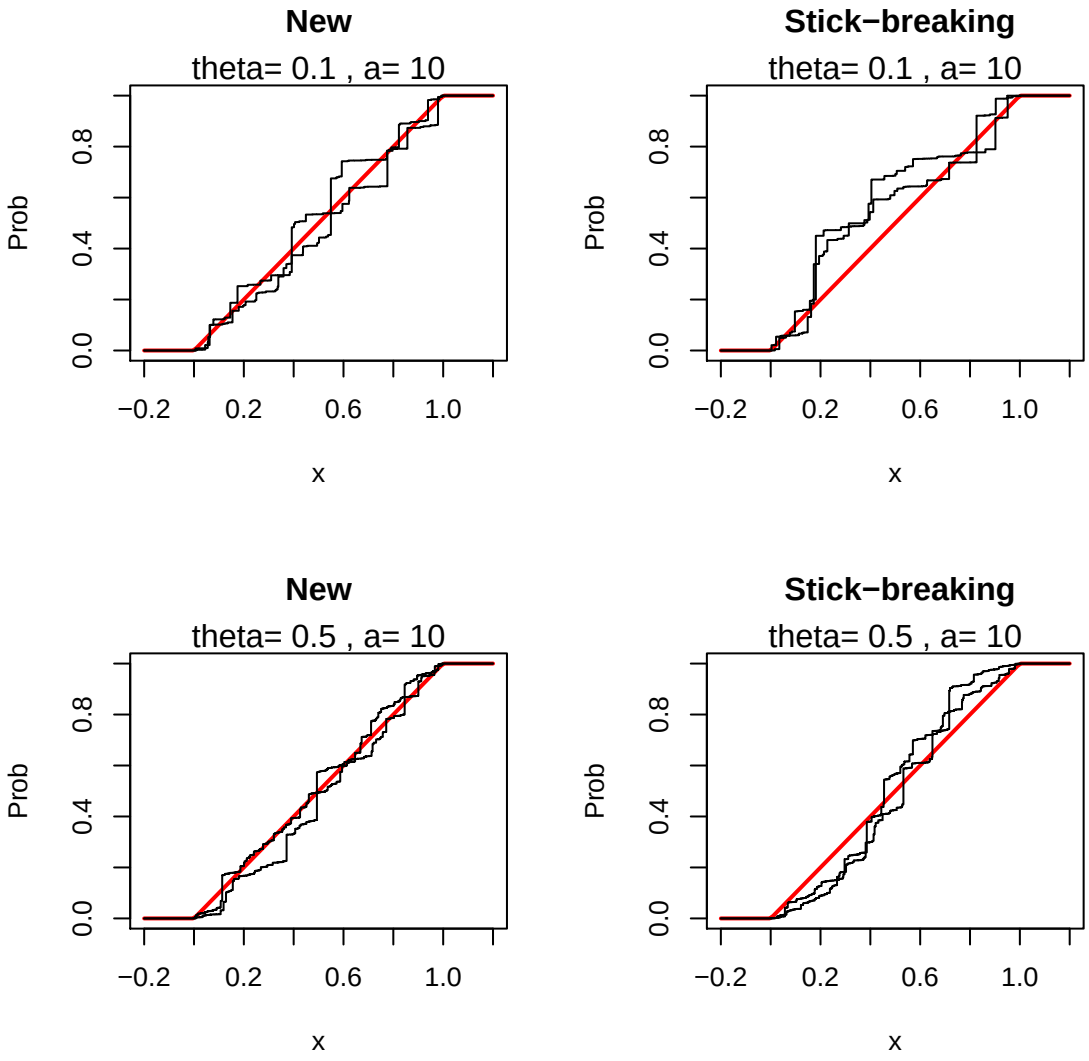

Figure 4.5: Sample paths of the two-parameter Poisson-Dirichlet process $P_{H,\theta,a}$, where H is the uniform distribution on $[0,1]$, $a = 10$ and $\theta = 0.1, 0.5$. The solid line denotes the cumulative distribution function of H.

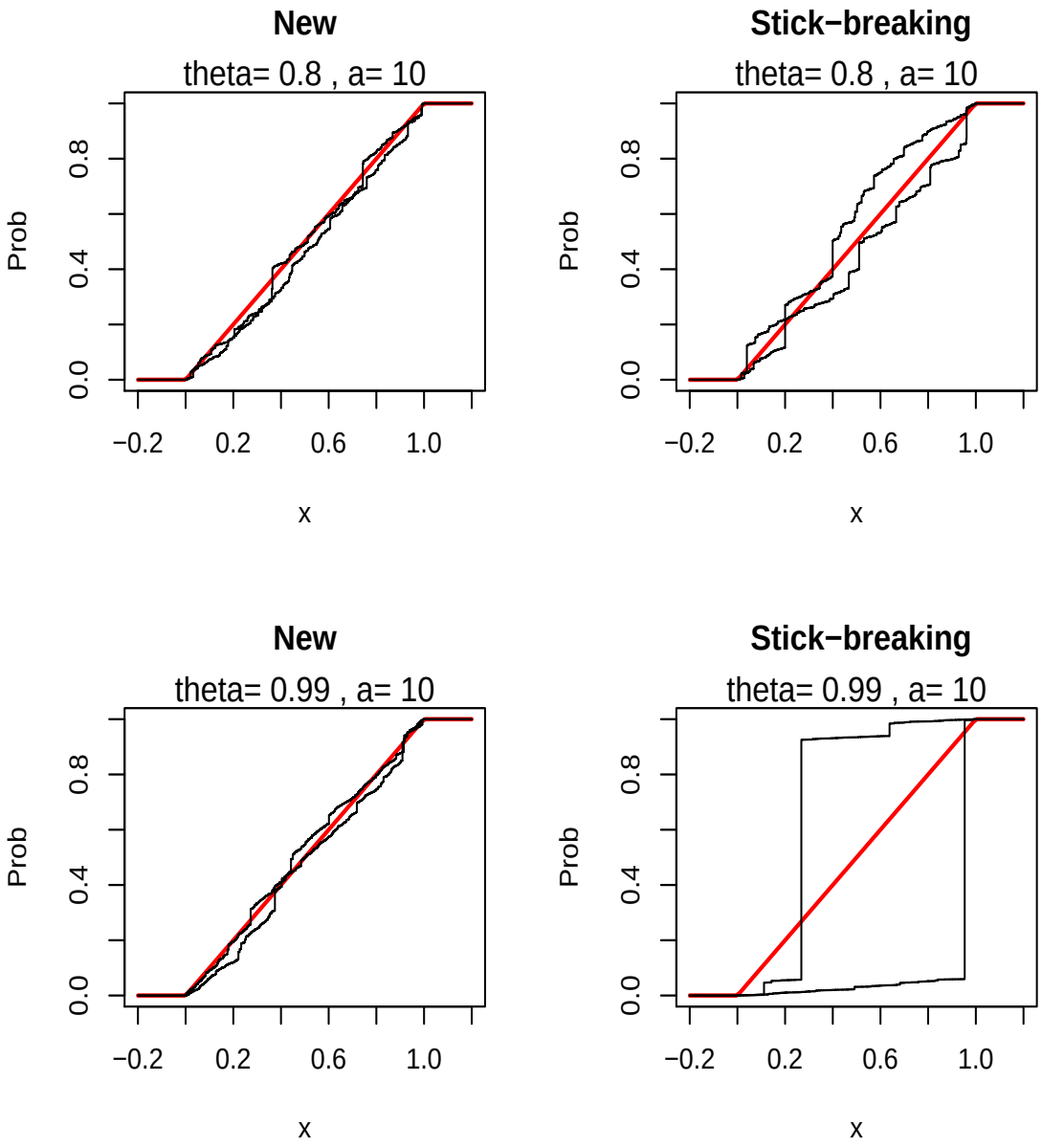

Figure 4.6: Sample paths of a two-parameter Poisson-Dirichlet process $P_{H,\theta,a}$, where H is the uniform distribution on $[0,1]$, $a = 10$ and $\theta = 0.8, 0.99$. The solid line denotes the cumulative distribution function of H.

Table 4.4: This table reports the maximum mean (max. mean) error and maximum standard deviation (max. sd.) error. That is, the absolute maximum difference between the approximated mean (standard deviation) and the actual mean (standard deviation) of $P_{H,\theta,a}(x) = P_{H,\theta,a}((-\infty, x])$ evaluated at $x = 0.1, 0.2, \ldots, 0.9, 1.0$, where H is a uniform distribution on $[0, 1]$.

		New		Stick-breaking	
θ	a	max. mean error	max. sd error	max. mean error	max. sd error
0.1	1	0.0115	0.6708	0.0233	0.6708
0.5	1	0.0098	0.5	0.01334	0.5
0.9	1	0.0056	0.2236	0.0130	0.2236
0.1	10	0.0099	0.2860	0.0110	0.2860
0.5	10	0.0037	0.2132	0.0045	0.2132
0.9	10	0.0025	0.0953	0.0115	0.1386
0.1	50	0.0024	0.1328	0.0034	0.1328
0.5	50	0.0013	0.0990	0.0023	0.0990
0.9	50	0.0009	0.04428	0.0097	0.0547

4.3 A New Algorithm to Generate the Beta Process

The new technique used for simulating the Dirichlet process in Section 4.1 can be extended to the beta process. This is the third contribution of this chapter. This section contains three subsections. In the first subsection, we describe the new method for simulating the beta process. In the second subsection, we review several well-known algorithms which can be adapted to sample from the beta process. An extensive simulation study evaluating the accuracy of our new algorithm and an extensive comparison with several other approaches is presented in the third subsection.

4.3.1 The New Algorithm

As it was mentioned in Chapter 3, the simulation of the beta process based on a series representation is very complex and may be difficult to apply in practice. For example, the Wolpert and Ickstadt representation (Section 3.2), which is computationally simpler than the Ferguson and Klass representation (Section 3.1), requires finding $M_{\theta_i}^{-1}(\Gamma_i)$ at each step, where $M_{\theta_i}(\Gamma_i)$ is defined in (3.2.3). Since $M_{\theta_i}(x)$ is not available as a standard mathematical function, finding $M_{\theta_i}^{-1}(\Gamma_i)$ at each step is not straightforward and is time consuming. To overcome this problem, we derive, in this subsection, a finite sum approximation which converges almost surely to the Wolpert and Ickstadt representation for the beta process.

Let $A \sim BP(c(\cdot), A_0(\cdot))$ be as in Definition 2.5.1 with a continuous A_0 (i.e $A_{0,c} = A_0$ and $A_c = A$). We restrict the domain of t to be the interval $[0, T]$, where $T > 0$ is fixed. We assume that $A_0(T) < \infty$. For any real number θ, define

$$L_{n,\theta}(x) = \frac{\Gamma(c(\theta))}{\Gamma(c(\theta)/n)\Gamma(c(\theta) - c(\theta)/n)} \int_x^1 s^{c(\theta)/n-1} (1-s)^{c(\theta)(1-1/n)-1} ds. \quad (4.3.1)$$

The following proposition summarizes some properties of $L_{n,\theta}(x)$.

Proposition 4.3.1 *For any real number θ, the function $L_{n,\theta}(x)$ defined in (4.3.1) has the following properties as $n \to \infty$:*

(i) For any $x \in (0,1)$, $nA_0(T)L_{n,\theta}(x) \to M_\theta(x)$, where

$$M_z(x) = A_0(T)c(z)\int_x^1 s^{-1}(1-s)^{c(z)-1}ds.$$

(ii) For any $y > 0$, $L_{n,\theta}^{-1}\left(y/(nA_0(T))\right) \to M_\theta^{-1}(y)$.

Proof: To prove (i), observe that, by (4.1.5),

$$\frac{n}{\Gamma\left(c(\theta)/n\right)} \to c(\theta).$$

Since $\Gamma(x)$ is a continuous function

$$\frac{\Gamma\left(c(\theta)\right)}{\Gamma\left(c(\theta) - c(\theta)/n\right)} \to 1.$$

Clearly, the integrand in the right hand side of (4.3.1) converges to $s^{-1}(1-s)^{c(\theta)-1}$. To apply the dominated convergence theorem, we need to show that this integrand is dominated by an integrable function. Since $x < s < 1$, we have $s^{-1} < x^{-1}$ and $s^{c(\theta)/n} < 1$. This implies that $s^{c(\theta)/n-1} < x^{-1}$. Therefore, the integrand is bounded above by the integrable function $x^{-1}\left(1-s\right)^{c(\theta)(1-1/n)-1}$. Thus, by the dominated convergence theorem, (i) follows. The proof of (ii) is similar to the proof of Proposition 4.1.1. ∎

The previous proposition gives a simple procedure to approximate both $M_\theta(x)$ and $M_\theta^{-1}(x)$. For computational simplicity, another convenient approximation is presented in the next corollary. Its proof follows straightforwardly by taking $y = \Gamma_i$ in Proposition 4.3.1 and using the fact that $\Gamma_{n+1}/n \stackrel{a.s.}{\to} 1$ (by the strong law of large numbers).

Corollary 4.3.2 *For a fixed $i \geq 1$, we have*

$$L_{n,\theta_i}^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}A_0(T)}\right) \stackrel{a.s.}{\to} M_{\theta_i}^{-1}(\Gamma_i),$$

as $n \to \infty$.

The next theorem gives a finite sum representation which converges almost surely to the series representation of the beta process given by (3.2.5). Convergence of random measures is considered with respect to the vague topology on the space of point measures on $[0, T]$. See Appendix B for background on convergence of random measures. The proof is similar to the proof of Lemma 4.1.3. We include the proof for the sake of completeness.

Theorem 4.3.3 *Let $(\theta_i)_{i\geq 1}$ be i.i.d. random variables with common distribution Π and $\Gamma_i = E_1 + \cdots + E_i$, where $(E_i)_{i\geq 1}$ are i.i.d. with exponential distribution with mean 1, independent of $(\theta_i)_{i\geq 1}$. Let $A \sim BP(c(\cdot), A_0(\cdot))$ on $[0, T]$, where $T > 0$ is fixed. We assume that A_0 is continuous with $A_0(T) < \infty$. Let $\Pi(dz) = \eta(dz)/A_0(T)$, where $\eta([0, t]) = A_0(t)$. Then as $n \to \infty$,*

$$A_n(t) = \sum_{i=1}^{n} L_{n,\theta_i}^{-1}\left(\frac{\Gamma_i}{A_0(T)\Gamma_{n+1}}\right)\delta_{\theta_i} \stackrel{a.s.}{\to} A(t) = \sum_{i=1}^{\infty} M_{\theta_i}^{-1}(\Gamma_i)\,\delta_{\theta_i}.$$

Proof: We will show that

$$\sum_{i=1}^{n} L_{n,\theta_i}^{-1}\left(\frac{\Gamma_i}{A_0(T)\Gamma_{n+1}}\right)\delta_{\theta_{(i)}} \stackrel{a.s.}{\to} \sum_{i=1}^{\infty} M_{\theta_i}^{-1}(\Gamma_i)\,\delta_{\theta_{(i)}},$$

as $n \to \infty$. By Lemma 2 of Grandell (1977) (also see the proof of Theorem 3), it is enough to show that, for all k fixed,

$$\sum_{i=1}^{k} L_{n,\theta_i}^{-1}\left(\frac{\Gamma_i}{A_0(T)\Gamma_{n+1}}\right)\delta_{\theta_{(i)}} \stackrel{a.s.}{\to} \sum_{i=1}^{k} M_{\theta_i}^{-1}(\Gamma_i)\,\delta_{\theta_{(i)}},$$

as $n \to \infty$. This is clear since $L_{n,\theta_i}^{-1}\left(\Gamma_i/\left(A_0(T)\Gamma_{n+1}\right)\right) \stackrel{a.s.}{\to} M_{\theta_i}^{-1}(\Gamma_i)$. ∎

The next algorithm is based on Theorem 4.4.3 and is used to generate samples from an approximation of the beta process with parameters $c(\cdot)$ and A_0 on $[0, T]$ with A_0 continuous.

Algorithm F: New algorithm for simulating an approximation for the prior beta process. Since the prior beta process has no fixed point of discontinuity, it is characterized entirely by its Lévy measure.

(1) Fix n large enough.

(2) Generate $\theta_i \overset{\text{i.i.d.}}{\sim} dA_0/A_0(T)$, for $i = 1, \ldots, n$.

(3) For $i = 1, \ldots, n+1$, generate E_i from an exponential distribution with mean 1, independent of $(\theta_i)_{1 \leq i \leq n}$ and let $\Gamma_i = E_1 + \cdots + E_i$.

(4) For $i = 1, \ldots, n$, compute $\left(L^{-1}_{n,\theta_i}\left(\Gamma_i/(A_0(T)\Gamma_{n+1})\right)\right)_{1 \leq i \leq n}$, which is simply the quantile function of the $Beta\left(c(\theta_i)/n, c(\theta_i)(1-1/n)\right)$ distribution evaluated at $1 - \Gamma_i/\left(A_0(T)\Gamma_{n+1}\right)$.

(5) Set $A_n(t)$ as in Theorem 4.4.3.

Figure 4.7 shows sample paths generated for the beta process $\{A(t)\}_{t \in [0,1]}$ with parameters $c(t) = 2$ and $A_0(t) = t$ by using the Wolpert-Ickstadt algorithm (solid line) and the new algorithm (dashed line). This figure clearly shows that the new representation converges to the Wolpert and Ickstadt representation of the beta process. Simulating the Wolpert and Ickstadt representation is performed through relatively complex numerical methods, which is very time consuming.

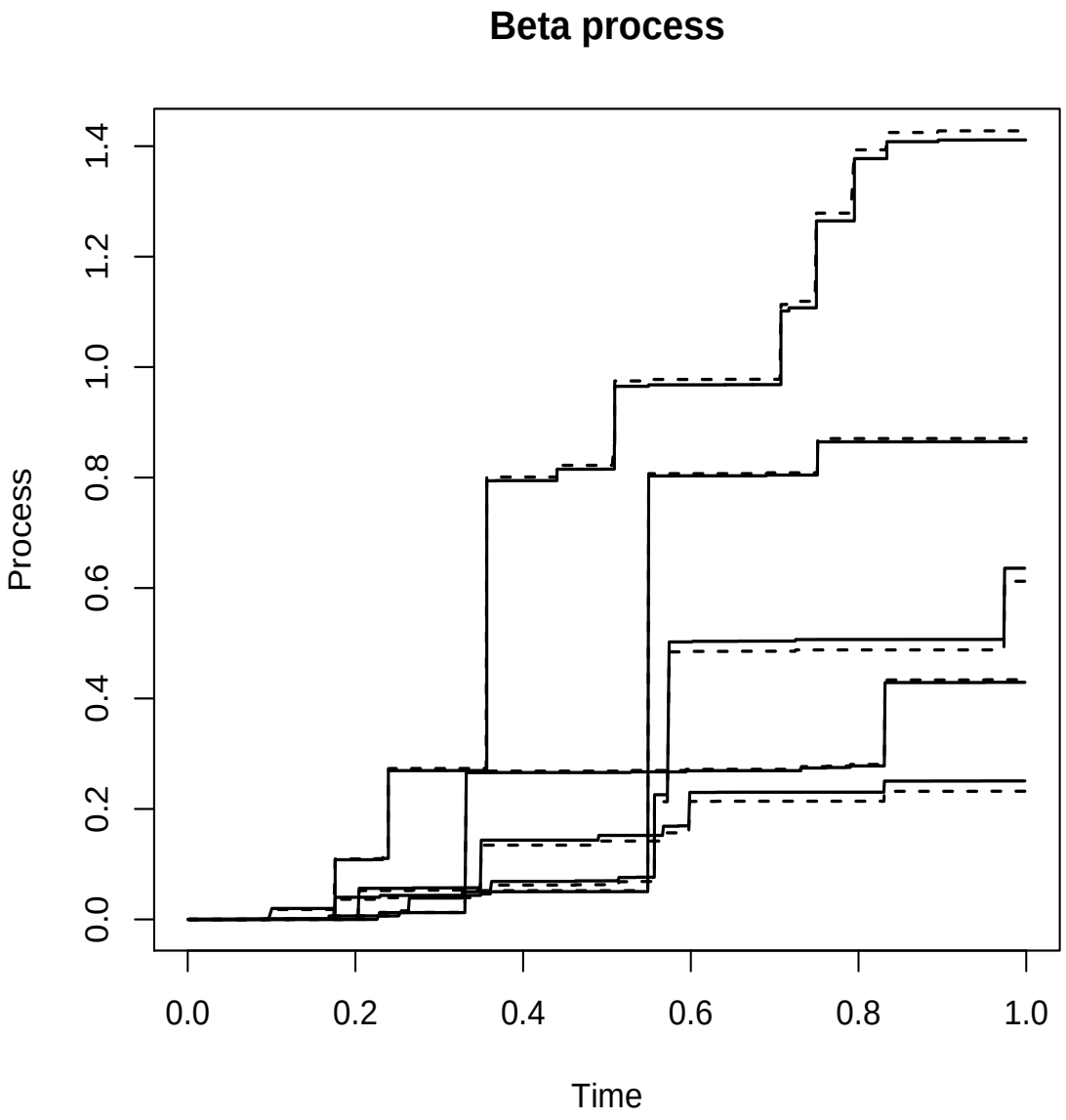

Figure 4.7: *Sample paths of a beta process with* $c(t) = 2$, *and* $A_0(t) = t$, *where* $t \in [0, 1]$. *The dashed lines denote the sample paths generated by the new algorithm with* $n = 1000$.

To sample from an approximation for the posterior beta process, given the right censored data $(T_1, \delta_1), \ldots, (T_m, \delta_m)$, one can use the following algorithm.

Algorithm G: New algorithm for simulating an approximation for the posterior beta process:

(1) To sample from the posterior process $\{A_c^*(t)\}_{t\in[0,T]}$, which is a pure jump process with no fixed point of discontinuity, use Algorithm F (for the prior beta process) with $\beta(\theta_i)$ replaced by $\beta(\theta_i) + Y(\theta_i)$.

(2) To generate the random jumps $(S_i^*)_{1\leq i\leq l}$ that are associated with the fixed points of discontinuity, apply Algorithm B in Section 3.3.

(3) Putting the simulation for the two components together results in a sample from the approximated posterior beta process.

4.3.2 Other Sampling Algorithms

In the literature, there exist several algorithms to sample from the beta process. In addition to the Ferguson and Klass (1972) algorithm and the Wolpert and Ickstadt (1998) algorithm, we consider the Damien, Laud, and Smith (1995) algorithm, the Walker and Damien (1998) algorithm, the Lee and Kim (2004) algorithm, and the Lee (2007) algorithm. Some of these algorithms are so general that they can be applied to any neutral to the right process. Some of them are specialized to certain processes. Below is a discussion of these algorithms.

• **Damien-Laud-Smith Algorithm:** Using the fact that the distributions of the increments of a nondecreasing Lévy process are infinitely divisible, Damien, Laud, and Smith (1995) derived an algorithm to generate approximations for infinitely divisible random variables and used it to generate a pure jump process $\{Z(t)\}_{t\in[0,T]}$. The Damien-Laud-Smith algorithm (in its general form) is described as follows. First, the

time interval $[0, T]$ is partitioned into small subintervals with endpoints $0 = \theta_0 < \theta_1 < \ldots < \theta_m = T$. Let J_i denotes the increment of $\{Z(t)\}_{t\in[0,T]}$ in the interval $\Delta_i = (\theta_{i-1}, \theta_i]$, i.e. $J_i = Z(\theta_i) - Z(\theta_{i-1})$. The random variable J_i follows an infinitely divisible distribution with the log-Laplace transform given by:

$$\log Ee^{-uJ_i} = \int_0^\infty \left(e^{-us} - 1\right) L_{\Delta_i}(ds),$$

where

$$L_{\Delta_i}(ds) = \left[\int_{\theta_{i-1}}^{\theta_i} K(s, z)\eta(dz)\right] ds. \tag{4.3.2}$$

See also Section 3.2. Define

$$dG_{\Delta_i}(s) = \frac{s}{s+1} L_{\Delta_i}(ds) = \frac{s}{s+1}\left[\int_{\theta_{i-1}}^{\theta_i} K(s, z)\eta(dz)\right] ds$$

and

$$\lambda_i = \int_0^\infty dG_{\Delta_i}(s) = \int_0^\infty \int_{\theta_{i-1}}^{\theta_i} \frac{s}{1+s} K(s, z)\eta(dz)ds.$$

The steps of the Damien-Laud-Smith algorithm for simulating an approximation for the jump J_i are:

(1) Fix a relatively large positive integer n.

(2) Generate independent random values x_{ij} from the probability density function $dG_{\Delta_i}(s)/\lambda_i$, for $j = 1, \ldots, n$.

(3) Generate y_{ij}: $y_{ij}|x_{ij} \sim Poisson\left(\lambda_i(1 + x_{ij})/nx_{ij}\right)$, for $j = 1, \ldots, n$.

(4) Set $J_{i,n} = \sum_{j=1}^n x_{ij}y_{ij}$.

Damien, Laud, and Smith (1995) showed that $J_{i,n} \xrightarrow{d} J_i$, as $n \to \infty$. That is, $J_{i,n}$ is an approximate sample from the ith increment of Z. Note that, the Damien-Laud-Smith algorithm generates only the increments of the process and not the entire process. To obtain the whole process, we set

$$Z_{m,n}(t) = \sum_{i=1}^m J_{i,n} I(\theta_i \le t).$$

For large m and n, $Z_{m,n}(t)$ is an approximation of $Z(t)$.

In general, one can replace $s/(1+s)$ in the algorithm by any nonnegative function $h(s)$ such that

$$\int_0^\infty h(s) L_{\Delta_i}(ds) < \infty$$

(De Blasi, 2007). By the regularity condition (iv) in Section 2.3, it is sufficient that $h(s) \leq s/(1+s)$. Equivalently, one can choose $h(s)$ such that $h(s) \leq \min(1, s)$. In particular, for the beta process with parameters $c(\cdot)$ and A_0 with A_0 continuous, Damien, Laud and Smith (1996) took $h(s) = s$ and suggested the following algorithm to generate an approximation for the jump J_i:

(1) Fix a relatively large positive integer n.

(2) Generate independent values z_{ij} from the probability density function $dA_0(t)/A_0(\Delta_i)$, for $j = 1, \ldots, n$.

(3) Generate $x_{ij} \sim beta(1, c(z_{ij}))$, for $j = 1, \ldots, n$.

(4) Generate y_{ij}: $y_{ij}|x_{ij} \sim Poisson(\lambda_i n^{-1} x_{ij}^{-1})$, for $j = 1, \ldots, n$.

(5) Set $J_{i,n} = \sum_{j=1}^n x_{ij} y_{ij}$. For large n, $J_{i,n}$ is an approximation of J_i.

• **Walker-Damien Algorithm:** This algorithm can be described as follows. Let $\Delta = [a, b) \subseteq [0, T]$ and $L_\Delta(ds)$ be the Lévy measure of the beta process restricted to Δ as defined by (4.3.2). Define

$$G_\epsilon(s) = L_\Delta((\epsilon, s]) = \int_\epsilon^s \int_a^b c(z) u^{-1} (1-u)^{c(z)-1} dA_0(z) du$$

and

$$\lambda_\epsilon = L_\Delta((\epsilon, 1)) = \int_\epsilon^1 \int_a^b c(z) u^{-1} (1-u)^{c(z)-1} dA_0(z) du.$$

The steps of the Walker-Damien algorithm for the beta process A with parameters $c(\cdot)$ and A_0 with A_0 continuous are:

(1) Fix a relatively small positive number ϵ.

(2) Generate the total number of jumps $n \sim Poisson\,(\lambda_\epsilon)$.

(3) Generate the jump sizes $J_1, \ldots, J_n \sim G_\epsilon/\lambda_\epsilon$.

(4) $A_\epsilon(\Delta) = \sum_{i=1}^{n} J_i$. For for a small ϵ, $A_\epsilon(\Delta)$ is an approximate of $A(\Delta)$.

• **Lee-Kim Algorithm:** Lee and Kim (2004) derived an approximate sampling algorithm specialized for the gamma process and the beta process. The Kim and Lee algorithm for the beta process with parameters $c(\cdot)$ and A_0 with A_0 continuous can be described as follows. First the Lévy measure L_t of the beta process given by (2.5.3) is approximated by:

$$L_{t,\epsilon}(ds) = \left[\int_0^t \frac{c(s)}{\epsilon} b(s : \epsilon, c(z)) dA_0(z)\right] ds,$$

where

$$b(x : a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad \text{for } 0 < x < 1, a > 0, b > 0. \tag{4.3.3}$$

Observe that, unlike L_t, $L_{t,\epsilon}$ is a finite measure. Let

$$\lambda_\epsilon = L_{T,\epsilon}\,((0,1)) = \int_0^T \int_0^1 \frac{c(z)}{\epsilon} b(s : \epsilon, c(z)) ds dA_0(z) = \frac{1}{\epsilon}\int_0^T c(z) dA_0(z)$$

and

$$dG_\epsilon(z) = \frac{c(z)}{\epsilon} dA_0(z) I(0 \le z \le T).$$

The steps of the Lee-Kim algorithm for the beta process A are:

(1) Fix a relatively small positive number ϵ.

(2) Generate the total number of jumps $n \sim Poisson\,(\lambda_\epsilon)$.

(3) Generate the jump times $\theta_1, \ldots, \theta_n$ form the probability density function $dG_\epsilon/\lambda_\epsilon$.

(4) Let $\theta_{(1)} \leq \ldots \leq \theta_{(n)}$ be the corresponding order statistics of $\theta_1, \ldots, \theta_n$.

(5) Generate the jump sizes $J_1, \ldots, J_n : J_i|\theta_{(i)} \sim Beta(\epsilon, c(\theta_{(i)}))$.

(6) Set $A_\epsilon(t) = \sum_{=1}^{n} J_i I(\theta_{(i)} \leq t)$.

Lee and Kim (2004) showed that $A_\epsilon(t) \xrightarrow{d} A(t)$, as $\epsilon \to 0$, on $D[0,T]$ with respect to the Skorohod topology, where $\{A(t)\}_{t\in[0,T]}$ is the beta process with parameters $c(\cdot)$ and A_0 with A_0 continuous.

• **Lee Algorithm:** This is a general algorithm to generate a pure jump process $\{Z(t)\}_{t\in[0,T]}$ with a Lévy measure taking the form (3.2.1). For each z, let $K'(s,z)$ be a probability density from which a random number can be easily drawn and whose support contains that of $K(s,z)$. The steps of the Lee algorithm are as follows:

(1) Fix a relatively large positive integer n.

(2) Generate i.i.d. random variables $\theta_1, \ldots, \theta_n$ from the probability density function $\Pi(dz) = \eta(dz)/\eta([0,T])$.

(3) For $i = 1, \ldots, n$, generate $x_i \sim K'(s, \theta_i)$.

(4) For $i = 1, \ldots, n$, set
$$\lambda_i = \frac{\eta([0,T])K(x_i,\theta_i)}{nK'(x_i,\theta_i)}.$$

(5) For $i = 1, \ldots, n$, generate $y_i \sim Poisson(\lambda_i)$.

(6) Set $Z_n(t) = \sum_{i=1}^{n} x_i y_i I(\theta_i \leq t)$.

Lee (2007) proved that, as $n \to \infty$, $Z_n(t) \xrightarrow{d} Z(t)$ on $D[0,T]$ with respect to the Skorohod topology, where $\{Z(t)\}_{t\in[0,T]}$ is a pure jump process. Lee (2007) implemented this algorithm to the gamma process and the beta process. For the beta process, he suggested to take

$$K'(s,z) = b(s : \epsilon, c(z)),$$

where $b(s:\epsilon, c(z))$ is defined in (4.3.3). Note that, for the beta process, $K(s,z) = c(z)s^{-1}(1-s)^{c(z)-1} = s^{-1}b(s:1,c(z))$. Thus, in the Lee algorithm for the beta process with parameters $c(\cdot)$ and A_0 with A_0 continuous, one can take

$$\lambda_i = \lambda_{i,\epsilon} = \frac{A_0(T)b(x_i:1,c(\theta_i))}{nx_i b(x_i:\epsilon,c(\theta_i))}.$$

4.3.3 Empirical Results: Comparison with Other Methods

In this section, we compare the new approximation of the beta process derived in Subsection 4.3.1 with the algorithms mentioned in the previous subsection. We consider two beta processes: $A_1(t) \sim BP(c(t)=2, A_0(t)=t)$ and $A_2(t) \sim BP(c(t)=2e^{-t}, A_0(t)=t)$, where $t \in [0,1]$. Observe that A_1 is a homogenous process (since $c(t)$ is independent of t) whereas A_2 is nonhomogeneous (since $c(t)$ depends on t).

To compare these algorithms we generate 3000 sample paths from each process. We compute the sample mean and the standard deviation of the generated process at $t = 0.1, 0.2, \ldots, 0.9, 1.0$. We obtain the absolute maximum of the differences between the sample mean and the sample standard deviation with the exact values. The exact values are given by (2.5.5). Observe that each algorithm is characterized by its own parameters. For example, the Damien-Laud-Smith algorithm is characterized by two parameters: the number m of subintervals and the number n of Poisson and beta random variables. In order to make comparisons between these different algorithms, we give equivalent setting for the parameters characterizing these processes. For instance, in the Kim-Lee algorithm, we set ϵ so that total number of jumps is equal to n. For the Lee algorithm and the new algorithm we choose the same n. We point out that in the Lee algorithm we select $\epsilon = 0.05$. A smaller value of ϵ gives undefined values in this algorithm. The results are summarized in Table 4.5 and Table 4.6. The tables show a perfect performance of the new approximation compared with other algorithms.

Table 4.5: This table reports the maximum mean (max. mean) error and maximum standard deviation (max. sd) error. That is, the absolute maximum difference between the approximated mean (standard deviation) and the actual mean (standard deviation) of $A_1(t)$ evaluated at $x = 0.1, 0.2, \ldots, 0.9, 1.0$, where $A_1(t) \sim BP(c(t) = 2, A_0(t) = t)$. DSL and KL stands respectively for the Damien, Laud, and Smith algorithm (1995) and the Lee and Kim algorithm (2004).

Algorithm	Parameters	max. mean error	max. sd error
DSL	$m = n = 200$	0.0110	0.0156
LK	$\epsilon = 0.01$	0.0128	0.0107
Lee	$n = 200, \epsilon = 0.05$	0.0162	0.0392
New	$n = 200$	0.0087	0.0061

Table 4.6: This table reports the maximum mean (max. mean) error and maximum standard deviation (max. sd) error. That is, the absolute maximum difference between the approximated mean (standard deviation) and the actual mean (standard deviation) of $A_1(t)$ evaluated at $x = 0.1, 0.2, \ldots, 0.9, 1.0$, where $A_2(t) \sim BP(c(t) = 2e^{-t}, A_0(t) = t)$. DSL and KL stands respectively for the Damien, Laud, and Smith algorithm (1995) and the Lee and Kim algorithm (2004).

Algorithm	Parameters	max. mean error	max. sd error
DSL	$m = n = 200$	0.0121	0.0110
LK	$\epsilon = 0.01$	0.0165	0.0175
Lee	$n = 200, \epsilon = 0.05$	0.0217	0.0420
New	$n = 200$	0.0077	0.0007

Remark 4.3.4 *In this example, applying Walker-Damien algorithm requires sampling from nonstandard distributions. Hence, this algorithm is not considered in this section. To overcome sampling from nonstandard distributions, Walker and Damien (1998) suggested a Gibbs/Metropolis sampling algorithm.*

4.4 A New Algorithm to Generate the Beta-Stacy Process

The approach used for simulating the Dirichlet and the beta processes can be extended to include the beta-Stacy process. This will be the fourth contribution of this Chapter. This section contains three subsections. In the first subsection, we describe a new method to simulate the beta-Stacy process. In the second subsection, we review several well-known algorithms which can be adapted to sample from the beta-Stacy process. In particular, we modify the Lee and Kim (2004) algorithm and the Lee (2007) algorithm to sample from the beta-Stacy process. This adds another contribution of this chapter. In the third subsection, we compare the new algorithm with several other approaches through survival data.

4.4.1 The New Algorithm

It follows from Definition 2.6.2 that simulating the beta-Stacy process can be accomplished through simulating the log-beta process. Let $\{Z(t)\}_{t\in[0,T]}$ be a log-beta process with parameters $\beta(\cdot)$ and α, where $\beta(\cdot)$ is a positive function on $[0,T]$ and α is a measure concentrated on $[0,T]$ which is absolutely continuous with respect to the Lebesgue measure such that $\int_0^T d\alpha(z)/\beta(z) = \infty$ (see Definition 2.6.1). The Lévy measure (2.6.1) of the log-beta process can be rewritten as follows. Using the

transformation $y = e^{-s}$, it follows that:

$$\begin{aligned} L_t(x) = L_t([x,\infty)) &= \int_0^t \int_x^\infty \frac{1}{1-e^{-s}} e^{-s\beta(z)} ds\alpha(dz) \\ &= \int_0^t \int_0^{e^{-x}} y^{\beta(z)-1}(1-y)^{-1} dy\alpha(dz) \\ &= \int_0^t \int_0^{e^{-x}} s^{\beta(z)-1}(1-s)^{-1} ds\alpha(dz), \quad x > 0. \end{aligned}$$

For any real number θ, define

$$S_{n,\theta}(x) = \frac{\Gamma(\beta(\theta))}{\Gamma(\beta(\theta)/n)\Gamma(\beta(\theta) - \beta(\theta)/n)} \int_0^{e^{-x}} s^{\beta(\theta)(1-1/n)-1} (1-s)^{\beta(\theta)/n-1} ds. \quad (4.4.1)$$

The following proposition summarizes some properties of $S_{n,\theta}(x)$.

Proposition 4.4.1 *For any real number θ, the function $S_{n,\theta}(x)$ defined in (4.4.1) has the following properties:*

(i) For any $x > 0$,

$$\frac{n\alpha([0,T])S_{n,\theta}(x)}{\beta(\theta)} \to M_\theta(x),$$

where

$$M_z(x) = \alpha([0,T]) \int_0^{e^{-x}} s^{\beta(z)-1}(1-s)^{-1} ds.$$

(ii) For any $y > 0$,

$$S_{n,\theta}^{-1}\left(\frac{y\beta(\theta)}{n\alpha([0,T])}\right) \to M_\theta^{-1}(y), \text{ as } n \to \infty,$$

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

Proof: To prove (i), observe that for any $x > 0$,

$$\Gamma(x) = \frac{\Gamma(x+1)}{x}. \quad (4.4.2)$$

With $x = \beta(\theta)/n$ in (4.4.2) we obtain:

$$\frac{n}{\Gamma(\beta(\theta)/n)} = \frac{\beta(\theta)}{\Gamma(\beta(\theta)/n+1)}. \quad (4.4.3)$$

Note that, the right hand of (4.4.3) converges to $\beta(\theta)$ as $n \to \infty$, since Γ is a continuous function and $\Gamma(1) = 1$. Hence,

$$\frac{n}{\Gamma\left(\beta(\theta)/n\right)} \to \beta(\theta).$$

Note also that

$$\frac{\Gamma\left(\beta(\theta)\right)}{\Gamma\left(\beta(\theta) - \beta(\theta)/n\right)} \to 1.$$

Clearly, the integrand in the right hand side of (4.4.1) converges to $s^{\beta(\theta)-1}(1-s)^{-1}$. To apply the dominated convergence theorem we need to show that this integrand is dominated by an integrable function. Since $0 < s < e^{-x}$, we have $(1-s)^{-1} < (1-e^{-x})^{-1}$. Also since $0 < 1 - s < 1$, we have $(1-s)^{\beta(\theta)/n} < 1$. This implies that $0 < (1-s)^{\beta(\theta)/n-1} < (1-e^{-x})^{-1}$. Therefore, the integrand is bounded above by $(1-e^{-x})^{-1} s^{\beta(\theta)(1-1/n)-1}$, which is integrable on $(0, e^{-x})$. Thus, by the dominated convergence theorem, (i) follows.

The proof of (ii) is similar to the proof of Proposition 4.1.1. ∎

The previous proposition gives a simple procedure to approximate both $M_\theta(x)$ and $M_\theta^{-1}(x)$. Another convenient approximation is offered in the following Corollary. The proof follows straightforwardly by taking $y = \Gamma_i$ in Proposition 4.4.1 (ii) and using the fact that $\Gamma_{n+1}/n \stackrel{a.s.}{\to} 1$ (by the strong law of large numbers).

Corollary 4.4.2 *For a fixed i, we have*

$$S_{n,\theta}^{-1}\left(\frac{\Gamma_i \beta(\theta)}{\Gamma_{n+1}\alpha([0,T])}\right) \stackrel{a.s.}{\to} M_\theta^{-1}(\Gamma_i),$$

as $n \to \infty$.

The next theorem gives a finite sum representation which converges almost surely to the Wolpert and Ickstadt representation for the log-beta process; see representation (3.2.5). Convergence of random measures is taken with respect to the vague

topology on the space of point measures on $[0,T]$. See Appendix B for background on convergence of random measures. The proof is similar to the proof of Lemma 4.1.3 and Theorem 4.4.3. We include the proof for the sake of completeness.

Theorem 4.4.3 *Let $(\theta_i)_{i\geq 1}$ be i.i.d. random variables with common distribution Π and $\Gamma_i = E_1 + \cdots + E_i$, where $(E_i)_i$ are i.i.d. with exponential distribution with mean 1, independent of $(\theta_i)_i$. Let α be a measure concentrated on $[0,T]$ and $\beta(\cdot)$ be positive function on $[0,T]$, where $T > 0$ is fixed. Let $\Pi(dz) = [\alpha([0,T])]^{-1}\,\alpha(dz)$. Then as $n \to \infty$,*

$$Z_n = \sum_{i=1}^{n} S_{n,\theta_i}^{-1}\left(\frac{\Gamma_i \beta(\theta_i)}{\alpha([0,T])\Gamma_{n+1}}\right)\delta_{\theta_i} \overset{a.s.}{\to} Z = \sum_{i=1}^{\infty} M_{\theta_i}^{-1}(\Gamma_i)\,\delta_{\theta_i}.$$

Proof: We will show that

$$\sum_{i=1}^{n} S_{n,\theta_{(i)}}^{-1}\left(\frac{\Gamma_i \beta(\theta_{(i)})}{\alpha([0,T])\Gamma_{n+1}}\right)\delta_{\theta_{(i)}} \overset{a.s.}{\to} \sum_{i=1}^{\infty} M_{\theta_{(i)}}^{-1}(\Gamma_i)\,\delta_{\theta_{(i)}},$$

as $n \to \infty$, where $\theta_{(1)} \leq \cdots \leq \theta_{(n)}$ represent the corresponding order statistics of $\theta_1, \ldots, \theta_n$. By Lemma 2 of Grandell (1977) (also see the proof of Theorem 3), it is enough to show that, for all k fixed,

$$\sum_{i=1}^{k} S_{n,\theta_{(i)}}^{-1}\left(\frac{\Gamma_i \beta(\theta_{(i)})}{\alpha([0,T])\Gamma_{n+1}}\right)\delta_{\theta_{(i)}} \overset{a.s.}{\to} \sum_{i=1}^{k} M_{\theta_{(i)}}^{-1}(\Gamma_i)\,\delta_{\theta_{(i)}},$$

$n \to \infty$. This is clear since $S_{n,\theta_{(i)}}^{-1}\left(\Gamma_i\beta(\theta_{(i)})/\left(\alpha([0,T])\Gamma_{n+1}\right)\right) \overset{a.s.}{\to} M_{\theta_{(i)}}^{-1}(\Gamma_i)$. ∎

Algorithm H: New algorithm for simulating an approximation for the prior beta-Stacy process. Since the prior log-beta process has no fixed point of discontinuity, it is characterized entirely by its Lévy measure. The steps for an approximate sample of the beta-Stacy process with parameters $k(\cdot)$ and F_0 on $[0,T]$ as in Definition 2.6.2 are:

(1) Generate first the log-beta process with the parameters $\alpha(dz) = k(z)F_0(dz)$ and $\beta(z) = k(z)F_0[z,\infty)$. For this:

(a) Fix n large enough.

(b) Generate $\theta_i \overset{\text{i.i.d.}}{\sim} d\alpha(t)/\alpha([0,T])$ for $i = 1, \ldots, n$.

(c) For $i = 1, \ldots, n+1$, generate E_i from an exponential distribution with mean 1, independent of $(\theta_i)_{1\leq i\leq n}$ and let $\Gamma_i = E_1 + \cdots + E_i$.

(d) Find $\left(S^{-1}_{n,\theta_i}\left(\Gamma_i\beta(\theta_i)/\alpha([0,T])\Gamma_{n+1}\right)\right)_{1\leq i\leq n}$, where $S_{n,\theta_i}(x)$ is defined in (4.4.1). Note that $S^{-1}_{n,\theta_i}(y) = -\ln Q_{n,\theta_i}(y)$, where $Q_{n,\theta_i}(y)$ given θ_i is the quantile function of the beta distribution with parameters $\beta(\theta_i)(1-1/n)$ and $\beta(\theta_i)/n$ evaluated at y.

(e) Set $Z_n(t)$ as in Theorem 4.4.3.

(2) Use the relation $F_n(t) = 1 - \exp(-Z_n(t))$ to find an approximate sample of the prior beta-Stacy process.

Figure 4.8 shows sample paths generated for the log-beta process $\{F(t)\}_{t\geq 0}$ with parameters $\beta(t) = \exp(-0.1t)$ and $\alpha(dt) = 0.1\exp(-0.1t)dt$ using the Wolpert-Ickstadt algorithm and the new algorithm. The figure clearly shows that the new representation converges to the Wolpert and Ickstadt representation of the log-beta process. In this figure, simulating the Wolpert and Ickstadt representation is performed through relatively complex numerical methods, which is very time consuming.

Algorithm I: Simulating an approximation for the posterior beta-Stacy process. Simulating an approximation for the posterior beta-Stacy process, given the right censored data $(T_1, \delta_1), \ldots, (T_m, \delta_m)$, consists of two tasks:

(1) Sampling from the posterior log-beta process. This consists of the following steps:

(a) Generate a sample path from the posterior process $\{Z^*_c(t)\}_{t\in[0,T]}$, which is a pure jump process with no fixed point of discontinuity whose Lévy measure is known. The updated Lévy measure is given by (2.6.4). Thus, the algorithm

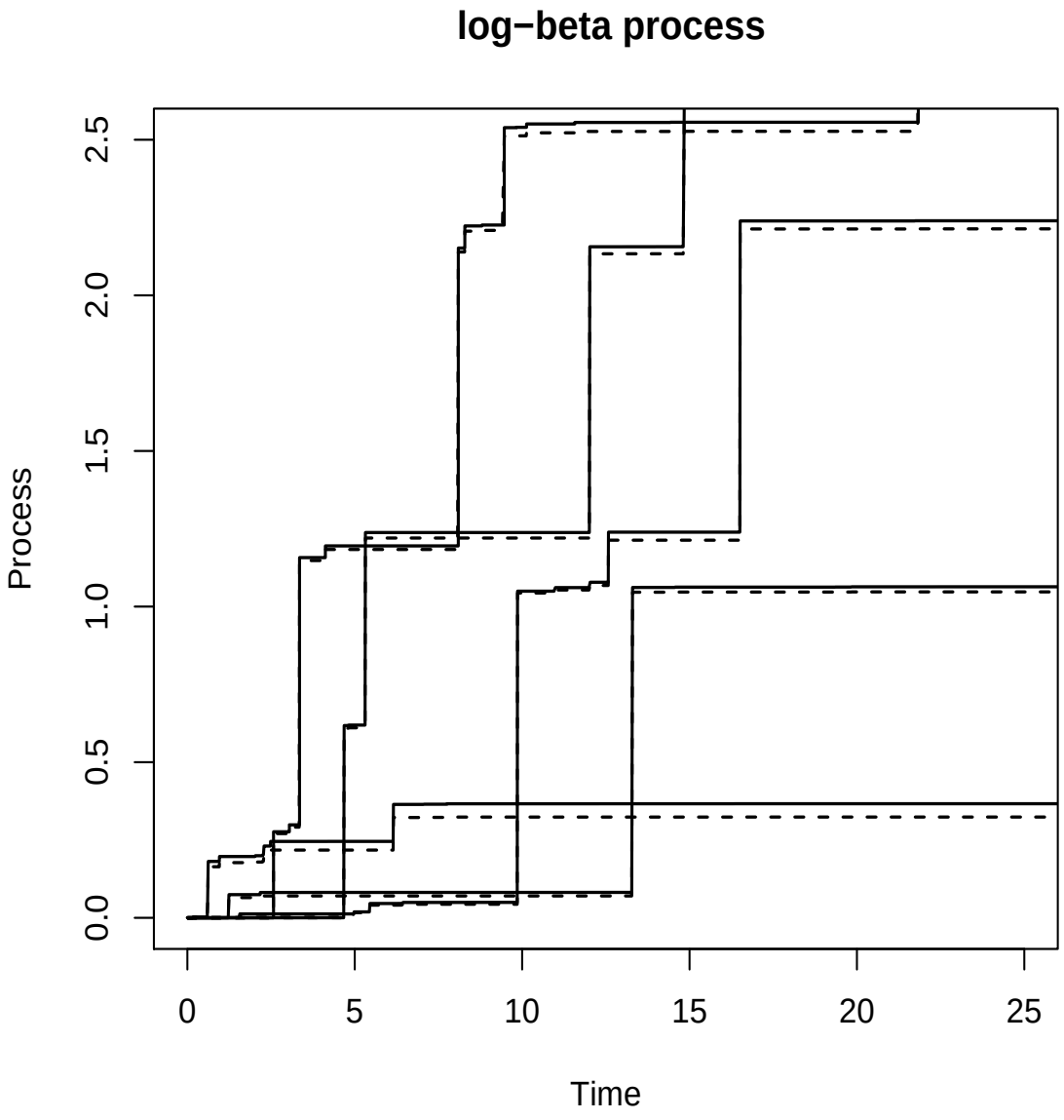

Figure 4.8: *Sample paths of a log-beta process with* $\alpha(dt) = 0.1\exp(-0.1t)dt$, *and* $\beta(t) = \exp(-0.1t)$. *The dashed lines denote the sample paths generated by the new algorithm with* $n = 1000$.

used for simulating the prior process can be used for simulating the posterior process with $\beta(\theta_i)$ replaced by $\beta(\theta_i)+Y(\theta_i)$. Denote the approximated process by $\left\{Z^*_{c,n}(t)\right\}_{t\in[0,T]}$.

(b) Generate the random jumps $(S^*_i)_{1\leq i\leq l}$ that are associated with the fixed points of discontinuity from the distribution given in (2.6.5), where $l \leq m$. Recall that these random jumps occur at points where the data is not right censored. Denote

$$Z^*_f(t) = \sum_{i=1}^{l} S^*_i I\left\{T_i \leq t\right\}.$$

(c) Put together the simulation results of the two components to obtain a sample from the posterior log-beta process. Hence, the posterior process $Z^*(t)$ is approximated by:

$$\begin{aligned} Z^*_n(t) &= Z^*_{c,n}(t) + Z^*_f(t) \\ &= \sum_{i=1}^{n} J^*_i I(\theta^*_i \leq t) + \sum_{i=1}^{l} S^*_i I\left\{T_i \leq t\right\}, \end{aligned}$$

where $(J^*_i)_{1\leq i\leq n}$ and $(\theta^*_i)_{1\leq i\leq n}$ are the jumps and the locations of the approximated process $Z^*_{c,n}(t)$.

(2) Use the relation $F^*_n(t) = 1 - \exp\left(-Z^*_n(t)\right)$ to find an approximate sample for the posterior beta-Stacy process, given the right censored data.

4.4.2 Other Sampling Algorithms

In the literature, there are several algorithms that can be used to sample from the log-beta process $\{Z(t)\}_{t\in[0,T]}$. We consider the Damien, Laud, and Smith algorithm (1995) and the Walker and Damien algorithm (1998). We also modify the Lee and Kim algorithm (2004) and the Lee (2007) algorithm to sample from the beta-Stacy process. Below is a discussion of these algorithms.

• **Damien-Laud-Smith Algorithm:** As is in the beta process, the time interval $[0, T]$ is partitioned into small subintervals with endpoints $0 = \theta_0 < \theta_1 < \ldots < \theta_m = T$. Let J_i denotes the increment of $Z(t)$ in the interval $\Delta_i = (\theta_{i-1}, \theta_i]$, i.e. $J_i = Z(\theta_i) - Z(\theta_{i-1})$. Let

$$L_{\Delta_i}(ds) = \left[\int_{\theta_{i-1}}^{\theta_i} (1 - e^{-s})^{-1} e^{-s\beta(z)} \alpha(dz)\right] ds.$$

Define

$$dG_{\Delta_i}(s) = \left(1 - e^{-s}\right) L_{\Delta_i}(ds) = \left[\int_{\theta_{i-1}}^{\theta_i} e^{-s\beta(z)} \alpha(dz)\right] ds$$

and

$$\lambda_i = \int_0^\infty dG_{\Delta_i}(s) = \int_0^\infty \int_{\theta_{i-1}}^{\theta_i} e^{-s\beta(z)} \alpha(dz) ds.$$

The steps of the Damien-Laud-Smith algorithm to generate an approximation for the log-beta process $\{Z(t)\}_{t\in[0,T]}$ are:

(1) Fix a relatively large positive integer n.

(2) Generate independent values x_{ij} from the probability density function $dG_{\Delta_i}(s)/\lambda_i$, for $j = 1, \ldots, n$.

(3) Generate y_{ij}: $y_{ij}|x_{ij} \sim Poisson\left(\lambda_i n^{-1}(1 - e^{-x_{ij}})^{-1}\right)$, for $j = 1, \ldots, n$.

(4) Set $J_{i,n} = \sum_{j=1}^n x_{ij} y_{ij}$.

(5) Set $Z_{m,n}(t) = \sum_{i=1}^m J_{i,n} I(\theta_i \leq t)$. For large m and n, $Z_{m,n}(t)$ is an approximation of $Z(t)$.

• **Walker-Damien Algorithm:** The steps of this algorithm for the log-beta process are similar to the steps of the beta process described in Subsection 4.3.2 but with G_ϵ and λ_ϵ defined as follows. Let $L_\Delta(ds)$ be the Lévy measure of the log-beta process restricted to $\Delta = [a, b) \subseteq [0, T]$, i.e.

$$L_\Delta(ds) = \left[\int_a^b (1 - e^{-s})^{-1} e^{-s\beta(z)} \alpha(dz)\right] ds.$$

Define

$$G_\epsilon(s) = L_\Delta\left((\epsilon, s]\right) = \int_\epsilon^s \int_a^b \frac{1}{1-e^{-u}} e^{-u\beta(z)} \alpha(dz)du$$

and

$$\lambda_\epsilon = L_\Delta\left((\epsilon, \infty)\right) = \int_\epsilon^\infty \int_a^b \frac{1}{1-e^{-u}} e^{-u\beta(z)} \alpha(dz)du.$$

• **Lee-Kim Algorithm:** According to Lee and Kim (2004) and Lee (2007) this algorithm can only be implemented for the gamma process and the beta process. Here we show that this algorithm can also be extended to sample from the log-beta process. The Lévy measure L_t in (2.5.3) of the log-beta process can be approximated by:

$$L_{t,\epsilon}(x) = \int_0^t \int_x^\infty \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} e^{-s\beta(z)}(1-e^{-s})^{\epsilon-1} ds\alpha(dz). \qquad (4.4.4)$$

Let

$$\lambda_\epsilon = L_{T,\epsilon}\left((0,\infty)\right) = \frac{1}{\epsilon}\int_0^T \int_0^\infty \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\epsilon\Gamma(\epsilon)\Gamma(\beta(z))} e^{-s\beta(z)}(1-e^{-s})^{\epsilon-1} ds\alpha(dz) = \frac{\alpha(T)}{\epsilon}$$

and

$$dG_\epsilon(z) = \frac{\alpha(dz)}{\epsilon} I(0 \le z \le T).$$

The steps of the adapted Lee-Kim algorithm for the log-beta process are:

(1) Fix a relatively small positive number ϵ.

(2) Generate the total number of jumps $n \sim Poisson\left(\lambda_\epsilon\right)$.

(3) For $i = 1, \ldots n$, generate the jump times θ_i from the probability density function $dG_\epsilon(z)/\lambda_\epsilon = \alpha(dz)/\alpha(T)$.

(4) Let $\theta_{(1)} \le \ldots \le \theta_{(n)}$ be the corresponding order statistics of $\theta_1, \ldots, \theta_n$.

(5) Generate the jump sizes $J_1, \ldots, J_n$:

$$J_i|\theta_{(i)} \sim \frac{\Gamma\left(\beta(\theta_{(i)}) + \epsilon\right)}{\Gamma(\epsilon)\Gamma(\beta(\theta_{(i)}))} e^{-s\beta(\theta_{(i)})}(1 - e^{-s})^{\epsilon-1}.$$

Equivalently, $1 - \exp(-J_i)|\theta_{(i)} \sim Beta(\epsilon, \beta(\theta_{(i)}))$.

(6) Set $Z_\epsilon(t) = \sum_{=1}^{n} J_i I(\theta_i \leq t)$.

The next theorem shows that $Z_\epsilon(t)$ converges in distribution to $Z(t)$, as $\epsilon \to 0$, on $D[0,T]$ with respect to the Skorohod topology. The proof is based on the following lemma of Lee and Kim (2004, page 444). Observe that, since $L_{t,\epsilon}((0,\infty)) < \infty$, Z_ϵ is a compound Poisson process. See Lee and Kim (1998) and Lee (2007) for a discussion about the relationship between the nondecreasing Lévy process and the Poisson random measure.

Lemma 4.4.4 *Let $\{Z(t)\}_{t\in[0,T]}$ be a pure jump process with Lévy measure $(L_t)_{t\in[0,T]}$. For each $\epsilon > 0$, let $\{Z_\epsilon(t)\}_{t\in[0,T]}$ be a compound Poisson process with Lévy measure $(L_{t,\epsilon})_{t\in[0,T]}$. Suppose for any bounded continuous function f with $f(s) = 0$ on $0 \leq s < \delta$ for some $\delta > 0$,*

$$\sup_{t\in[0,T]} \left| \int_0^\infty f(s) \left(L_t(ds) - L_{t,\epsilon}(ds)\right) \right| \to 0, \text{ as } \epsilon \to 0. \tag{4.4.5}$$

Then

$$Z_\epsilon(t) \xrightarrow{d} Z(t),$$

as $\epsilon \to 0$, on $D[0,T]$ with respect to the Skorohord topology.

Theorem 4.4.5 *Let $\{Z(t)\}_{t\in[0,T]}$ be a log-beta process with parameters $\alpha(t)$ and $\beta(t)$ as defined in Definition 2.6.1. Suppose that $\alpha([0,T]) < \infty$ and $0 < \inf_{t\in[o,T]} \beta(t) < \sup_{t\in[o,T]} \beta(t) < \infty$ for some $T > 0$. Then as $\epsilon \to 0$, we have $Z_\epsilon(t) \xrightarrow{d} Z(t)$ on $D[0,T]$ with respect to the Skorohord topology.*

Proof: We mimic the proof Theorem 2 of Lee and Kim (2004), which we obtained through personal communication with the authors (the proof is missing from the appendix of Lee and Kim, 2004). Let f be a bounded continuous function such that $f(x) = 0$ on $0 \leq x < \delta$ for some $\delta > 0$. We proof that (4.4.5) holds. Using (4.4.4) and (2.6.1) we have

$$\begin{aligned}
&\sup_{t\in[0,T]} \left| \int_0^\infty f(s) \left(L_{t,\epsilon}(ds) - L_t(ds)\right) \right| \\
= &\sup_{t\in[0,T]} \left| \int_0^t \int_0^\infty f(s) \Big(\frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} e^{-s\beta(z)} (1-e^{-s})^{\epsilon-1} \right. \\
&\qquad \left. - \left(1-e^{-s}\right)^{-1} e^{-s\beta(z)} \Big) ds\alpha(dz) \right| \\
\leq &\int_0^T \int_\delta^\infty |f(s)| \left| \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} e^{-s\beta(z)} (1-e^{-s})^{\epsilon-1} - \left(1-e^{-s}\right)^{-1} e^{-s\beta(z)} \right| ds\alpha(dz) \\
\leq &\int_0^T \int_\delta^\infty f_{\sup} \left| \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} (1-e^{-s})^{\epsilon} - 1 \right| (1-e^{-s})^{-1} e^{-s\beta(z)} ds\alpha(dz) \\
\leq &f_{\sup} \int_0^T \int_\delta^\infty \left| \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} (1-e^{-s})^{\epsilon} - (1-e^{-s})^{\epsilon} \right| (1-e^{-s})^{-1} e^{-s\beta(z)} ds\alpha(dz) \\
+ &f_{\sup} \int_0^T \int_\delta^\infty \left| (1-e^{-s})^{\epsilon} - 1 \right| (1-e^{-s})^{-1} e^{-s\beta(z)} ds\alpha(dz) \\
\leq &f_{\sup} \int_0^T \int_\delta^\infty \left| \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} - 1 \right| (1-e^{-s})^{\epsilon-1} e^{-s\beta(z)} ds\alpha(dz) \\
+ &f_{\sup} \int_0^T \int_\delta^\infty \left| (1-e^{-s})^{\epsilon} - 1 \right| (1-e^{-s})^{-1} e^{-s\beta(z)} ds\alpha(dz) := I_1 + I_2
\end{aligned}$$

where $f_{\sup} = \sup_{s\in(0,\infty)} |f(s)|$. Let $\beta_{\inf} = \inf_{t\in[0,T]} \beta(t)$ and $\beta_{\sup} = \sup_{t\in[0,T]} \beta(t)$. For I_1, since $0 < s < \infty$, we have $(1-e^{-s})^{\epsilon} < 1$ and $(1-e^{-s})^{-1} < (1-e^{-\delta})^{-1}$. It follows that $(1-e^{-s})^{\epsilon-1} < (1-e^{-\delta})^{-1}$. Hence we have

$$\begin{aligned}
I_1 &\leq f_{\sup}(1-e^{-\delta})^{-1} \int_0^T \int_\delta^\infty \left| \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} - 1 \right| e^{-s\beta(z)} ds\alpha(dz) \\
&\leq f_{\sup}(1-e^{-\delta})^{-1} \int_0^T \frac{1}{\beta(z)} \left| \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} - 1 \right| \alpha(dz) \\
&\leq f_{\sup} \frac{(1-e^{-\delta})^{-1}}{\beta_{\inf}} \int_0^T \left| \frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))} - 1 \right| \alpha(dz). \qquad (4.4.6)
\end{aligned}$$

Since $\Gamma(x)$ is continuous function, we have as

$$\left|\frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))}-1\right|\to 0,$$

as $\epsilon\to 0$. To apply the bounded convergence theorem, we need to show that the integrand in (4.4.6) is uniformly bounded. Observe that, for all $z\in[0,T]$, we have

$$\left|\frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))}-1\right|\leq\frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))}+1.$$

Using the fact that the function $\Gamma(x)$ is decreasing on $(0,y_0]$ and increasing on $[y_0,\infty)$, where $y_0=\min_{x>0}\Gamma(x)\approx 0.8856$ (Wrench, 1964), and the relation $\beta_{\inf}\leq\beta_{\inf}+\epsilon\leq\beta(z)+\epsilon\leq\beta_{\sup}+\epsilon\leq\beta_{\sup}+1$, we get $\Gamma\left(\beta(z)+\epsilon\right)\leq\max\left(\Gamma(\beta_{\inf}),\Gamma(\beta_{\sup}+1)\right))$ and $\Gamma(\epsilon+1)\geq y_0$. Hence

$$\left|\frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))}-1\right|\leq\frac{\max\left(\Gamma(\beta_{\inf}),\Gamma(\beta_{\sup}+1)\right))}{y_0\min\left(\Gamma(\beta_{\inf}),\Gamma(\beta_{\sup})\right)}+1.$$

Therefore, by the bounded convergence theorem,

$$\int_0^T\left|\frac{\Gamma\left(\beta(z)+\epsilon\right)}{\Gamma(\epsilon+1)\Gamma(\beta(z))}-1\right|\alpha(dz)\to 0,$$

as $\epsilon\to 0$. Thus, $I_1\to 0$. For I_2, we have: $1-e^{-\delta}<1-e^{-s}<1$. This implies that $\left(1-e^{-\delta}\right)^{\epsilon}-1<(1-e^{-s})^{\epsilon}-1<0$. Thus, $\left|\left(1-e^{-\delta}\right)^{\epsilon}-1\right|>\left|(1-e^{-s})^{\epsilon}-1\right|$. Hence

$$\begin{aligned}
I_2 &\leq f_{\sup}\int_0^T\int_\delta^\infty\left|\left(1-e^{-\delta}\right)^{\epsilon}-1\right|(1-e^{-\delta})^{-1}e^{-s\beta(z)}ds\alpha(dz)\\
&= f_{\sup}\int_0^T\left|\left(1-e^{-\delta}\right)^{\epsilon}-1\right|(1-e^{-\delta})^{-1}\left[\beta(z)\right]^{-1}e^{-\delta\beta(z)}\alpha(dz)\\
&\leq f_{\sup}\int_0^T\left|\left(1-e^{-\delta}\right)^{\epsilon}-1\right|(1-e^{-\delta})^{-1}\left[\beta_{\inf}\right]^{-1}\alpha(dz)\\
&\leq f_{\sup}\left|\left(1-e^{-\delta}\right)^{\epsilon}-1\right|(1-e^{-\delta})^{-1}\left[\beta_{\inf}\right]^{-1}\alpha(T). \qquad (4.4.7)
\end{aligned}$$

As $\epsilon\to 0$ the right hand side of (4.4.7) converges to 0. Therefore, by Lemma 4.4.4 the proof follows. ∎

• **Lee Algorithm:** This algorithm is a general algorithm to generate a sample from a nondecreasing Lévy process. Here we adapt this algorithm to sample form the log-beta process. The Lee algorithm for the log-beta process is similar to the algorithm used for the beta process described in Subsection 4.3.2, with the following modifications. Replace A_0 by α and take $K'(s,z) = \beta(z)\exp(-\beta(z)s)$. Therefore,

$$\lambda_i = \frac{\alpha(T)K(x_i,\theta_i)}{nK'(x_i,\theta_i)} = \frac{\alpha(T)\left(1-e^{-x_i}\right)^{-1}e^{-x_i\beta(\theta_i)}}{n\beta(\theta_i)e^{-\beta(\theta_i)x_i}} = \frac{\alpha(T)}{n\beta(\theta_i)\left(1-e^{-x_i}\right)}.$$

4.4.3 Empirical Results: Comparison with Other Methods

In this section, we compare the new approximation of the beta-Stacy process (Algorithms H and I) developed in Subsection 4.4.1 with the algorithms given in the previous subsection. We consider the data presented in Section 1.2 of Aalen, Borgan and Gjessing (2008). The data are described in the next table and they represent the remission duration (in months) for patients given the drug 6-mercaptopurine, where + denotes a right censored observation.

10	7	32+	23	22	6	16	34+	32+	25+	11+
20+	19+	6	17+	35+	6	13	9+	6+	10+	

The objective is to estimate the values of $F^*([0,6))$ and $F^*([6,7))$, where F^* is the posterior beta-Stacy process given the data. We follow Example 6.4 of Alan, Borgan and Gjessing (2008) and consider the prior beta-Stacy process with parameters $k(t) = 5$ and $F_0(t) = 1-\exp(-0.1t)$. That is, the parameters for the log-beta process are $\beta(z) = 5\exp(-0.1z)$ and $\alpha(dz) = 0.5\exp(-0.1z)dz$; see Definitions 2.6.1 and 2.6.2. It is also equivalent to say that the prior distribution is the Dirichlet process with concentration parameter 5 and base measure F_0; see Remark 2.6.6. We collect 2000 samples from the posterior beta-stacy process using three different approximations:

New with $n = 2000$, Lee-Kim with $\epsilon = 0.0025$, and Lee with $n = 2000$. The results (average) are given in Table 4.7.

Table 4.7: Estimates and true values of $F^*(I)$, where F^* is the posterior beta-Stacy process given the data and I is the time interval.

Interval	New	Lee-Kim	Lee	Exact
[0,6)	0.0867	0.0865	0.1007	0.0868
[6,7)	0.1259	0.1258	0.1246	0.1259

It follows from Table 4.7 that the new approximation performs very well as its estimated values are close to the exact values.

Remark 4.4.6 *In this example, applying Damien-Laud-Smith algorithm and Walker-Damien algorithm require sampling from nonstandard distributions. Hence, these algorithms are not considered in our comparison. We point out that, Walker and Damien (1998) suggested a Gibbs/Metropolis sampling algorithm to overcome sampling from nonstandard distributions.*

Chapter 5

The Distance between the Dirichlet Process and its Base Measure

In this chapter, we use the sum representations of the Dirichlet process to derive explicit expressions that are used to calculate the Kolmogorov, Lévy and Cramér-von Mises distances between the Dirichlet process and its base measure. The derived expressions of the distance are used in the next chapter for a goodness of fit test and for the selection of the concentration parameter of the Dirichlet process.

5.1 Probability Metrics

There are a host of metrics (distances) that can be found in the literature to measure the distance between two distribution functions. The most natural distance between two distribution functions is the Kolmogorov distance.

Definition 5.1.1 *Let F and G be two distribution functions. The Kolmogorov (or uniform) distance between F and G, denoted by $d_K(F, G)$, is defined as*

$$d_K(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|.$$

Note that $d_K(F,G) \in [0,1]$, for any two distribution functions F and G. While the Kolmogorov distance is easy to compute, it is not suitable to use in general probability problems as it does not metrize the weak convergence of probability measures (Huber, 1981, page 34). A more convenient metric is the Lévy metric.

Definition 5.1.2 *Let F and G be two distribution functions. The Lévy distance between F and G, denoted by $d_L(F,G)$, is defined as*

$$d_L(F,G) = \inf\{\delta > 0 : F(x-\delta) - \delta \leq G(x) \leq F(x+\delta) + \delta, \forall x \in \mathbb{R}\}.$$

The Lévy distance also takes values in $[0,1]$, but it metrizes the weak convergence of probability measures, or equivalently, the convergence in distribution of random variables. Specifically, if $(X_k)_{k\geq 1}$ denotes a sequence of random variables with probability distributions $(F_k)_{k\geq 1}$, $(X_k)_{k\geq 1}$ converges in distribution to a random variable X with probability distribution F if and only if

$$\lim_{k\to\infty} d_L(F_k, F) = 0$$

(Huber, 1981, page 25).

For any distribution functions F and G, we have:

$$d_L(F,G) \leq d_K(F,G). \tag{5.1.1}$$

If G has the density $g = G'$, then

$$d_K(F,G) \leq \left(1 + \sup_x G'(x)\right) d_L(F,G).$$

Form (5.1.1) it follows that if $(F_k)_{k\geq 1}$ converges to F in $(\mathscr{F}, d_K)$, then it also converges to F in $(\mathscr{F}, d_L)$, where $\mathscr{F}$ is the space of distribution functions. The converse is not true; see Pestman (2009, page 320) for a counter example. However, if the limit F is absolutely continuous, then it follows that $(F_k)_{k\geq 1}$ converges to F in $(\mathscr{F}, d_K)$ if and only if F_k converges to F in $(\mathscr{F}, d_L)$. For the proof of the last argument, see Pestman (2009, Theorem VII 5.6, page 321).

Another widely used distance is the Cramér-von Mises distance.

Definition 5.1.3 *Let F and G be two distribution functions. The Cramér-von Mises distance between F and G, denoted by $d_{CvM}(F,G)$, is defined as*

$$d_{CvM}(F,G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 \, g(x)dx,$$

where $g(x) = G'(x)$ is the probability density function.

Clearly, the Cramér-von Mises distance is not a metric since it is not symmetric. However, this distance is very useful in applications. This point will be clarified in the next chapter. For more details about metrics (distances) on probability measures consult Dudley (2002, pages 393-398) and Gibbs and Su (2002).

5.2 The Kolmogorov Distance

In this subsection, we derive explicit expressions for the Kolmogorov distance between the Dirichlet process and its base measure. These formulas are useful in some applications such as testing statistical hypothesis. Furthermore, they can be used to set a reasonable value of the concentration parameter in the Dirichlet process. More details are found in Chapter 6. We refer to the Kolmogorov distance between the Dirichlet process $P \sim DP(a,H)$ and the base measure H as the Kolmogorov distance. We denote this distance by $d_K(P,H)$, where

$$d_K(P,H) = \sup_{x\in\mathbb{R}} |P(-\infty,x] - H(-\infty,x]| := \sup_{x\in\mathbb{R}} |P(x) - H(x)|.$$

Let H be a continuous distribution and P_n be a discrete distribution with jump points $(\theta_k)_{1\leq k\leq n}$. Let $\theta_{(1)} \leq \ldots \leq \theta_{(n)}$ be the order statistics of $(\theta_k)_{1\leq k\leq n}$ and define $\theta_{(0)} = -\infty$. To compute $d_K(P_n,H)$ notice that the largest difference between P_n and H is achieved either before or after one of the jumps. Therefore,

$$d_K(P_n,H) = \max\{d_K^{(1)}, d_K^{(2)}\}, \tag{5.2.1}$$

where

$$d_K^{(1)} = \max_{1\le i\le n} |P_n(\theta_{(i-1)}) - H(\theta_{(i)})|, \quad d_K^{(2)} = \max_{1\le i\le n} |P_n(\theta_{(i)}) - H(\theta_{(i)})|. \tag{5.2.2}$$

The previous formulas for the Komogorov distance between P_n and H are justified formally by the following proposition.

Proposition 5.2.1 *Let H be a continuous distribution and P_n be a discrete distribution with jump points $(\theta_k)_{1\le k\le n}$. Then*

$$d_K(P_n, H) = \max\{d_K^{(1)}, d_K^{(2)}\},$$

where $d_K^{(1)}$ and $d_K^{(2)}$ are defined by (5.2.2), $\theta_{(1)} \le \ldots \le \theta_{(n)}$ are the order statistics of $(\theta_k)_{1\le k\le n}$ and $\theta_{(0)} = -\infty$.

Proof: Note that

$$P_n(x) = \sum_{i=1}^{n} P_n(\theta_{(i)}) I_{[\theta_{(i)},\theta_{(i+1)})}(x).$$

We write $\mathbb{R}$ as a union of $n+1$ disjoint intervals:

$$\mathbb{R} = (-\infty, \theta_{(1)}) \bigcup \bigcup_{i=1}^{n} [\theta_{(i)}, \theta_{(i+1)}),$$

where $\theta_{(n+1)} = \infty$. Then

$$d_K(P_n, H) = \max\{ \sup_{x<\theta_{(1)}} |P_n(x) - H(x)|, \max_{1\le i\le n} \sup_{\theta_{(i)}\le x<\theta_{(i+1)}} |P_n(x) - H(x)|\}. \tag{5.2.3}$$

If $x < \theta_{(1)}$, then $|P_n(x) - H(x)| = H(x)$. Hence,

$$\sup_{x<\theta_{(1)}} |P_n(x) - H(x)| = \sup_{x<\theta_{(1)}} H(x) = H(\theta_{(1)}), \tag{5.2.4}$$

since H is continuous.

If $x \in [\theta_{(i)}, \theta_{(i+1)})$ for some $1 \le i \le n$, then

$$|P_n(x) - H(x)| = |P_n(\theta_{(i)}) - H(x)|.$$

Let $A_i = \{x \in [\theta_{(i)}, \theta_{(i+1)}) : P_n(\theta_{(i)}) \geq H(x)\}$ and $B_i = \{x \in [\theta_{(i)}, \theta_{(i+1)}) : P_n(\theta_{(i)}) < H(x)\}$. Since H is continuous,

$$\begin{aligned} \sup_{x \in A_i} |P_n(\theta_{(i)}) - H(x)| &= |P_n(\theta_{(i)}) - H(\theta_{(i)})| \\ \sup_{x \in B_i} |P_n(\theta_{(i)}) - H(x)| &= |P_n(\theta_{(i)}) - H(\theta_{(i+1)})|. \end{aligned}$$

Hence

$$\begin{aligned} \sup_{\theta_{(i)} \leq x < \theta_{(i+1)}} |P_n(x) - H(x)| &= \sup_{\theta_{(i)} \leq x < \theta_{(i+1)}} |P_n(\theta_{(i)}) - H(x)| \\ &= \max\{\sup_{x \in A_i} |P_n(\theta_{(i)}) - H(x)|, \sup_{x \in B_i} |P_n(\theta_{(i)}) - H(x)|\} \\ &= \max\{|P_n(\theta_{(i)}) - H(\theta_{(i)})|,\ |P_n(\theta_{(i)}) - H(\theta_{(i+1)})|\} \end{aligned}$$

and

$$\begin{aligned} &\max_{1 \leq i \leq n} \sup_{\theta_{(i)} \leq x < \theta_{(i+1)}} |P_n(x) - H(x)| = \\ &\quad = \max_{1 \leq i \leq n} \max\{|P_n(\theta_{(i)}) - H(\theta_{(i)})|,\ |P_n(\theta_{(i)}) - H(\theta_{(i+1)})|\} \\ &\quad = \max\{\max_{1 \leq i \leq n} |P_n(\theta_{(i)}) - H(\theta_{(i)})|,\ \max_{1 \leq i \leq n} |P_n(\theta_{(i)}) - H(\theta_{(i+1)})|\}. \end{aligned} \tag{5.2.5}$$

The conclusion follows from (5.2.3), (5.2.4) and (5.2.5). ∎

In the next proposition, let $(J'_i)_{1 \leq i \leq n}$ be the reordered weights associated with $\left(\theta_{(i)}\right)_{1 \leq i \leq n}$ such that $\sum_{i=1}^n J'_i \delta_{\theta_{(i)}} = \sum_{i=1}^n J_i \delta_{\theta_i}$, where $\theta_{(1)} \leq \cdots \leq \theta_{(n)}$ represent the order statistics of $\theta_1, \ldots, \theta_n$.

Proposition 5.2.2 *Let H be a continuous distribution. Let $P_n = \sum_{i=k}^n J_k \delta_{\theta_k}$ where $(\theta_k)_{1 \leq k \leq n}$ are i.i.d. random variables with common distribution H. Then*

$$d_K(P_n, H) = \max\left\{d_K^{(1)}, d_K^{(2)}\right\},$$

where

$$d_K^{(1)} = \max_{1 \leq i \leq n} \left| \sum_{k=1}^{i-1} J'_k - H(\theta_{(i)}) \right| \quad \textit{and} \quad d_K^{(2)} = \max_{1 \leq i \leq n} \left| \sum_{k=1}^{i} J'_k - H(\theta_{(i)}) \right|,$$

with the convention that $\sum_{k=1}^0 J'_k = 0$.

Proof: Notice that

$$P_n = \sum_{k=1}^{n} J'_k \delta_{\theta_{(k)}}.$$

We apply (5.2.1). Then

$$d_K(P_n, H) = \max\{d_K^{(1)}, d_K^{(2)}\},$$

where $d_K^{(1)} = \max_{1\le i\le n} |P_n(\theta_{(i-1)}) - H(\theta_{(i)})|$ and $d_k^{(2)} = \max_{1\le i\le n} |P_n(\theta_{(i)}) - H(\theta_{(i)})|$. We can rewrite $d_K^{(1)}$ as follows:

$$\begin{aligned} d_K^{(1)} &= \max_{1\le i\le n} \left| \sum_{k=1}^{n} J'_k \delta_{\theta_{(k)}} \left(\left(-\infty, \theta_{(i-1)}\right] \right) - H(\theta_{(i)}) \right| \\ &= \max_{1\le i\le n} \left| \sum_{k=1}^{i-1} J'_k - H(\theta_{(i)}) \right|. \end{aligned}$$

Similarly, we get

$$d_K^{(2)} = \max_{1\le i\le n} \left| \sum_{k=1}^{i} J'_k - H(\theta_{(i)}) \right|.$$

This completes the proof. ∎

Our next goal is to prove that $d_K(P_n^{\text{F.D.}}, H) \xrightarrow{d} d_K(P, H)$, where $P_n^{\text{F.D.}}$ is the finite-dimensional Dirichlet prior defined by (2.1.13). Note that, since P is discrete a.s. and the Kolmogorov distance does not metrize the weak convergence, $d_K(P_n^{\text{F.D.}}, P)$ does not necessarily converge to zero. Therefore, $d_K(P_n^{\text{F.D.}}, H) \xrightarrow{d} d_K(P, H)$ does not follow from the triangle inequality

$$\left| d_K(P_n^{\text{F.D.}}, H) - d_K(P, H) \right| \le d_K(P_n^{\text{F.D.}}, P).$$

The next lemma is used to show that $d_K(P_n^{\text{F.D.}}, H) \xrightarrow{d} d_K(P, H)$.

Lemma 5.2.3 *Let d be a function on $\mathfrak{X} \times \mathfrak{X}$. Define the function $f : \mathfrak{X} \to \mathbb{R}^+$ by $f(x) = d(x, y)$ for fixed $y \in \mathfrak{X}$. Then f is a continuous function.*

Proof: Let $a \in \mathfrak{X}$ be arbitrary. Given any $\epsilon > 0$, we need to find a positive number δ such that

$$\text{if } 0 < d(x,a) \leq \delta, \text{ then } |f(x) - f(a)| = |d(x,y) - d(a,y)| < \epsilon.$$

By the triangle inequality we have

$$d(x,y) - d(a,y) \leq d(x,a) < \delta$$

and

$$d(x,y) - d(a,y) \geq -d(x,a) > -\delta.$$

Thus,

$$-\delta < d(x,y) - d(a,y) < \delta$$

which implies that

$$|d(x,y) - d(a,y)| < \delta.$$

Therefore, if we take $\delta = \epsilon$, we are guaranteed that $|f(x) - f(a)| < \epsilon$. ∎

Corollary 5.2.4 *For any probability metric d we have:*

(i) $d\left(P_n^{F.D.}, H\right) \xrightarrow{d} d\left(P, H\right)$, *where $P_n^{F.D.}$ is the finite-dimensional Dirichlet prior defined by (2.1.13) and $P \sim DP(a, H)$.*

(ii) $d\left(P_n^{new}, H\right) \xrightarrow{a.s.} d\left(P^{Ferg.}, H\right)$, *where P_n^{new} is defined in (4.1.6) and $P^{Ferg.}$ is the Ferguson representation of the Dirichlet process defined by (2.1.4).*

(iii) $d\left(P_n^{Seth.}, H\right) \xrightarrow{a.s.} d\left(P^{Seth.}, H\right)$, *where $P_n^{Seth.}$ is the truncated Sethuraman representation of the Dirichlet process defined by (2.1.11) and $P^{Seth.}$ is the Sethuraman representation of the Dirichlet process defined by (2.1.10).*

Proof: As $n \to \infty$, $P_n^{\text{new}} \xrightarrow{a.s.} P^{\text{Ferg.}}$, $P_n^{\text{Seth.}} \xrightarrow{a.s.} P^{\text{Seth.}}$ and $P_n^{\text{F.D.}} \xrightarrow{d} P$. By Lemma 5.2.3 the proof follows. See also Serfling (2002, page 24). ∎

The next two lemmas will be used in the proof of the next proposition. For the proof of the first lemma, see Breiman (1968, Section 13.6). The proof of the second lemma is included for the sake of completeness.

Lemma 5.2.5 *Let $(U_i)_{1\le i\le n}$ be a sequence of i.i.d. random variables with a uniform distribution on $[0,1]$, and $U_{(1)} \le \ldots \le U_{(n)}$ be the corresponding order statistics. Then*

$$\left(U_{(1)}, U_{(2)}, \ldots, U_{(n)}\right) \stackrel{d}{=} \left(\frac{\Gamma_1}{\Gamma_{n+1}}, \frac{\Gamma_2}{\Gamma_{n+1}}, \ldots, \frac{\Gamma_n}{\Gamma_{n+1}}\right),$$

where $\Gamma_i = E_1 + \cdots + E_i$ and $(E_i)_{i\ge 1}$ are i.i.d. random variables with exponential distribution of mean 1.

Lemma 5.2.6 *Let $(X_{i,n})_{1\le i\le n, n\ge 1}$ and $(Y_{i,n})_{1\le i\le n, n\ge 1}$ be two independent collections of random variables. Define*

$$Z_n := \sup_{1\le i\le n} |X_{i,n} - Y_{i,n}|.$$

Let $(Y'_{i,n})_{1\le i\le n, n\ge 1}$ be a collection of random variables such that $(X_{i,n})_{1\le i\le n, n\ge 1}$ and $(Y'_{i,n})_{1\le i\le n, n\ge 1}$ are independent, and $(Y_{1,n}, \ldots, Y_{n,n}) \stackrel{d}{=} (Y'_{1,n}, \ldots, Y'_{n,n})$ for all $n \ge 1$. Define

$$Z'_n := \sup_{1\le i\le n} |X_{i,n} - Y'_{i,n}|.$$

Then $Z_n \stackrel{d}{=} Z'_n$ for all $n \ge 1$.

Proof: Let $\mathbf{X}_n = (X_{i,n})_{1\le i\le n}$, $\mathbf{Y}_n = (Y_{i,n})_{1\le i\le n}$ and $\mathbf{Y}'_n = (Y'_{i,n})_{1\le i\le n}$. Then

$$Z_n = h(\mathbf{X}_n, \mathbf{Y}_n) \quad \text{and} \quad Z'_n = h(\mathbf{X}_n, \mathbf{Y}'_n),$$

where $h((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sup_{i\le n} |x_i - y_i|$.

Due to the independence between $\mathbf{X}_n$ and $\mathbf{Y}_n$, the law of Z_n is:

$$P \circ Z_n^{-1} = P \circ (\mathbf{X}_n, \mathbf{Y}_n)^{-1} \circ h^{-1} = [(P \circ \mathbf{X}_n^{-1}) \times (P \circ \mathbf{Y}_n^{-1})] \circ h^{-1}.$$

Similarly, the law of Z'_n is:

$$P \circ (Z'_n)^{-1} = P \circ (\mathbf{X}_n, \mathbf{Y'}_n)^{-1} \circ h^{-1} = [(P \circ \mathbf{X}_n^{-1}) \times (P \circ \mathbf{Y'}_n^{-1})] \circ h^{-1}.$$

Since $(P \circ \mathbf{Y}_n^{-1}) = (P \circ \mathbf{Y'}_n^{-1})$, it follows that $P \circ Z_n^{-1} = P \circ (Z'_n)^{-1}$. ∎

The next proposition finds an explicit formula for $d_K(P_n^{\text{F.D.}}, H)$.

Proposition 5.2.7 *If H is continuous and $P_n^{F.D.}$ is defined by (2.1.13), then*

$$d_K(P_n^{F.D.}, H) \stackrel{d}{=} \max\left\{d_K^{(1)}, d_K^{(2)}\right\},$$

where

$$d_K^{(1)} = \max_{1\le i\le n} \left| \sum_{k=1}^{i-1} \frac{G_{k,n}}{\sum_{k=1}^{n} G_{k,n}} - \sum_{k=1}^{i} \frac{E_k}{\sum_{k=1}^{n+1} E_k} \right|$$

and

$$d_K^{(2)} = \max_{1\le i\le n} \left| \sum_{k=1}^{i} \frac{G_{k,n}}{\sum_{k=1}^{n} G_{k,n}} - \sum_{k=1}^{i} \frac{E_k}{\sum_{k=1}^{n+1} E_k} \right|,$$

with the convention that $\sum_{k=1}^{0} \frac{G_{k,n}}{\sum_{k=1}^{n} G_{k,n}} = 0$. Here $(G_{k,n})_{k\le n}$ is a sequence of i.i.d. random variables with a Gamma$(a/n, 1)$ distribution and $(E_k)_{k\le n+1}$ is a sequence of i.i.d. random variables with exponential distribution of mean 1, independent of $(G_{k,n})_{k\le n}$.

Proof: Let $\theta_{(1)} \le \ldots \le \theta_{(n)}$ are the order statistics of $(\theta_k)_{1\le k\le n}$. Since the random weights $(p_{k,n})_{1\le k\le n}$ of the finite dimensional Dirichlet prior are exchangeable (Ishwaran and Zarepour, 2002), we have

$$P_n^{\text{F.D.}} \stackrel{d}{=} \sum_{k=1}^{n} p_{k,n}\delta_{\theta_{(k)}} =: Q_n^{\text{F.D.}}.$$

Hence, $d_K(P_n^{\text{F.D.}}, H) \stackrel{d}{=} d_K(Q_n^{\text{F.D.}}, H)$ since the map $h : M_1(\mathbb{R}) \to [0, \infty)$ defined by $h(P) = d_K(P, H)$ is measurable (Lemma 5.2.3), where $M_1(\mathbb{R})$ is the space of probability measures on $\mathbb{R}$ endowed with the topology of weak convergence. We use (5.2.1) for $Q_n^{\text{F.D.}}$. Then

$$d_K(Q_n^{\text{F.D.}}, H) = \max\{d_K^{(1)}, d_K^{(2)}\},$$

where $d_K^{(1)} = \max_{1\le i\le n} |Q_n^{\text{F.D.}}(\theta_{(i-1)}) - H(\theta_{(i)})|$ and $d_k^{(2)} = \max_{1\le i\le n} |Q_n^{\text{F.D.}}(\theta_{(i)}) - H(\theta_{(i)})|$. Note that

$$\begin{aligned} d_K^{(1)} &= \max_{1\le i\le n} \left| \sum_{k=1}^{n} p_{k,n} \delta_{\theta_{(k)}} \left(\left(-\infty, \theta_{(i-1)} \right] \right) - H(\theta_{(i)}) \right| \\ &= \max_{1\le i\le n} \left| \sum_{k=1}^{i-1} \frac{G_{k,n}}{\sum_{k=1}^{n} G_{k,n}} - H(\theta_{(i)}) \right|. \end{aligned}$$

Since $(\theta_i)_{i\ge 1}$ is a sequence of i.i.d. random variables with continuous distribution H, $(U_i := H(\theta_i))_{i\ge 1}$ is a sequence of i.i.d. random variables with a uniform distribution on $[0, 1]$. Let $U_{(1)} \le \ldots \le U_{(n)}$ be the sequence of order statistics of $(U_i)_{1\le i\le n}$. Since H is nondecreasing, $U_{(i)} = H(\theta_{(i)})$ for all $i \le n$. By Lemma 5.2.5,

$$\left(H(\theta_{(i)}\right)_{1\le i\le n} \stackrel{d}{=} \left(\frac{\Gamma_i}{\Gamma_{n+1}} \right)_{1\le i\le n} = \left(\sum_{k=1}^{i} \frac{E_k}{\sum_{k=1}^{n+1} E_k} \right)_{1\le i\le n}.$$

Using Lemma 5.2.6 with $X_i = \sum_{k=1}^{i} \frac{G_{k,n}}{\sum_{k=1}^{n} G_{k,n}}$ and $Y_i' = \sum_{k=1}^{i} \frac{E_k}{\sum_{k=1}^{n+1} E_k}$, we obtain:

$$d_K^{(1)} = \max_{1\le i\le n} \left| \sum_{k=1}^{i-1} \frac{G_{k,n}}{\sum_{k=1}^{n} G_{k,n}} - \sum_{k=1}^{i} \frac{E_k}{\sum_{k=1}^{n+1} E_k} \right|.$$

Similarly we get

$$d_K^{(2)} = \max_{1\le i\le n} \left| \sum_{k=1}^{i} \frac{G_{k,n}}{\sum_{k=1}^{n} G_{k,n}} - \sum_{k=1}^{i} \frac{E_k}{\sum_{k=1}^{n+1} E_k} \right|.$$

This completes the proof of the proposition. ∎

It is worth noting that, in Propositions 5.2.2 and 5.2.7 the distribution of the Kolmogorov distance $d_K(P_n, H)$ does not depend on H.

5.3 The Lévy Distance

In this subsection, we consider the Lévy distance between a Dirichlet process and its base measure. We show that this distance is equal to half of the Kolmogorov distance. Therefore, all the previous results for the Kolmogorov distance can be used to calculate the Lévy distance.

Lemma 5.3.1 *Let H be a continuous distribution and P_n be a discrete distribution with jump points $(\theta_k)_{1\leq k\leq n}$ where $(\theta_k)_k$ are i.i.d random variables with common distribution H. Let $\theta_{(1)} \leq \ldots \leq \theta_{(n)}$ be the order statistics of $(\theta_k)_{1\leq k\leq n}$ and $\theta_{(0)} = -\infty$. Then*

$$d_L(P_n, H) = \frac{1}{2} d_K(P_n, H).$$

Proof: The Lévy distance $d_L(P_n, H)$ is equal to the side of the largest square with sides parallel to the coordinate axes, which can be inscribed between the graphs of the functions P_n and H (Zolotarev, 1997, page 63). Equivalently, $d_L(P_n, H)$ is equal to $1/\sqrt{2}$ times the maximum distance between the graphs of P_n and H, measured along a 45 °-direction (Huber, 1981, page 25). Notice that this distance is achieved either before or after one of the jumps. Thus,

$$d_L(P_n, H) = \frac{1}{\sqrt{2}} \max\left(d_L^{(1)}, d_L^{(2)}\right), \tag{5.3.1}$$

where

$$d_L^{(1)} = \max_{1\leq i\leq n} d_i^{(1)} \quad d_L^{(2)} = \max_{1\leq i\leq n} d_i^{(2)},$$

$$d_i^{(1)} = \sqrt{\left(\theta_{(i)} - a_i\right)^2 + \left(P_n\left(\theta_{(i-1)}\right) - H\left(a_i\right)\right)^2}, \tag{5.3.2}$$

$$d_i^{(2)} = \sqrt{\left(\theta_{(i)} - b_i\right)^2 + \left(P_n\left(\theta_{(i)}\right) - H\left(b_i\right)\right)^2}, \tag{5.3.3}$$

and

$$\theta_{(i)} - a_i = H\left(a_i\right) - P_n\left(\theta_{(i-1)}\right), \tag{5.3.4}$$

$$\theta_{(i)} - b_i = H(b_i) - P_n\left(\theta_{(i)}\right) \tag{5.3.5}$$

(see Figure 5.1 below). Here the points $(a_i, H(a_i))$ and $(b_i, H(b_i))$ are on the graph of H. Substituting (5.3.4) in (5.3.2) and (5.3.5) in (5.3.3) we get,

$$d_i^{(1)} = \sqrt{2}|\theta_{(i)} - a_i|, \tag{5.3.6}$$

and

$$d_i^{(2)} = \sqrt{2}|\theta_{(i)} - b_i|. \tag{5.3.7}$$

Assume first that H is the uniform distribution on $[0,1]$. Since $(\theta_k)_{1\le k\le n}$ have distribution H, we have $\theta_k \in [0,1]$, for all $k = 1, 2, \ldots, n$. Thus, $H(a_i) = a_i$ and, from (5.3.4), $a_i = \left(\theta_{(i)} + P(\theta_{(i-1)})\right)/2$. Therefore, equation (5.3.6) reduces to

$$d_i^{(1)} = \frac{\sqrt{2}}{2}\left|P_n\left(\theta_{(i-1)}\right) - \theta_{(i)}\right|, \tag{5.3.8}$$

Similarly, $H(b_i) = b_i$. Thus, from (5.3.5), we get $b_i = \left(\theta_{(i)} + P(\theta_{(i)})\right)/2$. Substituting in (5.3.7), we obtain

$$d_i^{(2)} = \frac{\sqrt{2}}{2}\left|P_n\left(\theta_{(i)}\right) - \theta_{(i)}\right|, \tag{5.3.9}$$

From (5.3.1), (5.3.8), and (5.3.9) we obtain

$$\begin{aligned} d_L(P_n, H) &= \frac{1}{2}\max\left\{\max_{1\le i\le n}|P_n(\theta_{(i-1)}) - \theta_{(i)}|, \max_{1\le i\le n}|P_n(\theta_{(i)}) - \theta_{(i)}|\right\} \\ &= \frac{1}{2}\max\{d_K^{(1)}, d_K^{(2)}\} \\ &= \frac{1}{2}d_K(P_n, H). \end{aligned}$$

Thus, the lemma holds when H is the uniform distribution on $[0,1]$. The case of a general continuous distribution H follows since if $(\theta_i)_{1\le i\le n}$ are i.i.d. random variables with common distribution H, $(H(\theta_i))_{1\le i\le n}$ are i.i.d. random variables with a uniform distribution on $[0,1]$. This completes the proof of the lemma. ∎

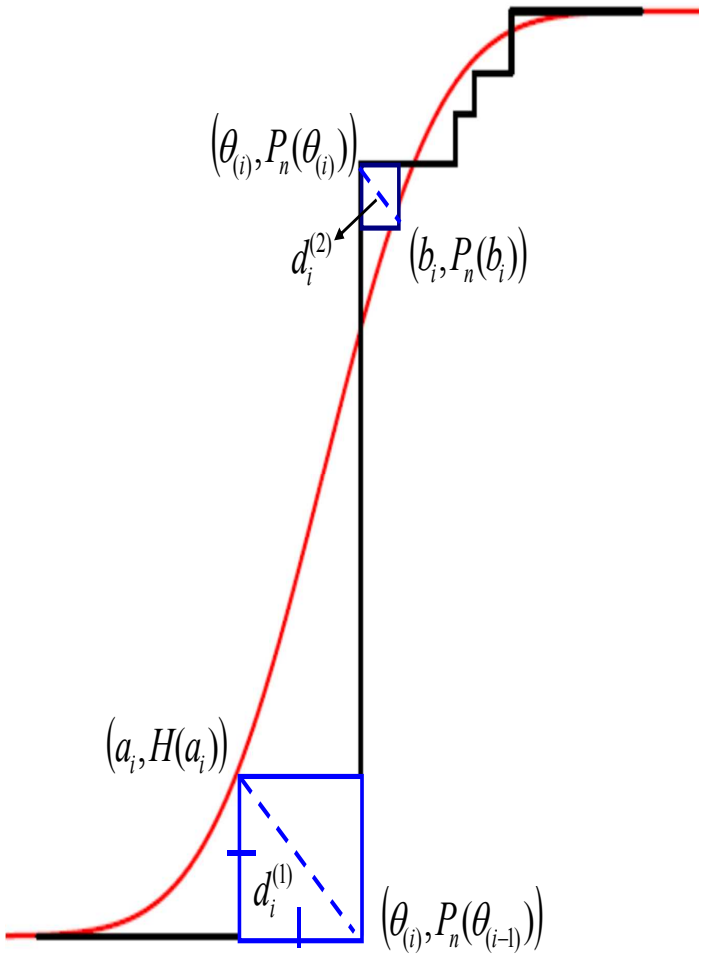

Figure 5.1: *Illustration of the Lévy distance calculation.*

5.4 The Cramér-von Mises Distance

In this subsection, we consider the Cramér-von Mises distance between a Dirichlet process and its base measure. The next proposition gives a simple formula to compute this distance.

Lemma 5.4.1 *Let H be a continuous distribution. Let $P_n = \sum_{i=1}^n J_i \delta_{\theta_i}$, where the jump points $(\theta_k)_{1\leq k\leq n}$ are i.i.d random variables with common distribution H. Let $\theta_{(1)} \leq \ldots \leq \theta_{(n)}$ be the order statistics of $(\theta_k)_{1\leq k\leq n}$ and $J'_1, \ldots, J'_n$ be the corresponding jump sizes. Then*

$$\begin{aligned} d_{CvM}(P_n, H) &= \frac{1}{3} - \sum_{i=1}^n J'_i + \left(\sum_{i=1}^n J'_i\right)^2 + \sum_{i=1}^n J'_i H^2(\theta_{(i)}) \\ &\quad - \sum_{i=1}^n J'^2_i H(\theta_{(i)}) - 2\sum_{i=2}^n \left[J'_i H(\theta_{(i)}) \sum_{k=1}^{i-1} J'_k \right]. \end{aligned}$$

Proof: Note that

$$P_n(x) = \begin{cases} 0 & \text{if} \quad x < \theta_{(1)} \\ P_n(\theta_{(i)}) & \text{if} \quad \theta_{(i)} \leq x < \theta_{(i+1)} \quad (i = 1, \ldots, n-1). \\ 1 & \text{if} \quad x \geq \theta_{(n)} \end{cases}$$

Let $\theta_{(0)} = -\infty$ and $\theta_{(n+1)} = +\infty$. Then

$$\begin{aligned} d_{CvM}(P_n, H) &= \int_{\theta_{(0)}}^{\theta_{(n+1)}} \left[P_n(x) - H(x)\right]^2 h(x)dx \\ &= \sum_{i=0}^n \int_{\theta_{(i)}}^{\theta_{(i+1)}} \left[P_n(\theta_{(i)}) - H(x)\right]^2 h(x)dx. \end{aligned}$$

We integrate by substitution. Let $y = H(x)$, we have $dy = h(x)dx$. We denote $U_{(i)} = H(\theta_{(i)})$. Thus,

$$\begin{aligned} d_{CvM}(P_n, H) &= \sum_{i=0}^n \int_{U_{(i)}}^{U_{(i+1)}} \left[P_n(\theta_{(i)}) - y\right]^2 dy \\ &= \frac{1}{3}\sum_{i=0}^n \left\{ \left[P_n(\theta_{(i)}) - U_{(i)}\right]^3 - \left[P_n(\theta_{(i)}) - U_{(i+1)}\right]^3 \right\}. \end{aligned}$$

$$\begin{aligned} &= \frac{1}{3}\sum_{i=0}^{n}\left[U_{(i+1)}^3 - U_{(i)}^3\right] \\ &\quad -\sum_{i=0}^{n} P_n(\theta_{(i)})\left[U_{(i+1)}^2 - U_{(i)}^2\right] \\ &\quad +\sum_{i=0}^{n} P_n^2(\theta_{(i)})\left[U_{(i+1)} - U_{(i)}\right] \\ &= I_1 - I_2 + I_3. \end{aligned}$$

Observe that

$$I_1 = \frac{1}{3}\left[\theta_{(n+1)}^3 - \theta_{(0)}^3\right] = \frac{1}{3}. \tag{5.4.1}$$

$$\begin{aligned} I_2 &= P_n(\theta_{(0)})\left[U_{(1)}^2 - U_{(0)}^2\right] + P_n(\theta_{(1)})\left[U_{(2)}^2 - U_{(1)}^2\right] \\ &\quad + \ldots + P_n(\theta_{(n)})\left[U_{(n+1)}^2 - U_{(n)}^2\right] \\ &= -U_{(0)}^2 P_n(\theta_{(0)}) + U_{(1)}^2\left[P_n(\theta_{(0)}) - P_n(\theta_{(1)})\right] + \ldots \\ &\quad + U_{(n)}^2\left[P_n(\theta_{(n-1)}) - P_n(\theta_{(n)})\right] + U_{(n+1)}^2 P_n(\theta_{(n)}). \end{aligned}$$

Since $P_n(\theta_{(0)}) = 0$, $U_{(n+1)} = 1$, $P_n(\theta_{(i)}) - P_n(\theta_{(i+1)}) = J_i'$, and $P_n(\theta_{(n)}) = \sum_{i=1}^{n} J_i'$ we have:

$$\begin{aligned} I_2 &= -\sum_{i=1}^{n} J_i' U_{(i)}^2 + \sum_{i=1}^{n} J_i' \\ &= -\sum_{i=1}^{n} J_i' H^2(\theta_{(i)}) + \sum_{i=1}^{n} J_i'. \end{aligned} \tag{5.4.2}$$

Similarly,

$$\begin{aligned} I_3 &= P_n^2(\theta_{(0)})\left[U_{(1)} - U_{(0)}\right] + P_n^2(\theta_{(1)})\left[U_{(2)} - U_{(1)}\right] \\ &\quad + \ldots + P_n^2(\theta_{(n)})\left[U_{(n+1)} - U_{(n)}\right] \\ &= -U_{(0)} P_n^2(\theta_{(0)}) + U_{(1)}\left[P_n^2(\theta_{(0)}) - P_n^2(\theta_{(1)})\right] + \ldots \\ &\quad + U_{(n)}\left[P_n^2(\theta_{(n-1)}) - P_n^2(\theta_{(n)}\right] + U_{(n+1)} P_n^2(\theta_{(n)}) \end{aligned}$$

$$= -U_{(1)}J'^2_1 + \sum_{i=1}^{n-1} U_{(i+1)}\left[P^2(\theta_{(i)}) - P^2(\theta_{(i+1)})\right] + \left(\sum_{i=1}^{n} J'_i\right)^2.$$

Note that

$$\begin{aligned} P_n^2(\theta_{(i)}) - P_n^2(\theta_{(i+1)}) &= \left[P_n(\theta_{(i)}) - P_n(\theta_{(i+1)})\right]\left[P_n(\theta_{(i)}) + P_n(\theta_{(i+1)})\right] \\ &= -J'_{i+1}[2P_n(\theta_{(i)}) + J'_{i+1}] \\ &= -J'_{i+1}\left[2\sum_{k=1}^{i} J'_k + J'_{i+1}\right] \\ &= -J'^2_{i+1} - 2J'_{i+1}\sum_{k=1}^{i} J'_k. \end{aligned}$$

Therefore,

$$\begin{aligned} I_3 &= -U_{(1)}J'^2_1 - \sum_{i=1}^{n-1} U_{(i+1)}\left[J'^2_{i+1} + 2J'_{i+1}\sum_{k=1}^{i} J'_k\right] + \left(\sum_{i=1}^{n} J'_i\right)^2 \\ &= -\sum_{i=1}^{n} U_{(i)}J'^2_i + \left(\sum_{i=1}^{n} J'_i\right)^2 - 2\sum_{i=2}^{n} U_{(i)}J'_i\sum_{k=1}^{i-1} J'_k \\ &= -\sum_{i=1}^{n} H(\theta_{(i)})J'^2_i + \left(\sum_{i=1}^{n} J'_i\right)^2 - 2\sum_{i=2}^{n} H(\theta_{(i)})J'_i\sum_{k=1}^{i-1} J'_k. \qquad (5.4.3) \end{aligned}$$

The lemma follows by adding (5.4.1), (5.4.2), and (5.4.3). ∎

Remark 5.4.2 *Let $P \sim DP(a, H)$ and P_n be an approximation of P. For n large, the sum of the jumps of P_n is approximately 1. Hence, for n large,*

$$d_{CvM}(P_n, H) = \frac{1}{3} + \sum_{i=1}^{n} J'_i H^2(\theta_{(i)}) - \sum_{i=1}^{n} J'^2_i H(\theta_{(i)}) - 2\sum_{i=2}^{n}\left[J'_i H(\theta_{(i)})\sum_{k=1}^{i-1} J'_k\right].$$

Finally, results similar to that given in Corollary 5.2.4 can also be derived for the Cramér-von Mises distance.

Lemma 5.4.3 *If $P \sim DP(a, H)$, then*

(i) $d_{CvM}\left(P_n^{F.D.}, H\right) \stackrel{a.s.}{\to} d_{CvM}\left(P, H\right)$, *where $P_n^{F.D.}$ is the finite-dimensional Dirichlet prior defined by (2.1.13).*

(ii) $d_{CvM}\left(P_n^{new}, H\right) \stackrel{a.s.}{\to} d_{CvM}\left(P^{Ferg.}, H\right)$, *where P_n^{new} is defined in (4.1.6) and $P_n^{Ferg.}$ is the Ferguson representation of the Dirichlet process defined by (2.1.4).*

(iii) $d_{CvM}\left(P_n^{Seth.}, H\right) \stackrel{a.s.}{\to} d_{CvM}\left(P, H\right)$, *where $P_n^{Seth.}$ is the truncated Sethuraman representation of the Dirichlet process defined by (2.1.11) and $P^{Seth.}$ is the Sethuraman representation of the Dirichlet process defined by (2.1.10).*

Proof: We only prove (i). The other parts of the lemma follow similarly. Notice that

$$d_{CvM}(P_n^{\mathrm{F.D.}}, H) \;=\; \int_{-\infty}^{\infty} \left(P_n^{\mathrm{F.D.}}(x) - H(x)\right)^2 h(x)dx.$$

By Theorem 2.1.11, there exists a set $\widetilde{\Omega}$ with $\Pr(\widetilde{\Omega}) = 1$ such that $P_n^{\mathrm{F.D.}} \stackrel{v}{\to} P$, where $\stackrel{v}{\to}$ denotes vague convergence (see Appendix B). Hence for every $\omega \in \Omega$, $P_n^{\mathrm{F.D.}}(\omega, (-\infty, x]) \to P(\omega, (-\infty, x])$ for every continuity point x of $P(\omega, \cdot)$. Since the set of discontinuity points is at most countable and any countable set has measure zero (Resnick 1998, page 248), $P_n^{\mathrm{F.D.}}(\omega, x) \to P(\omega, x)$ almost everywhere x for any $\omega \in \widetilde{\Omega}$ fixed. Note that $\left(P_n^{\mathrm{F.D.}}(\omega, x) - H(x)\right)^2 \leq 1$ and $h(x)$ is integrable as $h(x)$ is a probability density function. By the dominated convergence theorem, we obtain

$$\int_{-\infty}^{\infty} \left(P_n^{\mathrm{F.D.}}(\omega, x) - H(x)\right)^2 h(x)dx \to \int_{-\infty}^{\infty} \left(P_n(\omega, x) - H(x)\right)^2 h(x)dx,$$

for every $\omega \in \widetilde{\Omega}$. ∎

Chapter 6

Applications

This chapter is divided into four sections. In the first section, a suitable concentration parameter of the Dirichlet process is selected. In the second and the third sections, we examine a Bayesian goodness of fit test for a simple and composite hypothesis for non-censored observations, respectively. In the last section, we discuss a Bayesian goodness of fit test for right censored data.

6.1 The Concentration Parameter

The use of the Dirichlet process requires a selection of a concentration parameter and a base measure. In the literature, the concentration parameter of the Dirichlet process is selected with almost no justification for the selection. Thus, for a fixed base measure H, it is important to assess the selection of the concentration parameter. One way to attain this is through computing one of the following probabilities:

$$\Pr\left\{d_K\left(P,H\right)\leq\epsilon\right\},\qquad \Pr\left\{d_L\left(P,H\right)\leq\epsilon\right\},\quad \text{or}\quad \Pr\left\{d_{CvM}\left(P,H\right)\leq\epsilon\right\}$$

from some $\epsilon\in(0,1)$ and comparing the value with a fixed number q, where $0<q<1$. Here q is chosen based on the prior belief. By Lemma 5.3.1, the algorithms for estimating the probabilities $\Pr\left\{d_K\left(P,H\right)\leq\epsilon\right\}$ and $\Pr\left\{d_L\left(P,H\right)\leq\epsilon\right\}$ are similar. Thus, we

only consider the probability of the Kolmogorov distance and the Cramér-von Mises distance. To compute such probabilities a Monte Carlo approach is used. Throughout this chapter, let P_n be any of the representations $P_n^{\text{Seth.}}$, $P_n^{\text{F.D.}}$, or P_n^{new} defined by (2.1.11), (2.1.13) and (4.1.6), respectively.

Algorithm J: Assessing the concentration parameter. Given the concentration parameter a, we can assess whether our choice of a is reasonable by applying the following steps:

(1) Approximate the probability $\Pr\{d_K(P, H) \leq \epsilon\}$ as follows:

   (a) Generate a random sample from P_n. In particular, when $P_n = P_n^{\text{new}}$ apply Algorithm C in Subsection 4.1.2.

   (b) Compute $d_K(P_n, H)$ as defined in Propositions 5.2.2 or 5.2.7.

   (c) Repeat steps (a) and (b) to obtain r i.i.d. samples of $d_K(P_n, H)$. For large n and r, the empirical distribution of these values is an approximation to the distribution of $d_K(P, H)$. Hence, the required probability is estimated by the proportion of $d_K(P_n, H)$ that are less than or equal to ϵ.

(2) Based on the estimated probability from (c) and whether it is closed from q where $0 < q < 1$, we decide if the selection of the concentration parameter is reasonable. Note that, if we judge that the selected concentration parameter is not an appropriate one (i.e. the approximated probability is more (less) than q), then one can increase or decrease the value of the concentration parameter to reach the value of q.

Remark 6.1.1 *The steps for approximating* $\Pr\{d_{CvM}(P, H) \leq \epsilon\}$ *are similar to that of the Kolmogorov distance. The distance* $d_{CvM}(P_n, H)$ *is computed either by applying Lemma 5.4.1 or Remark 5.4.2.*

Example 6.1.2 In this example, we want to simulate the distances $d_K(P_n, H)$ and $d_{CvM}(P_n, H)$ and the probabilities $\Pr\{d_K(P_n, H) \leq \epsilon\}$ and $\Pr\{d_{CvM}(P_n, H) \leq \epsilon\}$. In the calculations, we set $n = 2000$, $r = 3000$, and $H = N(0, 1)$ with different values of a and ϵ. The results are reported in Tables 6.1-6.4.

Table 6.1: Estimates for 95% prediction interval (P. I.) for the Kolmogorov prior distance. The 95% P. I. is obtained by eliminating the 2.5% of the lowest and highest distance. In this table $d_K(P_n^{\text{Seth.}}, H)$, $d_K(P_n^{\text{F.D.}}, H)$, and $d_K(P_n^{\text{new}}, H)$ stand for 95% P. I. of $d_K(P_n^{\text{Seth.}}, H)$, 95% P. I. of $d_K(P_n^{\text{F.D.}}, H)$, and 95% P. I. of $d_K(P_n^{\text{new}}, H)$, respectively.

a	$d_K(P_n^{\text{Seth.}}, H)$	$d_K(P_n^{\text{F.D.}}, H)$	$d_K(P_n^{\text{new}}, H)$
0.1	$(0.4319, 0.9807)$	$(0.4264, 0.9776)$	$(0.4335, 0.9804)$
1	$(0.2687, 0.8510)$	$(0.2681, 0.8816)$	$(0.2664, 0.8643)$
30	$(0.0764, 0.2527)$	$(0.0781, 0.2608)$	$(0.0785, 0.2582)$
50	$(0.0617, 0.2013)$	$(0.0618, 0.2015)$	$(0.0625, 0.2037)$
106	$(0.0437, 0.1387)$	$(0.0451, 0.1429)$	$(0.0443, 0.1442)$

Table 6.2: Estimates for 95% prediction interval (P. I.) for the Cramér-von Mises distance. The 95% P. I. is obtained by eliminating the 2.5% of the lowest and highest distance. In this table $d_{CvM}(P_n^{\text{Seth.}}, H)$, $d_{CvM}(P_n^{\text{F.D.}}, H)$, and $d_{CvM}(P_n^{\text{new}}, H)$ stand for 95% P. I. of $d_{CvM}(P_n^{\text{Seth.}}, H)$, 95% P. I. of $d_{CvM}(P_n^{\text{F.D.}}, H)$, and 95% P. I. of $d_{CvM}(P_n^{\text{new}}, H)$, respectively.

a	$d_{CvM}(P_n^{\text{Seth.}}, H)$	$d_{CvM}(P_n^{\text{F.D.}}, H)$	$d_{CvM}(P_n^{\text{new}}, H)$
0.1	$(0.0443, 0.3120)$	$(0.0436, 0.3139)$	$(0.0414, 0.3184)$
1	$(0.0148, 0.2429)$	$(0.0149, 0.2391)$	$(0.0137, 0.2364)$
30	$(0.0010, 0.0181)$	$(0.0010, 0.0192)$	$(0.0010, 0.0192)$
50	$(0.0006, 0.1117)$	$(0.0006, 0.0123)$	$(0.0006, 0.0122)$
106	$(0.0003, 0.0052)$	$(0.0003, 0.0059)$	$(0.0003, 0.0058)$

Table 6.3: Estimates for the Kolmogorov probability. In this table, In this table, $\Pr\left\{d_K^{\text{Seth.}} \leq \epsilon\right\} = \Pr\left\{d_K\left(P_n^{\text{Seth.}}, H\right) \leq \epsilon\right\}$, $\Pr\left\{d_K^{\text{F.D.}} \leq \epsilon\right\} = \Pr\left\{d_K\left(P_n^{\text{F.D.}}, H\right) \leq \epsilon\right\}$ and $\Pr\left\{d_K^{\text{new}} \leq \epsilon\right\} = \Pr\left\{d_K\left(P_n^{\text{new}}, H\right) \leq \epsilon\right\}$.

a	ϵ	$\Pr\left\{d_K^{\text{Seth.}} \leq \epsilon\right\}$	$\Pr\left\{d_K^{\text{F.D.}} \leq \epsilon\right\}$	$\Pr\left\{d_K^{\text{new}} \leq \epsilon\right\}$
0.1	0.04	0.0000	0.0000	0.0000
	0.08	0.0000	0.0000	0.0000
1	0.04	0.0000	0.0000	0.0000
	0.08	0.0000	0.0000	0.0000
30	0.04	0.0000	0.0000	0.0000
	0.08	0.0310	0.0303	0.0330
50	0.04	0.0000	0.0000	0.0000
	0.08	0.1463	0.1353	0.1410
106	0.04	0.0000	0.0000	0.0000
	0.08	0.5500	0.4980	0.5110

Table 6.4: Estimates for the Cramér-von Mises distance probability. In this table, $\Pr\left\{d_{CvM}^{\text{Seth.}} \leq \epsilon\right\} = \Pr\left\{d_{CvM}\left(P_n^{\text{Seth.}}, H\right) \leq \epsilon\right\}$, $\Pr\left\{d_{CvM}^{\text{F.D.}} \leq \epsilon\right\} = \Pr\left\{d_{CvM}\left(P_n^{\text{F.D.}}, H\right) \leq \epsilon\right\}$ and $\Pr\left\{d_{CvM}^{\text{new}} \leq \epsilon\right\} = \Pr\left\{d_{CvM}\left(P_n^{\text{new}}, H\right) \leq \epsilon\right\}$.

a	ϵ	$\Pr\left\{d_{CvM}^{\text{Seth.}} \leq \epsilon\right\}$	$\Pr\left\{d_{CvM}^{\text{F.D.}} \leq \epsilon\right\}$	$\Pr\left\{d_{CvM}^{\text{new}} \leq \epsilon\right\}$
0.1	0.004	0.0000	0.0000	0.0000
	0.008	0.1037	0.0937	0.1090
1	0.004	0.0000	0.0000	0.0000
	0.008	0.5753	0.5877	0.5977
30	0.004	0.5160	0.5013	0.4920
	0.008	1.0000	1.0000	1.0000
50	0.004	0.7347	0.7360	0.7097
	0.008	1.0000	1.0000	1.0000
106	0.004	0.9417	0.9240	0.9260
	0.008	1.0000	1.0000	1.0000

Tables 6.1-6.4 show that, for large values of a, the value of the distance decreases toward 0 (equivalently, the value of the probability increases toward 1). On the other hand, for small values of a, the value of the distance increases toward 1 (equivalently, the value of the probability decreases toward 0). It is also clear from the tables that the distances obtained by using different approximations of the Dirichlet process are quite similar, as n is large. However, with the same concentration parameter a, the computed Cramér-von Mises distance is much less than that of the Kolmogorov distance.

6.2 A Goodness of Fit Test for Non-censored Data: A Simple Hypothesis

The problem considered in this section is to test the null hypothesis $\mathcal{H}_0 : \ P = H$, where P is the true underlying distribution, H is a completely specified continuous distribution (such as $N(0,1)$), and the data consist of a sample $X_1, \ldots, X_m$ from P. In the literature, there are two different Bayesian approaches for goodness of fit tests. The first one consists of embedding the proposed model in the null hypothesis into a larger family of models, the alternative family. After that, a prior is placed on the alternative family. Then, the Bayes factor of the null hypothesis to the alternative is computed. For example, Folrens, Richard, and Rolin (1996) used a Dirichlet process prior for the alternative distribution. Another form of prior, Polya tree process (Lavine 1992), was considered by Berger and Gugliemli (2001). The other approach of goodness of fit tests focuses on measuring the distance of the null distribution from the true underlying distribution. This approach was initiated by Muliere and Tardella (1998) and discussed in more details in Swartz (1999). In this chapter, we pursue the approach of Muliere and Tardella (1998) and Swartz (1999). In these two papers, the authors considered the Kolmogorov distance and used the truncated stick-breaking representation to simulate the Dirichlet process. Explicit expressions for calculating the distance were not provided in these two papers. The derived closed formulas of the distance make it simple to apply this approach. On the other hand, considering other distances such as the Cramér-von Mises distance makes this approach more efficient. See the forthcoming examples for more details. Since calculating the distance depends on an approximation of the Dirichlet process, using approximation methods of the Dirichlet process other than the truncated stick-breaking representation will improve this approach. For instance, since the variability among the weights in the new representation (4.1.6) is much less than that of the stick-breaking representation,

the results obtained by using this approximation are much less affected by the random factor associated with the weights of the Dirichlet process. Hence, more robust conclusions are obtained by using this representation. In addition, since the rate of convergence of the new representation is very high, less computational effort is needed when calculating the distance.

To construct a goodness of fit test, it is required to calculate the posterior distance. Let P_n^* be an approximation for the posterior of the Drichlet process P given the data. The posterior distribution P_n^* is obtained from the Dirichlet process with parameters given in Theorem 2.1.4. Let $(\theta_k^*)_{1\leq k\leq n}$ be the jump points of P_n^*. Under the null hypothesis $\mathcal{H}_0$, $(\theta_k^*)_{1\leq k\leq n}$ are i.i.d random variables with common distribution H. Thus, all the distance formulas derived in Chapter 5 can be applied for the posterior distribution.

Algorithm K: A simple goodness of fit test for non-censored data. The following steps are required for a goodness of fit test:

(1) For a given distribution function H, find an appropriate value of ϵ as follows (Swartz, 1999):

   (a) From the data, find the maximum value x_0 such that a measurement $x \pm x_0$ could be considered nearby x (x_0 is known as the measurement precision).

   (a) The required ϵ is:

$$\epsilon = \max_{x\in\mathbb{R}}(H(x+x_0) - H(x)). \tag{6.2.1}$$

(2) Find an appropriate value of the concentration parameter a such that

$$\Pr\left\{d_K\left(P_n, H\right) \leq \epsilon\right\} = q,$$

where $0 < q < 1$. This can be done by using the steps mentioned in Algorithm J in the previous section. Here q represents the prior belief that the underlying distribution P is practically equivalent to H. Usually we take $q = 0.5$.

(3) Generate a random sample from P_n^*.

(4) Compute $d_K\left(P_n^*, H\right)$ as given in Propositions 5.2.2 or 5.2.7.

(5) Repeat steps 3 and 4 to obtain r i.i.d. samples of $d_K\left(P_n^*, H\right)$. For large n and r, the empirical distribution of these values is an approximation to the distribution of $d_K(P^*, H)$. Hence, the probability $\Pr\left\{d_K\left(P^*, H\right) \leq \epsilon\right\}$ is estimated by the proportion of $d_K(P_n^*, H)$ that are less than or equal to ϵ. Based on the estimated probability we decide whether the true underlying distribution P is in proximity of the hypothesized distribution H. That is, if the value of the estimated posterior probability is greater than the prior probability q, then there is a good evidence not to reject $\mathcal{H}_0$ (as this means that data come from H). Otherwise, we reject the null hypothesis $\mathcal{H}_0$.

Remark 6.2.1 *To apply the Cramér-von Mises distance in the previous algorithm, one should replace (6.2.1) by*

$$\epsilon = \int_{-\infty}^{\infty} \left(H(x + x_0) - H(x)\right)^2 h(x)dx. \tag{6.2.2}$$

Example 6.2.2 We consider the data generated from a standard normal distribution of size $m = 200$. We test if $H = N(0, 1)$. First, we consider the Kolmogorov distance. We start by determining the value of the measurement precision x_0. For the generated data we set $x_0 = 0.2$. Applying (6.2.1), we get $\epsilon = 0.08$. Next we compute the concentration parameter a so that

$$\Pr\left\{d_K\left(P_n^{\text{Seth.}}, H\right) \leq \epsilon\right\} = \Pr\left\{d_K\left(P_n^{\text{F.D.}}, H\right) \leq \epsilon\right\} = \Pr\left\{d_K\left(P_n^{\text{new}}, H\right) \leq \epsilon\right\} = 0.5.$$

For $\epsilon = 0.08$, an appropriate a is 106 (see Table 6.3). Applying Algorithm K, we obtain the corresponding estimated posterior probabilities (Sethuraman, Finite-Dimensional and New) are 0.8900, 0.8530 and 0.8417, respectively. Therefore, we cannot reject the hypothesis that $H = N(0, 1)$.

Now we consider the Cramér-von Mises distance. For $x_0 = 0.2$, by applying (6.2.2), we obtain $\epsilon = 0.004$. We find the concentration parameter a so that

$$\begin{aligned}\Pr\left\{d_{CvM}\left(P_n^{\text{Seth.}}, H\right) \leq \epsilon\right\} = \Pr\left\{d_{CvM}\left(P_n^{\text{F.D.}}, H\right) \leq \epsilon\right\} &= \Pr\left\{d_{CvM}\left(P_n^{\text{new}}, H\right) \leq \epsilon\right\} \\ &= 0.5.\end{aligned}$$

For $\epsilon = 0.004$, an appropriate a is 30 (see Table 6.4). The estimated posterior probabilities (Sethuraman, Finite-Dimensional and New) are 0.9907, 0.9843 and 0.9867, respectively. Therefore, we cannot reject the hypothesis that $H = N(0, 1)$. See also Figure 6.1 and Figure 6.2 which give the plot of H and 5 sample paths for each of the prior Dirichlet process and the posterior Dirichlet process. It is clear from the two figures that the plots of the sample paths for the posterior process move toward the plot of the cumulative distribution function H, which supports our previous conclusion about the null hypothesis.

In the previous example, standard (frequentist) goodness of fit tests such as the frequentist Kolmogorov-Smirnov test give p-value near 0.7605, from which we conclude not to reject the null hypothesis. See also the qq-plot given in Figure 6.3.

Example 6.2.3 We consider the data generated from a standard Cauchy distribution of size $m = 200$. We want to test if $H = N(0, 1)$. For the same x_0 and ϵ obtained in Example 6.2.2, the estimated posterior probabilities (Sethuraman, Finite-Dimensional and New) for the Kolmogorov distance are equal to 0.0266, 0.0253 and 0.0283, respectively. On the other hand, the estimated posterior probabilities for the Cramér-von Mises distance are equal to 0.1463, 0.1557 and 0.1617, respectively. Thus, the normality hypothesis of the data is rejected. See also Figure 6.4 and Figure 6.5 which give the plot of H and 5 sample paths for each of the prior Dirichlet process and the posterior Dirichlet process. It is clear from the figures that the plots of the sample paths for the posterior process depart from the plot of the cumulative distribution function H. This supports the previous conclusion about the null hypothesis. Our

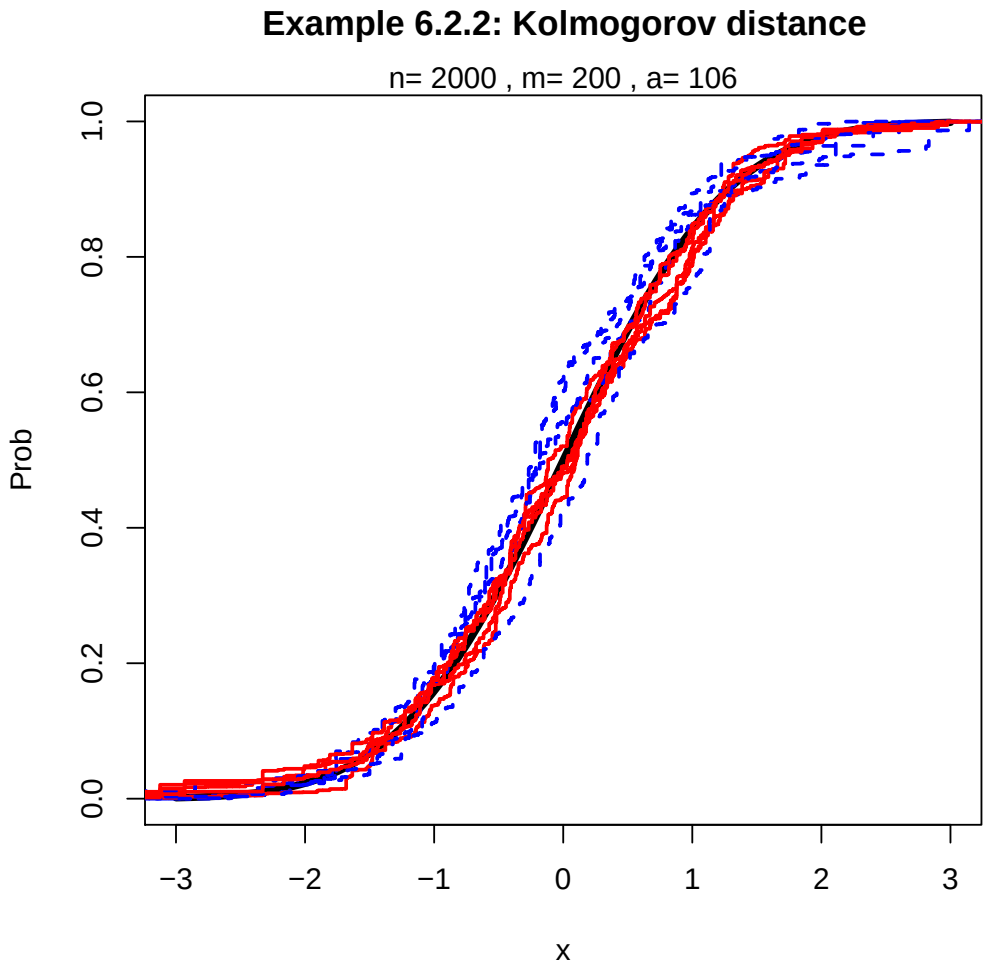

Figure 6.1: *The solid line (thick) represents the plot of the cumulative distribution function* $H = N(0,1)$*, the dashed lines represent sample paths of the prior Dirichlet process process and the other lines represent sample paths of the posterior Dirichlet process.*

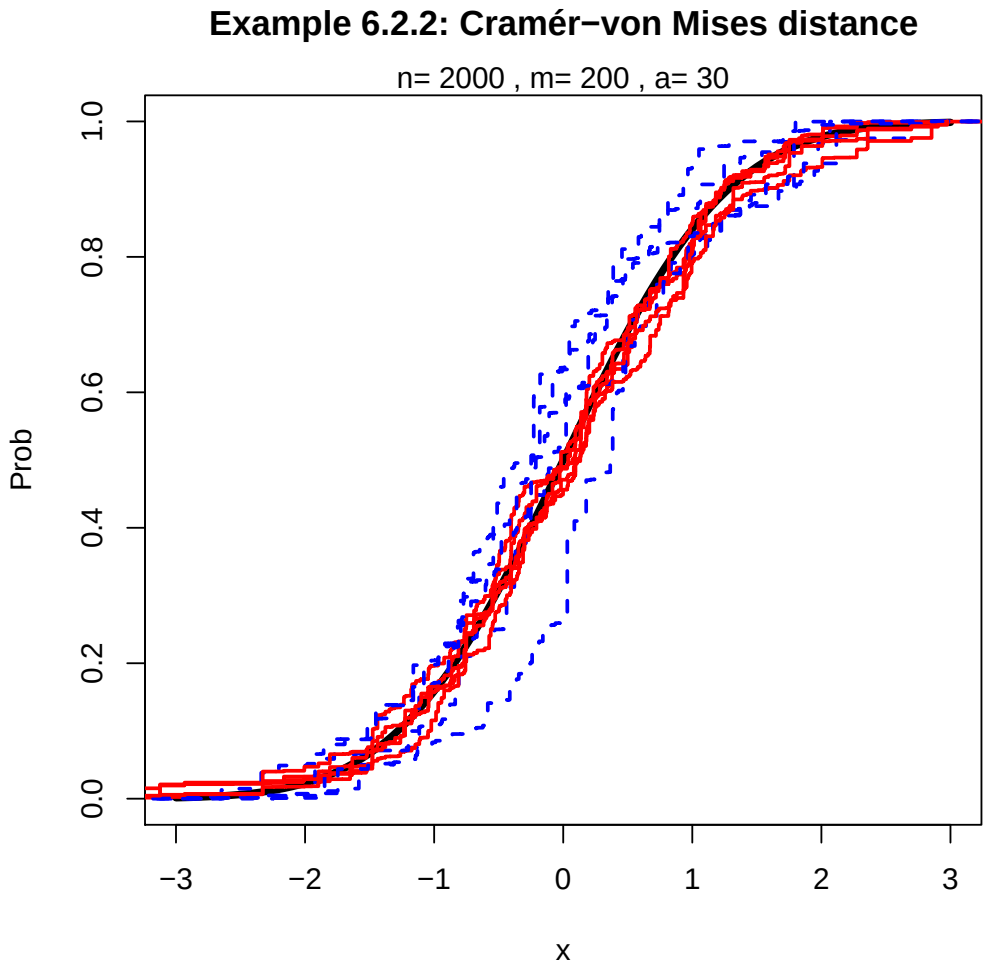

Figure 6.2: *The solid (thick) line represents the plot of the cumulative distribution function* $H = N(0,1)$ *the dashed lines represent sample paths of the prior Dirichlet process process and the other lines represent sample paths of the posterior Dirichlet process.*

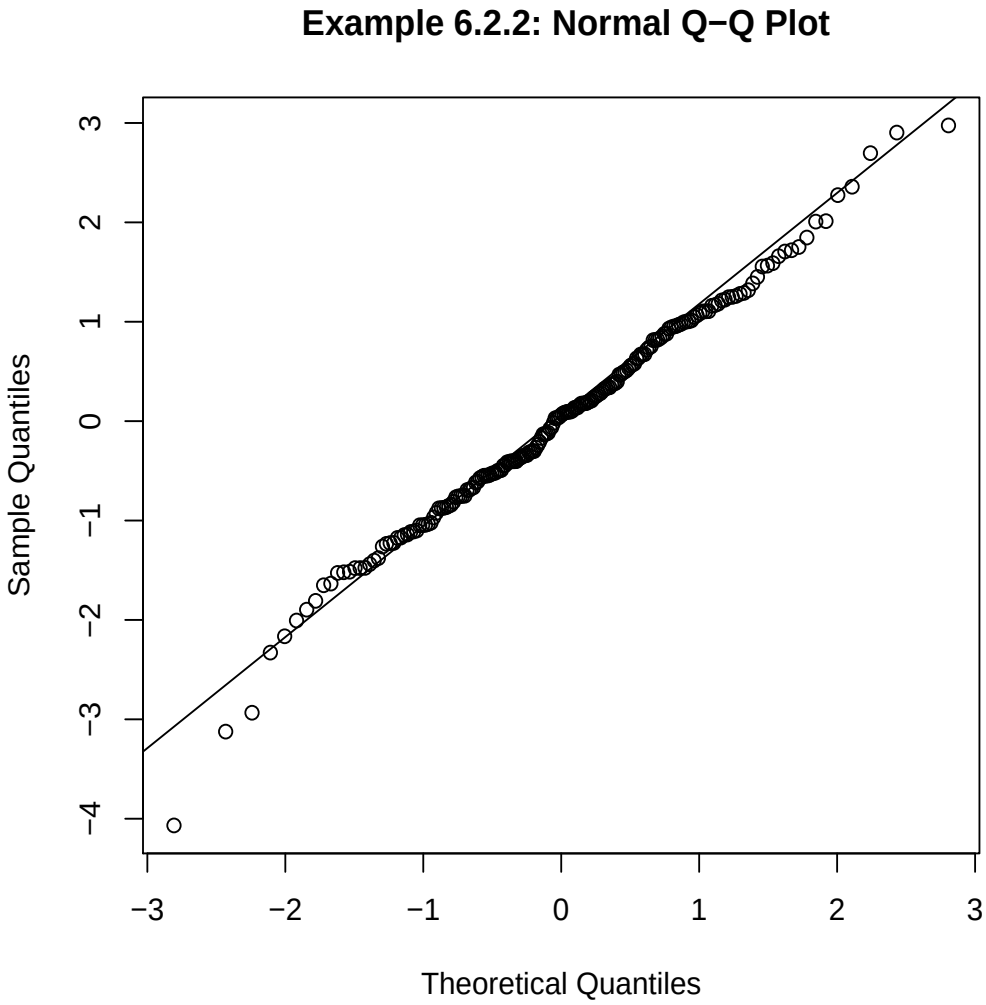

Figure 6.3: *qq plots for a data from normal distribution.*

conclusion is consistent with the frequentist Kolmogorov-Smirnov test which gives a p-value of 0.0003 and with the qq-plot given in Figure 6.6.

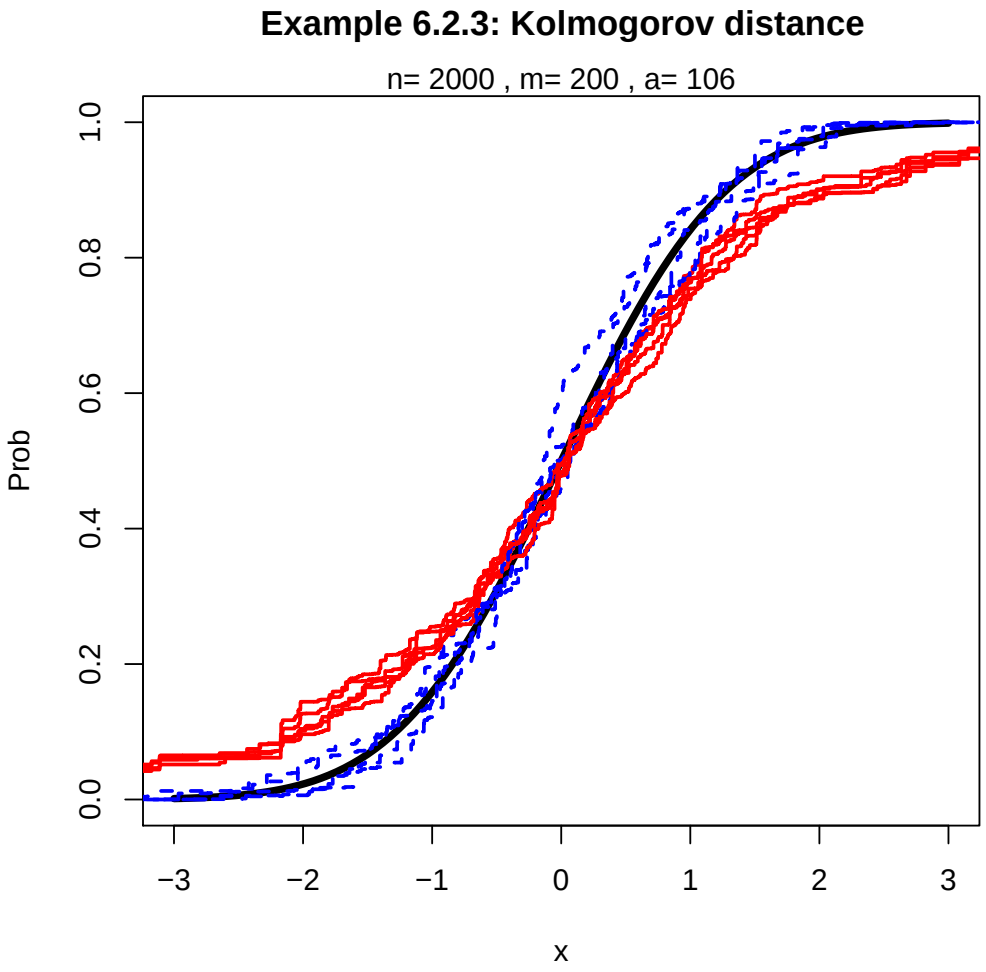

Figure 6.4: *The solid (thick) line represents the plot of the cumulative distribution function $H = N(0,1)$ the dashed lines represent sample paths of the prior Dirichlet process process and the other lines represent sample paths of the posterior Dirichlet process.*

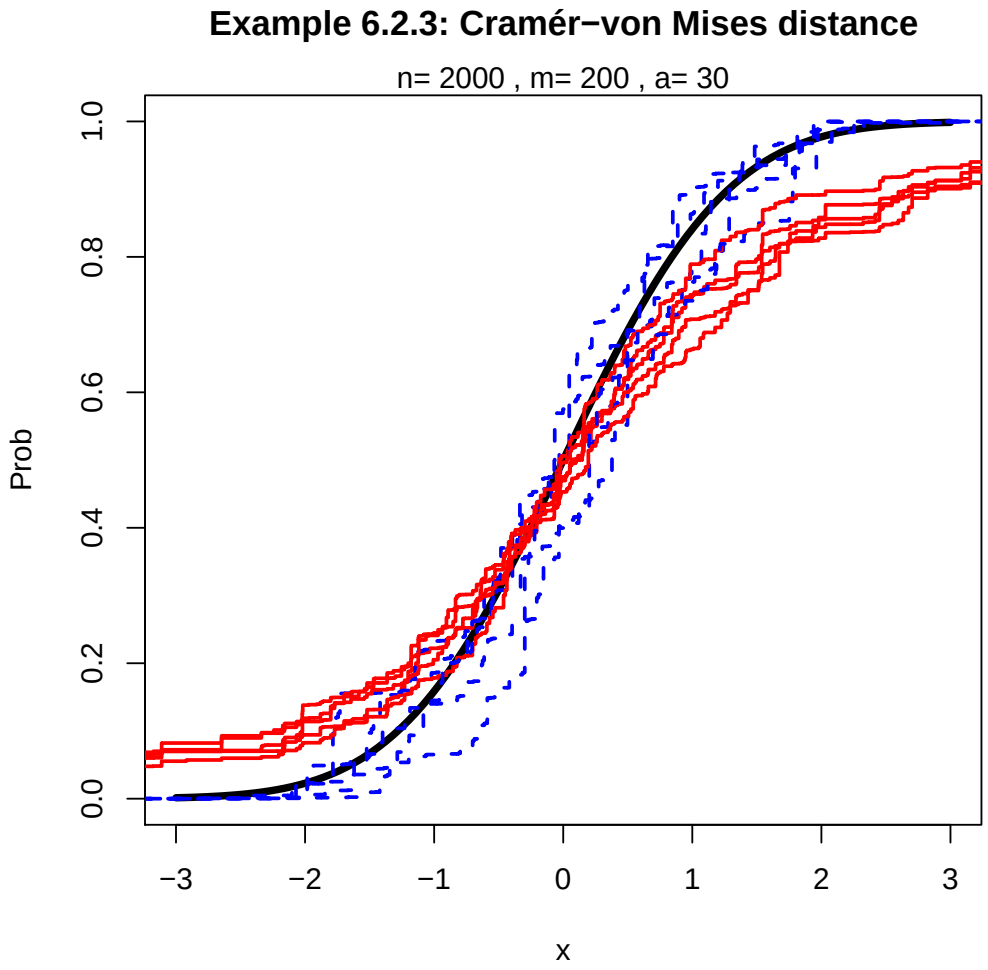

Figure 6.5: *The solid (thick) line represents the plot of the cumulative distribution function* $H = N(0,1)$ *the dashed lines represent sample paths of the prior Dirichlet process process and the other lines represent sample paths of the posterior Dirichlet process.*

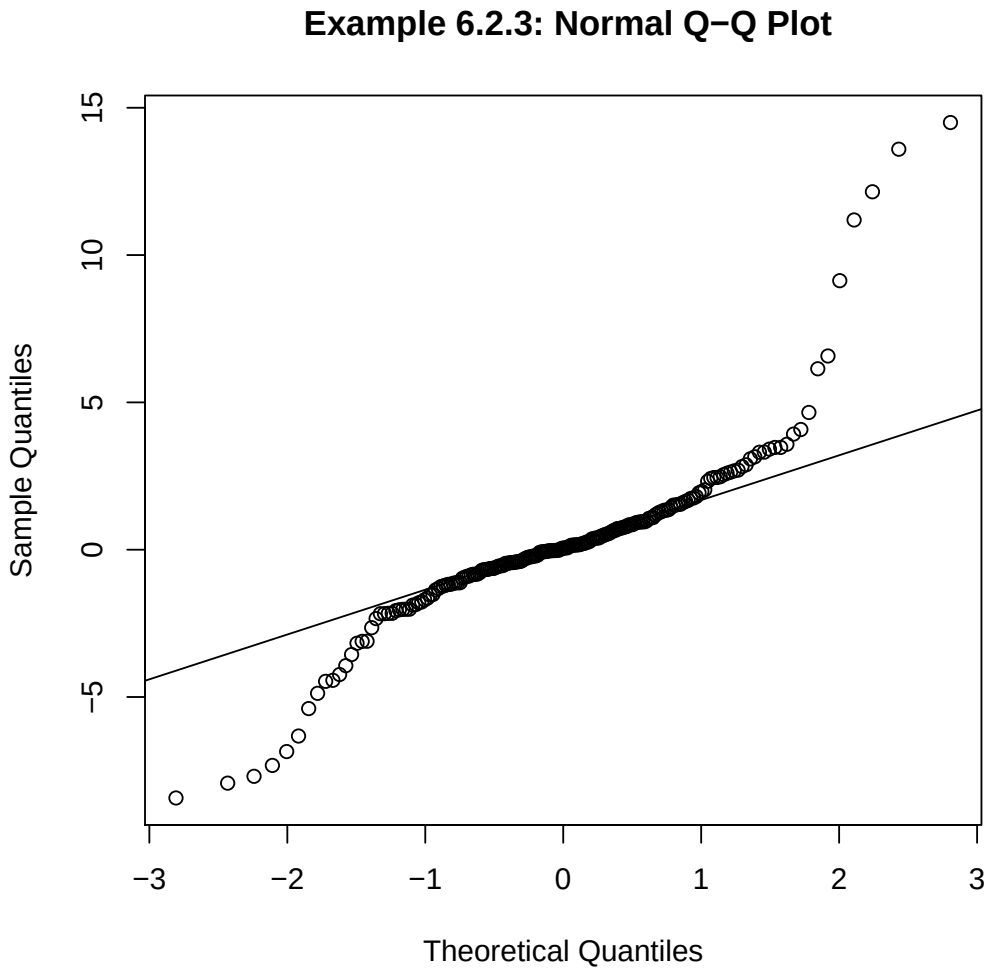

Figure 6.6: *qq plots for a data from Cauchy distribution.*

Observe that, in the previous two examples, for the same measurement precision x_0, the values of ϵ and a obtained using the Cramér-von Mises distance are much less than these obtained using the Kolmogorov distance. When the concentration parameter a is large, to get a more precise approximation of the Dirichlet process, we need to consider more terms in this approximation. This increases the computational time. Thus, applying the Cramér-von Mises distance leads to faster results.

6.3 A Goodness of Fit Test: A Composite Hypothesis

The approach used in testing a simple hypothesis can be generalized to the composite hypothesis $\mathcal{H}_0 : \ P = H_\theta$, where H_θ is a member of a family of distributions indexed by a finite dimensional parameter θ which has a prior distribution $\pi(\theta)$.

Algorithm L: A composite goodness of fit test for non-censored data. The following algorithm is used to test a composite null hypothesis:

(1) Choose ϵ as explained in the previous subsection (Algorithm K) with formula (6.2.1) replaced by (Swartz, 1999)

$$\epsilon = \max_{x \in \mathbb{R}} (H_{\widehat{\theta}}(x + x_0) - H_{\widehat{\theta}}(x)), \tag{6.3.1}$$

where $\widehat{\theta} = E(\theta)$.

(2) Find an appropriate value of a such that $\Pr\left\{d_K\left(P_n, H_{\widehat{\theta}}\right) \leq \epsilon\right\} = q$, where $0 < q < 1$. Here, q represents the prior belief that the underlying distribution P is equivalent to the specified distribution $H_{\widehat{\theta}}$. This can be done similar to Algorithm J in Section 6.1. Note that the choice of a in this step is independent of H and θ (see, for example, Proposition 5.2.7).

(3) Generate R realizations $\left(\theta_0^{(i)}\right)_{1 \leq i \leq R}$ from the distribution of θ given $X_1, \ldots, X_n$, where the density of this distribution is given by

$$f(\theta \mid X_1, \ldots, X_n) \propto \left(\prod_{k=1}^{n} h_\theta(X_i)\right) \pi(\theta),$$

where h_θ is the probability density function corresponding to H_θ and $\pi(\theta)$ is the prior distribution of θ. Note that sampling from nonstandard distributions is done by applying the Metropolis-Hastings algorithm. For more details about this

algorithm, see Robert and Casella (1999). Here each realization $\theta_0^{(i)}$ is obtained by using a different Metropolis-Hastings chain and it represents the final variate obtained from each chain.

(4) Find $\theta_{\text{opt}} = \arg\min_{\theta_0^{(i)}} d_K\left(P_n^*, H_{\theta_0^{(i)}}\right)$, for $i = 1, \ldots, R$, where P_n^* is an approximation for the posterior process given the data with the base measure $H^*_{\theta_0^{(i)}}$ as described in (2.1.2) with H replaced by $H_{\theta_0^{(i)}}$.

(5) Compute $d_K\left(P_n^*, H_{\theta_{\text{opt.}}}\right)$ and follow the steps of Algorithm K in the previous subsection with H replaced by $H_{\theta_{\text{opt.}}}$.

Remark 6.3.1 *To apply the Cramér-von Mises distance in the previous algorithm, one should replace (6.3.1) by*

$$\epsilon = \int_{-\infty}^{\infty} \left(H_{\widehat{\theta}}(x + x_0) - H_{\widehat{\theta}}(x)\right)^2 h(x)dx.$$

The above algorithm is different from the approach used in Swartz (1999) for composite hypothesis. Instead of finding $\theta_{\text{opt.}}$, Swartz used different values of θ when computing the distance.

Example 6.3.2 This example is taken from Hamada, Wilson, Reese and Martz (2008, Example 3.6). This example analyzes the lifetimes of $m = 31$ liquid crystal display (LCD) projector lamps. We test the null hypothesis:

$\mathcal{H}_0$: the lifetimes distribution of the observed data is exponential with mean $1/\theta$,

where $\theta > 0$ and θ has a $Gamma(1.7, 2\,550)$ prior distribution (i.e. $\pi(\theta)$ is $Gamma(1.7, 2550)$ distribution). The posterior distribution of θ is a gamma with mean $32.7/20457$ and variance $32.7/20457^2$. We consider first the Kolmogorov distance. Set $x_0 = 50$ in (6.2.1). This gives $\epsilon = 0.08$. For $q = 0.5$ an appropriate a is 106. As described in Algorithm L, sampling from the distribution of θ given $X_1, \ldots, X_n$ is carried out via

the Metropolis-Hastings algorithm. In details, we start by generating an arbitrary point $\theta^{(0)} \sim \pi(\theta)$. Then, for $i = 1, \ldots, 1000$, generate a candidate point $\theta^{(i)} \sim \pi(\theta)$ and u_i from a uniform distribution on $[0, 1]$. If

$$\begin{aligned} u_i &< \min\left(\frac{f(\theta^{(i)} \mid X_1, \ldots, X_n)\pi(\theta^{(i-1)})}{f(\theta^{(i-1)} \mid X_1, \ldots, X_n)\pi(\theta^{(i)})}, 1\right) \\ &= \min\left(\left(\frac{\theta^{(i)}}{\theta^{(i-1)}}\right)^n \exp\left\{-\sum_{k=1}^{n} X_k\left(\frac{1}{\theta^{(i)}} - \frac{1}{\theta^{(i-1)}}\right)\right\}, 1\right), \end{aligned}$$

set $\theta^{(i+1)} = \theta^{(i)}$. Otherwise set $\theta^{(i+1)} = \theta^{(i-1)}$. We take $\theta_0^{(1)} = \theta^{(1000)}$ as a realization from the distribution of θ given $X_1, \ldots, X_n$. We repeat the previous step 1000 times to obtain $\left(\theta_0^{(i)}\right)_{1 \le i \le 1000}$ realizations from the distribution of θ given $X_1, \ldots, X_n$. We find $\theta_{\text{opt.}} = \arg\min_{\theta_0^{(i)}} d_K\left(P_n^*, H_{\theta_0^{(i)}}\right)$, for $i = 1, \ldots, 1000$. Then we calculate $d_K\left(P_n^*, H_{\theta_{\text{opt.}}}\right)$. We repeat the calculations 3000 times where each time a new sample form P_n^* is used. The estimated posterior probabilities are reported in Table 6.5. We also consider the Cramér-von Mises distance. For $x_0 = 50$ and $q = 0.5$ we obtain $\epsilon = 0.002$ and $a = 67$. The estimated posterior probabilities are reported in Table 6.6. It follows from the tables that $\mathcal{H}_0$ is not rejected.

Table 6.5: Example 6.3.2: Testing composite hypothesis using the Kolmogorov distance.

P_n^*	$\theta_{\text{opt.}}$	$Pr\left\{d_K\left(P_n^*, H_{\theta_{\text{opt.}}}\right) \le \epsilon\right\}$
$P_n^{*\text{Seth.}}$	668.2181	0.6210
$P_n^{*\text{F.D.}}$	543.6369	0.6093
$P_{n\,n}^{*\text{new}}$	602.8773	0.6073

Table 6.6: Example 6.3.2: Testing composite hypothesis using the Cramér-von Mises distance.

P_n^*	$\theta_{\text{opt.}}$	$Pr\left\{d_{CvM}\left(P_n^*, H_{\theta_{\text{opt.}}}\right) \leq \epsilon\right\}$
$P_n^{*\text{Seth.}}$	717.2328	0.6463
$P_n^{*\text{F.D.}}$	701.9066	0.6443
$P_{n\,n}^{*\text{new}}$	535.695	0.6463

6.4 A Bayesian Nonparametric Goodness of Fit Test for Right Censored Data

As explained in Section 2.6, the Dirichlet process is not a conjugate prior for the survival time distribution when the sample contains right censored observations. So in the presence of right censored data, we use the beta-Stacy process and apply the method described in the previous section for a goodness fit test.

Let $X_1, \ldots, X_m$ be an i.i.d. sample from F. Suppose that

$$(T_1, \delta_1), \ldots, (T_m, \delta_m) \tag{6.4.1}$$

is observed, where $T_i = \min(X_i, C_i)$, $\delta_i = I\{X_i \leq C_i\}$ and $C_1, \ldots, C_m$ are censoring times. Clearly, $\delta_i = 1$ if X_i is observed, and $\delta_i = 0$ if X_i is right censored. The testing problem considered in this section is: $\mathcal{H}_0 : \ F = F_0$, where F_0 is a completely specified continuous distribution, based on a realization of the data in (6.4.1). In testing the previous null hypothesis, we assume that F is a beta-Stacy process as in Definition 2.6.2. The distance formulas derived in the previous chapter can be easily modified for the beta-Stacy process to measure the distance between the null distribution F_0 and the true underlying distribution F. Specifically, we need to replace P_n, an approximation of the Dirichlet process, by F_n, an approximation of the beta-Stacy

process. For instance, the Kolmogorov distance between F_n and F_0 is given in the following proposition. The proof is similar to the proof of Proposition 5.2.2 and is provided for the sake of completeness.

Proposition 6.4.1 *Let F_0 be a continuous distribution. Let $F_n(t) = 1-\exp\left(-Z_n(t)\right)$, where $Z_n(t) = \sum_{k=1}^{n} J_k \delta_{\theta_k}$ and $(\theta_k)_{1\leq k\leq n}$ are i.i.d. random variables with common distribution F_0. Then*

$$d_K(F_n, F_0) = \max\left\{d_K^{(1)}, d_K^{(2)}\right\},$$

where

$$d_K^{(1)} = \max_{1\leq i\leq n}\left|1 - \exp\left(-\sum_{k=1}^{i-1} J_k'\right) - F_0(\theta_{(i)})\right|$$

and

$$d_K^{(2)} = \max_{1\leq i\leq n}\left|1 - \exp\left(-\sum_{k=1}^{i} J_k'\right) - F_0(\theta_{(i)})\right|,$$

with the convention that $\sum_{k=1}^{0} J_k' = 0$.

Proof: Notice that

$$Z_n = \sum_{k=1}^{n} J_k' \delta_{\theta_{(k)}}.$$

Applying (5.2.1) we obtain

$$d_K(F_n, F_0) = \max\{d_K^{(1)}, d_K^{(2)}\},$$

where $d_K^{(1)} = \max_{1\leq i\leq n}|F_n(\theta_{(i-1)}) - F_0(\theta_{(i)})|$ and $d_k^{(2)} = \max_{1\leq i\leq n}|F_n(\theta_{(i)}) - F_0(\theta_{(i)})|$. Now, rewrite $d_K^{(1)}$ as follows:

$$\begin{aligned} d_K^{(1)} &= \max_{1\leq i\leq n}\left|1 - \exp\left(-\sum_{k=1}^{n} J_k' \delta_{\theta_{(k)}}\left(\left(-\infty, \theta_{(i-1)}\right]\right)\right) - F_0(\theta_{(i)})\right| \\ &= \max_{1\leq i\leq n}\left|1 - \exp\left(-\sum_{k=1}^{i-1} J_k'\right) - F_0(\theta_{(i)})\right|. \end{aligned}$$

Similarly, we get

$$d_K^{(2)} = \max_{1\leq i\leq n} \left| 1 - \exp\left(-\sum_{k=1}^{i-1} J_k' \right) - F_0(\theta_{(i)}) \right|.$$

This completes the proof. ∎

The algorithm needed to construct a Bayesian nonparametric goodness of fit test for right censored data is similar to Algorithms K and L given in the previous section with P_n, H, a and Proposition 5.2.2 are replaced by F_n, F_0, k and Proposition 6.4.1, respectively. Here we take $k(s) = k$, for all $s > 0$.

Example 6.4.2 This example is taken from Lee and Wang (2003, Example 9.3) and involves a set of remission times (in months) from 137 cancer patients. We test the following null hypothesis:

$\mathcal{H}_0$: the underlying distribution of the observed data is exponential with mean 10.

We start by finding the value of the measurement precision x_0. We set $x_0 = 1$. Applying (6.2.1) we get $\epsilon = 0.094$. Next we compute the concentration parameter k so that

$$\Pr\left\{ d_K\left(F_n, F_0\right) \leq \epsilon \right\} = 0.5.$$

For $\epsilon = 0.094$, an appropriate k is 73. The estimated posterior probability is 0.8363. Therefore, we cannot reject the null hypothesis. Figure 6.7 gives the plot of F_0 and 5 sample paths for each of the prior beta-Stacy process and the posterior beta-Stacy process. It is clear from this figure that the plots of the sample paths for the posterior process converge to the plot of the cumulative distribution function F_0, which supports our previous conclusion about the null hypothesis.

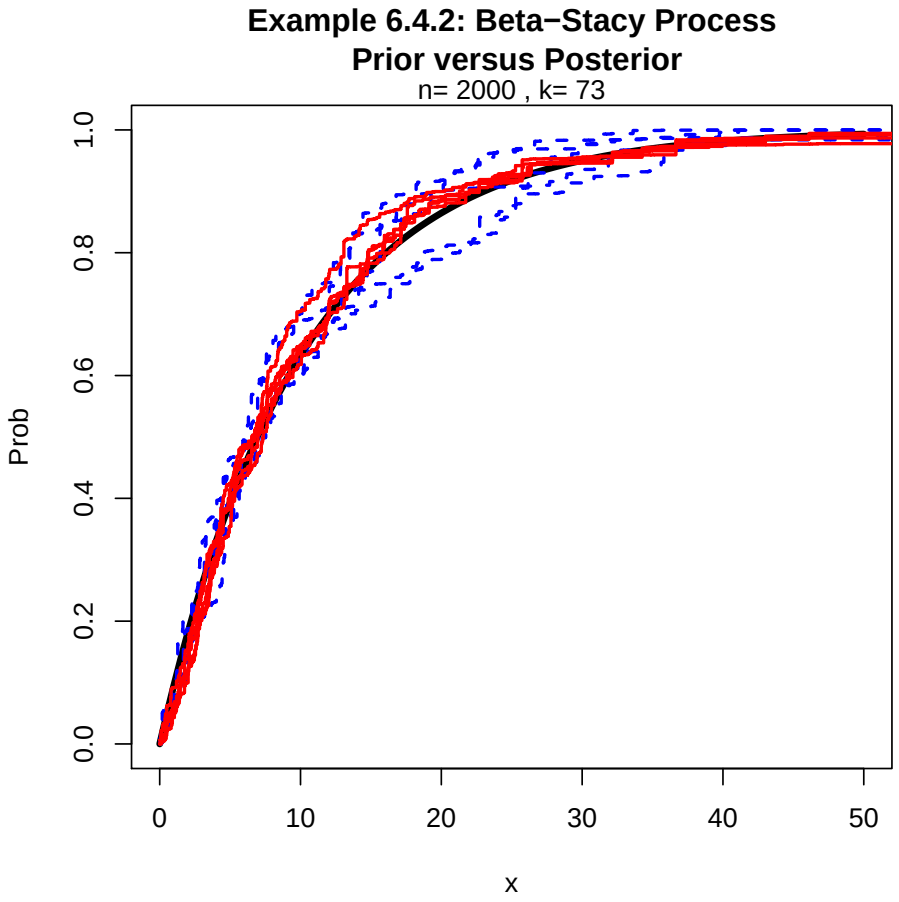

Figure 6.7: *The solid (thick) line represents the plot of the cumulative distribution function* $F_0(t) = 1 - \exp(-0.1t)$, *the dashed lines represent sample paths of the prior beta-Stacy process and the other lines represent sample paths of the posterior beta-Stacy process.*

Example 6.4.3 Suppose in Example 6.4.2 we are interested in testing the null hypothesis

$\mathcal{H}_0$: the underlying distribution of the observed data is Weibull with shape parameter 0.1 and scale parameter 1.

As in the previous example, we set $x_0 = 1$. We obtain $\epsilon = 0.094$ and $k = 73$. The

estimated posterior probability is 0. Hence the null hypothesis is rejected. Figure 6.8 gives the plot of F_0 and 5 sample paths for each of the prior beta-Stacy process and the posterior beta-Stacy process. It is clear from this figure that the plots of the sample paths for the posterior process deviate from the plot of the cumulative distribution function F_0. This supports our conclusion about the null hypothesis.

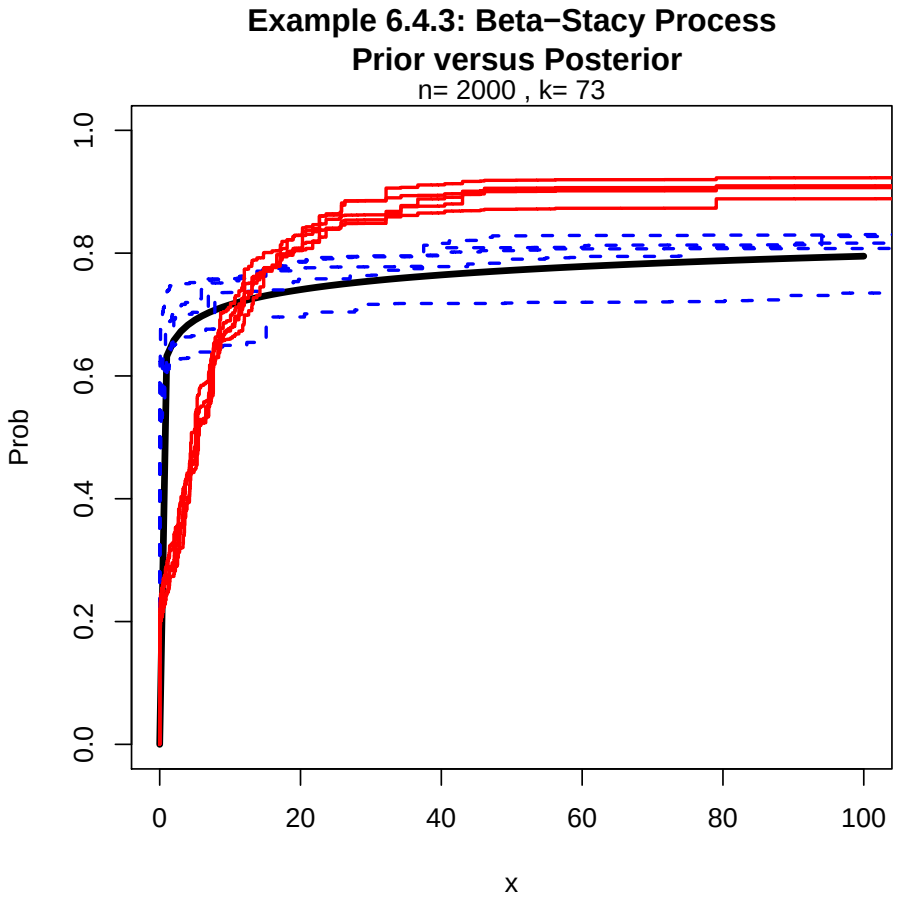

Figure 6.8: *The solid (thick) line represents the plot of the cumulative distribution function* $F_0(t) = 1 - \exp(-t^{0.1})$, *the dashed lines represent sample paths of the prior beta-Stacy process and the other lines represent sample paths of the posterior beta-Stacy process.*

Chapter 7

The Interplay of Frequentist and Bayesian Nonparametric Inference

In this chapter, we describe the striking relationship between the frequentist and the Bayesian nonparametric inference. We show that when the concentration parameter is large, the two-parameter Poisson-Dirichlet process and its corresponding quantile process share many asymptotic properties with the empirical process and the quantile process, respectively. More precisely, we show that an analogue of the empirical central limit theorem continues to hold for the two-parameter Poisson-Dirichlet process.

7.1 Asymptotic Properties of the Two-parameter Poisson-Dirichlet Process

Sethuraman and Tiwari (1982) studied the convergence and tightness of the Dirichlet process when the parameters converge in a certain sense. They showed that when the concentration parameter a converges to 0, the Dirichlet process $P_{H,0,a}$ converges to a degenerate probability measure at a point randomly chosen from H. In a remarkable paper, Lo (1987) combined the Dirichlet process with the Bayesian bootstrap, and

gave an approximation of the empirical process when the sample size is large.

The main objective of the present section is to study the weak convergence of the centered and scaled two-parameter Poisson-Dirichlet process defined by

$$D_{H,\theta,a}(\cdot) = \sqrt{a(1-\theta)^{-1}}\left(P_{H,\theta,a}(\cdot) - H(\cdot)\right), \tag{7.1.1}$$

as $a \to \infty$. We point out that our approach is significantly different form that considered in Lo (1987).

Let $\mathscr{S}$ be a collection of Borel sets in $\mathbb{R}$ and H be a probability measure on $\mathbb{R}$. We recall the definition of a Brownian bridge indexed by $\mathscr{S}$. A Gaussian process $\{B_H(S) : S \in \mathscr{S}\}$ is called a *Brownian bridge with parameter measure* H if $E\left[B_H(S)\right] = 0$ for any $S \in \mathscr{S}$ and

$$Cov\left(B_H(S_i), B_H(S_j)\right) = H(S_i \cap S_j) - H(S_i)H(S_j) \tag{7.1.2}$$

for any $S_i, S_j \in \mathscr{S}$ (Kim and Bickel, 2003). To see that (7.1.2) defines indeed a covariance function, note that

$$\begin{aligned}
\sum_{i,j=1}^{n} a_i a_j \left(H(S_i \cap S_j) - H(S_i)H(S_j)\right) &= \sum_{i,j=1}^{n} a_i a_j \int_{\mathbb{R}} I_{S_i} I_{S_j} dH - \left(\sum_{i=1}^{n} a_i H(S_i)\right)^2 \\
&= \int_{\mathbb{R}} \left(\sum_{i=1}^{n} a_i I_{S_i}\right)^2 dH - \left(\int_{\mathbb{R}} \sum_{i=1}^{n} a_i I_{S_i} dH\right)^2 \\
&\geq 0
\end{aligned}$$

using the fact that $\int_{\mathbb{R}} f dH \leq \left(\int_{\mathbb{R}} f^2 dH\right)^{1/2}$ for $f = \sum_{i=1}^{n} a_i I_{S_i}$.

The next lemma gives the limiting distribution of the process (7.1.1) for any finite Borel sets $S_1, \ldots, S_k \in \mathscr{S}$, as $a \to \infty$.

Lemma 7.1.1 *Let $D_{H,\theta,a}$ be defined by (7.1.1). Assume that $\theta = 0$ or $\theta = 1/2$. For any fixed sets $S_1, \ldots, S_k$ in $\mathscr{S}$ we have*

$$\left(D_{H,\theta,a}(S_1), D_{H,\theta,a}(S_2), \ldots, D_{H,\theta,a}(S_k)\right) \xrightarrow{d} \left(B_H(S_1), B_H(S_2), \ldots, B_H(S_k)\right),$$

as $a \to \infty$, where B_H is the Brownian bridge with parameter measure H.

Proof: We prove the result for $k = 2$. The general case is similar. Let S_1 and S_2 be two sets in $\mathscr{S}$. Assume first $S_1 \cap S_2 = \emptyset$. We treat separately the cases $\theta = 0$ and $\theta = 1/2$.

Case 1: $\theta = 0$ Note that

$$\begin{aligned}(P_{H,0,a}(S_1), P_{H,0,a}(S_2), 1 - P_{H,0,a}(S_1) - P_{H,0,a}(S_2)) &\sim D(aH(S_1), aH(S_2), \\ &\quad a(1 - H(S_1) - H(S_2))).\end{aligned}$$

Set $X^i_{H,0,a} = P_{H,0,a}(S_i)$ and $l_i = H(S_i),$ for $i = 1, 2$. Thus, the joint density function of $X^1_{H,0,a}$ and $X^2_{H,0,a}$ is:

$$f_{X^1_{H,0,a},X^2_{H,0,a}}(x_1, x_2) = \frac{\Gamma(a)}{\Gamma(al_1)\Gamma(al_2)\Gamma(a(1-l_1-l_2))} x_1^{al_1-1} x_2^{al_2-1}(1-x_1-x_2)^{a(1-l_1-l_2)-1}.$$

The joint probability density function of $D^1_{H,0,a} = \sqrt{a}\left(X^1_{H,0,a} - l_1\right) = \sqrt{a}(P_{H,0,a}(S_1) - H(S_1))$ and $D^2_{H,0,a} = \sqrt{a}\left(X^2_{H,0,a} - l_2\right) = \sqrt{a}\left(P_{H,0,a}(S_2) - H(S_2)\right)$ is:

$$\begin{aligned}f_{D^1_{H,0,a},D^2_{H,0,a}}(y_1, y_2) &= \frac{\Gamma(a)}{a\Gamma(al_1)\Gamma(al_2)\Gamma(a(1-l_1-l_2))} \left(\frac{1}{\sqrt{a}}y_1 + l_1\right)^{al_1-1} \\ &\quad \left(\frac{1}{\sqrt{a}}y_2 + l_2\right)^{al_2-1} \left(1 - \frac{y_1+y_2}{\sqrt{a}} - l_1 - l_2\right)^{a(1-l_1-l_2)-1}.\end{aligned}$$

By Scheffé's theorem (Billingsley 1999, page 29), it is enough to show that:

$$f_{D^1_{H,0,a},D^2_{H,0,a}}(y_1, y_2) \to f(y_1, y_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-(y_1\ y_2)\Sigma^{-1}(y_1\ y_2)^T/2\right\},$$

where $\Sigma = \begin{bmatrix} l_1(1-l_1) & -l_1l_2 \\ -l_1l_2 & l_2(1-l_2) \end{bmatrix}$.

Use Stirling's formula (Wilks 1963, page 177)

$$\Gamma(z) \approx \sqrt{2\pi} z^{z-\frac{1}{2}} e^{-z}, \text{ as } z \to \infty,$$

where we use the notation $f(z) \approx g(z)$ as $z \to \infty$ if $\lim_{z\to\infty} f(z)/g(z) = 1$. We get:

$$\begin{aligned}
\lim_{a\to\infty} f_{D^1_{H,0,a},D^2_{H,0,a}}(y_1,y_2) &= \frac{1}{2\pi}\lim_{a\to\infty}\left[\frac{\left(\frac{1}{\sqrt{a}}y_1+l_1\right)^{al_1-1}\left(\frac{1}{\sqrt{a}}y_2+l_2\right)^{al_2-1}}{l_1^{al_1-\frac{1}{2}}l_2^{al_2-\frac{1}{2}}}\right.\\
&\quad\left.\frac{\left(1-\frac{1}{\sqrt{a}}y_1-\frac{1}{\sqrt{a}}y_2-l_1-l_2\right)^{a(1-l_1-l_2)-1}}{(1-l_1-l_2)^{(1-l_1-l_2)a-\frac{1}{2}}}\right]\\
&= \frac{1}{2\pi\sqrt{l_1l_2(1-l_1-l_2)}}\lim_{a\to\infty}\left[\frac{\left(\frac{1}{\sqrt{a}}y_1+l_1\right)^{al_1-1}}{l_1^{al_1-1}}\right.\\
&\quad\left.\frac{\left(\frac{1}{\sqrt{a}}y_2+l_2\right)^{al_2-1}\left(1-\frac{1}{\sqrt{a}}y_1-\frac{1}{\sqrt{a}}y_2-l_1-l_2\right)^{a(1-l_1-l_2)-1}}{l_2^{al_2-1}(1-l_1-l_2)^{a(1-l_1-l_2)-1}}\right]\\
&= \frac{1}{2\pi\sqrt{l_1l_2(1-l_1-l_2)}}\lim_{a\to\infty}\left[\left(1+\frac{y_1}{\sqrt{a}l_1}\right)^{al_1}\right.\\
&\quad\left.\left(1+\frac{y_2}{\sqrt{a}l_2}\right)^{al_2}\left(1-\frac{y_1+y_2}{\sqrt{a}(1-l_1-l_2)}\right)^{a(1-l_1-l_2)}\right]\\
&= \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}}\exp\left\{\lim_{a\to\infty}a\ln v_a\right\}, \qquad (7.1.3)
\end{aligned}$$

where

$$\sigma_{11}=l_1(1-l_1),\quad \sigma_{22}=l_2(1-l_2),\quad \rho_{12}=-\sqrt{\frac{l_1l_2}{(1-l_1)(1-l_2)}}, \qquad (7.1.4)$$

and

$$v_a=\left(1+\frac{y_1}{\sqrt{a}l_1}\right)^{l_1}\left(1+\frac{y_2}{\sqrt{a}l_2}\right)^{l_2}\left(1-\frac{y_1+y_2}{\sqrt{a}(1-l_1-l_2)}\right)^{1-l_1-l_2}.$$

Observe that,

$$\begin{aligned}
\lim_{a\to\infty}a\ln v_a &= \lim_{a\to\infty}\frac{1}{1/a}\left[l_1\ln\left(1+\frac{y_1}{\sqrt{a}l_1}\right)+l_2\ln\left(1+\frac{y_2}{\sqrt{a}l_2}\right)\right.\\
&\quad\left.+(1-l_1-l_2)\ln\left(1-\frac{y_1+y_2}{\sqrt{a}(1-l_1-l_2)}\right)\right].
\end{aligned}$$

Using L'Hospital's rule we obtain $\lim_{a\to\infty}a\ln v_a$ equals:

$$\lim_{a\to\infty}\left[\frac{l_1\frac{-y_1}{2l_1a^{3/2}}}{\left(1+\frac{y_1}{l_1\sqrt{a}}\right)}+\frac{l_2\frac{-y_2}{2l_2a^{3/2}}}{\left(1+\frac{y_2}{l_2\sqrt{a}}\right)}-\frac{(1-l_1-l_2)\frac{-(y_1+y_2)}{2(1-l_1-l_2)a^{3/2}}}{\left(1-\frac{y_1+y_2}{(1-l_1-l_2)\sqrt{a}}\right)}\right](-a^2)$$

$$
\begin{aligned}
&= \lim_{a\to\infty} \frac{a}{2}\left[\frac{l_1 y_1}{l_1\sqrt{a}+y_1} + \frac{l_2 y_2}{l_2\sqrt{a}+y_2} - \frac{(1-l_1-l_2)(y_1+y_2)}{(1-l_1-l_2)\sqrt{a}-(y_1+y_2)}\right] \\
&= \lim_{a\to\infty} \frac{a}{2}\Bigg[\frac{l_1 y_1 y_2 + (1-l_2)y_1^2}{(l_1\sqrt{a}+y_1)\left((1-l_1-l_2)\sqrt{a}-(y_1+y_2)\right)} \\
&\quad + \frac{l_2 y_1 y_2 + (1-l_1)y_2^2}{(l_2\sqrt{a}+y_2)\left((1-l_1-l_2)\sqrt{a}-(y_1+y_2)\right)}\Bigg] \\
&= \lim_{a\to\infty} \frac{a}{2}\Bigg[\frac{l_1 y_1 y_2 + (1-l_2)y_1^2}{(l_1\sqrt{a}+y_1)\left((1-l_1-l_2)\sqrt{a}-(y_1+y_2)\right)} \\
&\quad + \frac{l_2 y_1 y_2 + (1-l_1)y_2^2}{(l_2\sqrt{a}+y_2)\left((1-l_1-l_2)\sqrt{a}-(y_1+y_2)\right)}\Bigg] \\
&= -\frac{l_2(1-l_2)y_1^2 + 2l_1 l_2 y_1 y_2 + l_1(1-l_1)y_2^2}{2l_1 l_2(1-l_1-l_2)} \\
&= -\frac{(1-l_1)(1-l_2)}{2(1-l_1-l_2)}\Bigg[\left(\frac{y_1}{\sqrt{l_1(1-l_1)}}\right)^2 + \left(\frac{y_2}{\sqrt{l_2(1-l_2)}}\right)^2 \\
&\quad + \frac{2y_1 y_2}{(1-l_1)(1-l_2)}\Bigg] \\
&= -\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{y_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{y_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{y_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{y_1}{\sqrt{\sigma_{11}}}\right)\right], \quad (7.1.5)
\end{aligned}
$$

where σ_{11}, σ_{22} and ρ_{12} are defined in (7.1.4). The proof in this case follows by using (7.1.3).

Case 2: $\theta = 1/2$ Note that

$$
\begin{aligned}
&\big(P_{H,1/2,a}(S_1), P_{H,1/2,a}(S_2), 1 - P_{H,1/2,a}(S_1) - P_{H,1/2,a}(S_2)\big) \\
&\qquad\qquad \sim RS\big(H(S_1), H(S_2), 1 - H(S_1) - H(S_2); 1/2, a\big),
\end{aligned}
$$

where the RS distribution is given by Definition 2.2.2. Set $X^i_{H,1/2,a} = P_{H,1/2,a}(S_i)$ and $l_i = H(S_i),\ i = 1, 2$. Thus, the joint density function of $X^1_{H,1/2,a}$ and $X^2_{H,1/2,a}$ is:

$$
f_{X^1_{H,1/2,a}, X^2_{H,1/2,a}}(x_1, x_2) = \frac{l_1 l_2 (1-l_1-l_2)\Gamma(a+\frac{3}{2})}{\pi^{\frac{3-1}{2}}\Gamma(a+\frac{1}{2})} \times \frac{x_1^{-3/2} x_2^{-3/2}(1-x_1-x_2)^{-3/2}}{\left(\frac{l_1^2}{x_1} + \frac{l_2^2}{x_2} + \frac{(1-l_1-l_2)^2}{1-x_1-x_2}\right)^{a+\frac{3}{2}}}
$$

$$= \frac{l_1 l_2 (1-l_1-l_2)\left(a+\frac{1}{2}\right)}{\pi} \times \frac{x_1^{-3/2} x_2^{-3/2} (1-x_1-x_2)^{-3/2}}{\left(\frac{l_1^2}{x_1}+\frac{l_2^2}{x_2}+\frac{(1-l_1-l_2)^2}{1-x_1-x_2}\right)^{a+\frac{3}{2}}}.$$

The joint probability density function of $D^1_{H,1/2,a} = \sqrt{2a}\left(X^1_{H,1/2,a} - l_1\right)$ and $D^2_{H,1/2,a} = \sqrt{2a}\left(X^2_{H,1/2,a} - l_2\right)$ is:

$$f_{D^1_{H,1/2,a}, D^2_{H,1/2,a}}(y_1, y_2) = \frac{a+\frac{1}{2}}{a} \times \frac{l_1 l_2 (1-l_1-l_2)}{2\pi} \times \left(\frac{1}{\sqrt{2a}} y_1 + l_1\right)^{-3/2} \times \frac{\left(\frac{1}{\sqrt{2a}} y_2 + l_2\right)^{-3/2} \left(1 - \frac{1}{\sqrt{2a}} y_1 - \frac{1}{\sqrt{2a}} y_2 - l_1 - l_2\right)^{-3/2}}{\left(\frac{l_1^2}{\frac{1}{\sqrt{2a}} y_1 + l_1} + \frac{l_2^2}{\frac{1}{\sqrt{2a}} y_2 + l_2} + \frac{(1-l_1-l_2)^2}{1 - \frac{1}{\sqrt{2a}} y_1 - \frac{1}{\sqrt{2a}} y_2 - l_1 - l_2}\right)^{a+\frac{3}{2}}}.$$

Notice that

$$\frac{a+\frac{1}{2}}{a} \times \frac{l_1 l_2 (1-l_1-l_2)}{2\pi} \times \frac{\left(\frac{1}{\sqrt{2a}} y_1 + l_1\right)^{-3/2} \left(\frac{1}{\sqrt{2a}} y_2 + l_2\right)^{-3/2}}{\left(\frac{l_1^2}{\frac{1}{\sqrt{2a}} y_1 + l_1} + \frac{l_2^2}{\frac{1}{\sqrt{2a}} y_2 + l_2} + \frac{(1-l_1-l_2)^2}{1 - \frac{1}{\sqrt{2a}} y_1 - \frac{1}{\sqrt{2a}} y_2 - l_1 - l_2}\right)^{\frac{3}{2}}}$$

converges to $1/\left(2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}\right)$, where σ_{11}, σ_{22} and ρ_{12} are defined in (7.1.4). To prove the lemma, it remains to show that

$$\left(\frac{l_1^2}{\frac{1}{\sqrt{2a}} y_1 + l_1} + \frac{l_2^2}{\frac{1}{\sqrt{2a}} y_2 + l_2} + \frac{(1-l_1-l_2)^2}{1 - \frac{1}{\sqrt{2a}} y_1 - \frac{1}{\sqrt{2a}} y_2 - l_1 - l_2}\right)^{-a}$$

converges to (7.1.9). The argument is similar to case 1 and is omitted.

Now we tackle the case when the sets S_1 and S_2 are not necessarily disjoint. Observe that the sets $S_1 \cap S_2^c$, $S_1^c \cap S_2$ and $S_1 \cap S_2$ are disjoint. Therefore, by the case which was already treated,

$$\begin{aligned} \mathbf{X}_\alpha &= \left(D_{H,\theta,a}(S_1 \cap S_2^c), D_{H,\theta,a}(S_1^c \cap S_2), D_{H,\theta,a}(S_1 \cap S_2)\right) \\ &\xrightarrow{d} \left(B_H(S_1 \cap S_2^c), B_H(S_1^c \cap S_2), B_H(S_1 \cap S_2)\right). \end{aligned}$$

Notice that $(B_H(S_1 \cap S_2^c), B_H(S_1^c \cap S_2), B_H(S_1 \cap S_2))$ has the 3−variate normal distribution with mean vector $\mathbf{0} = (0,0,0)^T$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix},$$

where $\sigma_{11} = H(S_1 \cap S_2^c) - H^2(S_1 \cap S_2^c)$, $\sigma_{12} = -H(S_1 \cap S_2^c)H(S_1^c \cap S_2)$, $\sigma_{13} = -H(S_1 \cap S_2^c)H(S_1 \cap S_2)$, $\sigma_{22} = H(S_1^c \cap S_2) - H^2(S_1^c \cap S_2)$, $\sigma_{23} = -H(S_1^c \cap S_2)H(S_1 \cap S_2)$ and $\sigma_{33} = H(S_1 \cap S_2) - H^2(S_1 \cap S_2)$. By application A of Serfling (2002, page 26) we have

$$B\mathbf{X}_\alpha^T \xrightarrow{d} N_3\left(B\mathbf{0}, B\Sigma B^T\right),$$

where $B = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$. Since $D_{H,\theta,a}(\cdot)$ is finitely additive, we get

$$B\mathbf{X}_\alpha^T = \begin{bmatrix} D_{H,\theta,a}(S_1 \cap S_2^c) + D_{H,\theta,a}(S_1 \cap S_2) \\ D_{H,\theta,a}(S_1^c \cap S_2) + D_{H,\theta,a}(S_1 \cap S_2) \end{bmatrix} = \begin{bmatrix} D_{H,\theta,a}(S_1) \\ D_{H,\theta,a}(S_2) \end{bmatrix}.$$

and

$$B\Sigma B^T = \begin{bmatrix} H(S_1) - H^2(S_1) & H(S_1 \cap S_2) - H(S_1)H(S_2) \\ H(S_1 \cap S_2) - H(S_1)H(S_2) & H(S_2) - H^2(S_2) \end{bmatrix}.$$

That is,

$$(D_{H,\theta,a}(S_1), D_{H,\theta,a}(S_2)) \xrightarrow{d} (B_H(S_1), B_H(S_2)).$$

∎

The next theorem shows that the process $D_{H,\theta,a}$ defined by (7.1.1) converges to the process B_H on $D[-\infty,\infty]$ with respect to the Skorokhod topology, where $D[-\infty,\infty]$ is the space of cadlag functions (right continuous with left limits) on $[-\infty,\infty]$. Right continuity at $-\infty$ can be achieved by setting $D_{H,\theta,a}(-\infty) = 0$; the

left limit at ∞ also equals zero, the natural value of $D_{H,\theta,a}(\infty)$. For more details, consult Pollard (1984, Chapter 5). If X and $(X_a)_{a>0}$ are random variables with values in a metric space M, we say that $(X_a)_a$ converges to X as $a \to \infty$ (and we write $X_a \xrightarrow{d} X$) if for any sequence $(a_n)_n$ converging to ∞, X_{a_n} converges in distribution to X.

Theorem 7.1.2 *For $\theta \in \{0, 1/2\}$, as $a \to \infty$, we have:*

$$D_{H,\theta,a}(\cdot) = \sqrt{a(1-\theta)^{-1}}\,(P_{H,\theta,a}(\cdot) - H(\cdot)) \xrightarrow{d} B_H(\cdot) \tag{7.1.6}$$

on $D[-\infty, \infty]$ with respect to the Skorokhod topology, where B_H is the Brownian bridge with parameter measure H.

Proof: Let (a_n) be an arbitrary sequence such that $a_n \to \infty$. To simplify the notation, in the argument below, we omit writing the index n of a_n. Assume first that $H(t) = \lambda(t) = t$ (i.e. λ is the Lebesgue measure on $[0, 1]$). Thus the process (7.1.6) reduces to

$$D_{\lambda,\theta,a}(t) = \sqrt{a(1-\theta)^{-1}}\,(P_{\lambda,\theta,a}(t) - t)\,.$$

To prove the theorem, we use Lemma 7.1.1 and Theorem 13.5 of Billingsley (1999). Therefore, we only need to show that for any $0 \le t_1 \le t \le t_2 \le 1$,

$$E\left[|D_{\lambda,\theta,a}(t) - D_{\lambda,\theta,a}(t_1)|^{2\beta}\,|D_{\lambda,\theta,a}(t_2) - D_{\lambda,\theta,a}(t)|^{2\beta}\right] \le |F(t_2) - F(t_1)|^{2\alpha}\,, \tag{7.1.7}$$

for some $\beta \ge 0$, $\alpha > 1/2$, and a nondecreasing continuous function F on $[0, 1]$. Take $\alpha = \beta = 1$ and $F(t) = t$. We show that

$$E\left((D_{\lambda,\theta,a}(t) - D_{\lambda,\theta,a}(t_1))^2\,(D_{\lambda,\theta,a}(t_2) - D_{\lambda,\theta,a}(t))^2\right) \le 8a^2(1-\theta)^{-2}\,(t_2 - t_1)^2\,. \tag{7.1.8}$$

Observe that

$$D_{\lambda,\theta,a}(t) - D_{\lambda,\theta,a}(t_1) = D_{\lambda,\theta,a}((t_1, t]) \text{ and } D_{\lambda,\theta,a}(t_2) - D_{\lambda,\theta,a}(t) = D_{\lambda,\theta,a}((t, t_2]).$$

Thus, the expectation in the left-hand side of (7.1.8) is equal to

$$a^2(1-\theta)^{-2}E\left(\left\{P_{\lambda,\theta,a}((t_1,t])-\lambda((t_1,t])\right\}^2\left\{P_{\lambda,\theta,a}((t,t_2])-\lambda((t,t_2])\right\}^2\right), \quad (7.1.9)$$

where $\lambda((t,t_2])=t_2-t$ and $\lambda((t_1,t])=t-t_1$. Expanding the expression

$$\left\{P_{\lambda,\theta,a}((t_1,t])-\lambda((t_1,t])\right\}^2\left\{P_{\lambda,\theta,a}((t,t_2])-\lambda((t,t_2])\right\}^2$$

gives

$$\begin{aligned}
&P^2_{\lambda,\theta,a}((t_1,t])P^2_{\lambda,\theta,a}((t,t_2])-2\lambda((t,t_2])P^2_{\lambda,\theta,a}((t_1,t])P_{\lambda,\theta,a}((t,t_2])\\
&\quad+\lambda^2((t,t_2])P^2_{\lambda,\theta,a}((t_1,t])-2\lambda((t_1,t])P_{\lambda,\theta,a}((t_1,t])P^2_{\lambda,\theta,a}((t,t_2])\\
&\quad+4\lambda((t_1,t])\lambda((t,t_2])P_{\lambda,\theta,a}((t_1,t])P_{\lambda,\theta,a}((t,t_2])-2\lambda((t_1,t])\lambda^2((t,t_2])P_{\lambda,\theta,a}((t_1,t])\\
&\quad+\lambda^2((t_1,t])P^2_{\lambda,\theta,a}((t,t_2])-2\lambda^2((t_1,t])\lambda((t,t_2])P_{\lambda,\theta,a}((t,t_2])+\lambda^2((t_1,t])\lambda^2((t,t_2])\\
&\quad\le P^2_{\lambda,\theta,a}((t_1,t])P^2_{\lambda,\theta,a}((t,t_2])+\lambda^2((t,t_2])P^2_{\lambda,\theta,a}((t_1,t])\\
&\quad+4\lambda((t_1,t])\lambda((t,t_2])P_{\lambda,\theta,a}((t_1,t])P_{\lambda,\theta,a}((t,t_2])+\lambda^2((t_1,t])P^2_{\lambda,\theta,a}((t,t_2])\\
&\quad+\lambda^2((t_1,t])\lambda^2((t,t_2])\\
&\quad\le P_{\lambda,\theta,a}((t_1,t])P_{\lambda,\theta,a}((t,t_2])+\lambda((t,t_2])P_{\lambda,\theta,a}((t_1,t])\\
&\quad+4\lambda((t_1,t])\lambda((t,t_2])+\lambda((t_1,t])P_{\lambda,\theta,a}((t,t_2])+\lambda((t,t_2])\lambda((t_1,t]),
\end{aligned}$$

using the fact that $P_{\lambda,\theta,a}(\cdot)$ is a probability measure and $0\le t_1\le t\le t_2\le 1$. By (2.2.2) and (2.2.3) we obtain

$$E\left(\left\{P_{\lambda,\theta,a}((t_1,t])-\lambda((t_1,t])\right\}^2\left\{P_{\lambda,\theta,a}((t,t_2])-\lambda((t,t_2])\right\}^2\right)\le 8\lambda((t_1,t])\lambda((t,t_2]). \quad (7.1.10)$$

Thus, using (7.1.9) and (7.1.10), we have

$$\begin{aligned}
E\left(\left(D_{\lambda,\theta,a}(t)-D_{\lambda,\theta,a}(t_1)\right)^2\left(D_{\lambda,\theta,a}(t_2)-D_{\lambda,\theta,a}(t)\right)^2\right) &= 8a^2(1-\theta)^{-2}\lambda((t_1,t])\lambda((t,t_2])\\
&= 8a^2(1-\theta)^{-2}(t-t_1)(t_2-t)\\
&\le 8a^2(1-\theta)^{-2}(t_2-t_1)^2,
\end{aligned}$$

for $0 \leq t_1 \leq t \leq t_2 \leq 1$. This proves the theorem in the case when $H(t) = t$, i.e. H is the uniform distribution. Observe that, the quantile function $H^{-1}(s) = \inf\{t : H(t) \geq s\}$ has the property: $H^{-1}(s) \leq t$ if and only if $s \leq H(t)$. If U_i is uniformly distributed over $[0, 1]$, then $H^{-1}(U_i)$ has distribution H. Thus, we can use the representation $\theta_i = H^{-1}(U_i)$, where $(U_i)_{i\geq 1}$ is a sequence of i.i.d. random variables with uniform distribution on $[0, 1]$, to have:

$$P_{H,\theta,a}(t) = P_{\lambda,\theta,a}(H(t)) \text{ and } D_{H,\theta,a}(t) = D_{\lambda,\theta,a}(H(t)) = D_{\lambda,\theta,a} \circ H(t), \quad t \in \mathbb{R},$$

where $P_{\lambda,\theta,a} = \sum_{i=1}^{\infty} p_i \delta_{U_i}$, is a two-parameter Poisson-Dirichlet process with discount parameter θ, concentration parameter a and Lebesgue base measure λ on $[0, 1]$. From the uniform case, which was already treated, we have $D_{\lambda,\theta,a}(\cdot) = \sqrt{a(1-\theta)^{-1}}$ $(P_{\lambda,\theta,a}(\cdot) - \lambda(\cdot)) \xrightarrow{d} B_\lambda(\cdot)$. Define $\Psi : D[0, 1] \to D[-\infty, \infty]$ by $(\Psi x)(t) = x(H(t))$. Since the function Ψ is uniformly continuous (Billingsley 1999, page 150; Pollard, 1984, page 97), it follows, from the continuous mapping theorem and the fact that $D_{\lambda,\theta,a} \xrightarrow{d} B_\lambda$, that $D_{H,\theta,a} = \Psi(D_{\lambda,\theta,a}) \xrightarrow{d} \Psi(B_\lambda) = B_H$. This completes the proof of the theorem. ∎

Remark 7.1.3 *In the proof of Lemma 7.1.1, the finite-dimensional convergence is in fact convergence in total variation, which is stronger than the convergence in distribution (Billingsley 1999, page 29).*

Remark 7.1.4 *The tightness condition described by relation (7.1.7) (in the proof of Theorem 7.1.2) holds for any $\theta \in [0, 1)$ and any $a > -\theta$. We conjuncture that Theorem 7.1.2 holds for any $\theta \in [0, 1)$. For more discussion, see Section 7.3.*

7.2 Asymptotic Properties of the Two-Parameter Poisson-Dirichlet Quantile Process

Similar to the frequentist asymptotic theory, in this section we establish large sample theory for the two-parameter Poisson-Dirichlet quantile process. Specifically, we derive the limiting distribution of the two-parameter Poisson-Dirichlet quantile process

$$Q_{H,\theta,a}(\cdot) = \sqrt{a(1-\theta)^{-1}}\left(P^{-1}_{H,\theta,a}(\cdot) - H^{-1}(\cdot)\right), \tag{7.2.1}$$

where the inverse of a distribution function F is defined by

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

Corollary 7.2.1 *Let $0 < p < q < 1$, and H be a continuous function with positive derivative h on the interval $[H^{-1}(p) - \epsilon, H^{-1}(q) + \epsilon]$ for some $\epsilon > 0$. Let $Q_{H,\theta,a}$ be the two-parameter Poisson-Dirichlet quantile process defined in (7.2.1) and let λ be the Lebesgue measure on $[0,1]$. For $\theta \in \{0, 1/2\}$, as $a \to \infty$, we have*

$$Q_{H,\theta,a}(\cdot) \xrightarrow{d} -\frac{B_\lambda(\cdot)}{h(H^{-1}(\cdot))} = Q(\cdot),$$

on $D[p,q]$. That is, the limiting process is a Gaussian process with zero-mean and covariance function

$$Cov\left(Q(s), Q(t)\right) = \frac{\lambda(s \wedge t) - \lambda(s)\lambda(t)}{h(H^{-1}(s))h(H^{-1}(t))}, \quad s,t \in \mathbb{R}.$$

Proof: By Theorem 7.1.2 the process $\sqrt{a(1-\theta)^{-1}}\left(P_{H,\theta,a} - H\right)$ converges in distribution to the process $B_H = B_\lambda(H) = B_\lambda \circ H$. (Notice that, almost all sample paths of the limiting process are continuous on the interval $[H^{-1}(p) - \epsilon, H^{-1}(q) + \epsilon]$.) By Lemma 3.9.23 page 386 of van der Vaart and Wellner (1996), the inverse map $H \mapsto H^{-1}$ is Hadamard-differentiable at H tangentially to the subspace of functions that are continuous on this interval. By the functional delta method (van der Vaart

and Wellner, 1996, Theorem 3.9.4, page 374) we have

$$Q_{H,\theta,a}(\cdot) \xrightarrow{d} -\frac{B_\lambda \circ H \circ H^{-1}(\cdot)}{h(H^{-1}(\cdot))} = -\frac{B_\lambda(\cdot)}{h(H^{-1}(\cdot))}$$

on $D[p,q]$. This completes the proof of the corollary. ∎

Remark 7.2.2 *Similar to Remark 1 of Bickel and Freedman (1981), if* $H^{-1}(0+) > -\infty$ *and* $H^{-1}(1) < \infty$ *and* h *is continuous on* $[H^{-1}(0+),\, H^{-1}(1)]$, *the conclusion of the corollary holds on* $D\left[H^{-1}(0+), H^{-1}(1)\right]$. *For example, if* H *is a uniform distribution on* $[0,1]$, *then the convergence holds on* $D[0,1]$.

The next two examples are direct applications of Corollary 7.2.1.

Example 7.2.3 (Median) Let M_a be the median of $P_{H,0,a}$ and m be the median of H (i.e. $P^{-1}_{H,0,a}(0.5) = M_a$ and $H^{-1}(0.5) = m$). From Corollary 7.2.1 we have

$$\sqrt{a}\,(M_a - m) \xrightarrow{d} N\left(0, \frac{1}{4h^2(m)}\right),$$

where $h = H'$ is the probability density function. Consequently, the asymptotic distribution of the median for Dirichlet process coincides with that of the sample median (DasGupta, 2008, Theorem 7.2, page 93).

Example 7.2.4 (Interquartile Range) In this example, we find the asymptotic distribution of the interquantile range $IQR = Q_{3,a} - Q_{1,a}$, where $Q_{3,a}$ and $Q_{1,a}$ are the third and the first quartiles of $P_{H,0,a}$ (i.e. $P^{-1}_{H,0,a}(0.75) = Q_{3,a}$ and $P^{-1}_{H,0,a}(0.25) = Q_{1,a}$). Let q_3 and q_1 be the third and the first quartiles of H. From Corollary 7.2.1 we have

$$\left(\sqrt{a}\,(Q_{3,a} - q_3)\,, \sqrt{a}\,(Q_{1,a} - q_1)\right) \xrightarrow{d} N\left(\mathbf{0}, \Sigma\right),$$

where

$$\Sigma = \begin{bmatrix} \frac{3}{16h^2(q_3)} & \frac{1}{16h(q_1)h(q_3)} \\ \frac{1}{16h(q_1)h(q_3)} & \frac{3}{16h^2(q_1)} \end{bmatrix},$$

where $h = H'$ is the probability density function. Using the continuous mapping theorem, it follows by a simple calculation that

$$\sqrt{a}\,(IQR - (q_3 - q_1)) \xrightarrow{d} N\left(0, \frac{3}{h^2(q_3)} + \frac{3}{h^2(q_1)} - \frac{2}{h(q_1)h(q_3)}\right).$$

This gives the asymptotic distribution of the sample interquartile range (DasGupta, 2008, Example 7.3, page 93).

7.3 Conjecture: The General Case

Unfortunately, the finite-dimensional distributions of the two-parameter Poisson-Dirichlet process are not tractable for cases other than $\theta \in \{0, 1/2\}$. To extend Theorem 7.1.2 to $\theta \in [0, 1)$, one may study the convergence of the moments of the process $D_{H,\theta,a}$, as $a \to \infty$. For this purpose, we recall first the following theorem from van der Vaart (1998, Theorem 2.22, page 18).

Theorem 7.3.1 *Let X_n and X be random variables such that $E[X_n^r] \to E[X^r] < \infty$ for $r \in \mathbb{N}$. If the distribution of X is uniquely determined by its moments, then $X_n \xrightarrow{d} X$.*

Using the fact that the normal distribution is uniquely determined by its moments (van der Vaart, 1998, example 2.23, page 18), by Theorem 7.3.1, one may prove Theorem 7.1.2 for any $\theta \in [0, 1)$ by showing that

$$\left(a(1-\theta)^{-1}\right)^{r/2} E\left[(P_{H,\theta,a}(S) - H(S))^r\right] \to \begin{cases} 0 & r \text{ odd} \\ \frac{r!}{(r/2)!2^{r/2}} \left(H(S)(1-H(S))\right)^{r/2} & r \text{ even} \end{cases}. \tag{7.3.1}$$

Note that the limit in (7.3.1) represents the moments of the normal distribution with mean 0 and variance $H(S)\,(1 - H(S))$. A general formula for computing $E\left[P_{H,\theta,a}^r(S)\right]$ was derived by Carleton (1999, Theorem 4.4). More precisely,

$$E\left[P_{H,\theta,a}^r(S)\right] = \sum{}' \frac{N(\mathbf{b})}{(a+1)_{n-1}} \times \prod_{l=1}^{k(\mathbf{b})-1} (a + l\theta) \times \prod_{j=2}^{r} \left[(1-\theta)_{j-1}\right]^{b_j} \times \left[H(S)\right]^{k(\mathbf{b})}, \tag{7.3.2}$$

where $\sum'$ runs over all $\mathbf{b} = (b_1, \ldots, b_r) \in \mathbb{N}^r$ which satisfy $\sum_{j=1}^r jb_j = r$, $k(\mathbf{b}) = \sum_{j=1}^r b_j$, $N(\mathbf{b}) = r!/\prod_{j=1}^r \left(j!^{b_j} b_j!\right)$ and $(x)_r = \Gamma(x+r)/\Gamma(x)$.

For example, using formula (7.3.2), the third and the fourth moments of $P_{H,\theta,a}(S)$ are:

$$E\left[P^3_{H,\theta,a}(S)\right] = H(S)\frac{(2-\theta)(1-\theta)}{(a+2)(a+1)} + 3H^2(S)\frac{(a+\theta)(1-\theta)}{(a+2)(a+1)} + H^3(S)\frac{(a+\theta)(a+2\theta)}{(a+2)(a+1)}$$

and

$$\begin{aligned} E\left[P^4_{H,\theta,a}(S)\right] &= H(S)\frac{(3-\theta)(2-\theta)(1-\theta)}{(a+3)(a+2)(a+1)} + 4H^2(S)\frac{(a+\theta)(2-\theta)(1-\theta)}{(a+3)(a+2)(a+1)} \\ &\quad + 6H^3(S)\frac{(a+\theta)(a+2\theta)(1-\theta)}{(a+3)(a+2)(a+1)} + 3H^2(S)\frac{(a+\theta)(1-\theta)^2}{(a+3)(a+2)(a+1)} \\ &\quad + H^4(S)\frac{(a+\theta)(a+2\theta)(a+3\theta)}{(a+3)(a+2)(a+1)}. \end{aligned}$$

Hence, after some algebra, the third and the fourth central moments of $P_{H,\theta,a}(S)$ simplify to:

$$E\left[\left(P_{H,\theta,a}(S) - H(S)\right)^3\right] = H(S)(1-H(S))(1-2H(S))\frac{(1-\theta)(2-\theta)}{(a+1)(a+2)} \qquad (7.3.3)$$

and

$$\begin{aligned} E\left[\left(P_{H,\theta,a}(S) - H(S)\right)^4\right] &= \frac{1}{(a+1)(a+2)(a+3)}\Big[3H^2(S)(1-H(S))^2(1-\theta)^2 a \\ &\quad + 6H(S)^2(1-H(S))^2(\theta-1)(\theta^2-3\theta+3) \\ &\quad + H(S)(H(S)-1)(\theta-1)(\theta^2-5\theta+6)\Big]. \qquad (7.3.4) \end{aligned}$$

It follows from (2.2.2), (7.3.3) and (7.3.4) that, for $r = 1$ and $r = 3$, the left-hand side of (7.3.1) converges to 0 as $a \to \infty$. On the other hand, the same quantity converges to $H(S)(1-H(S))$ or $3H^2(S)(1-H(S))^2$ for $r = 2$ or $r = 4$, respectively. This gives a strong indication that Theorem 7.1.2 may be generalized to any $\theta \in [0, 1)$, which corresponds to our conjecture. Unfortunately, the moments of order 5 and above are too complex and we do not pursue the question of their convergence here.

Notice that, from (7.3.3), the right-hand side of (7.3.1) does not converge to zero as $\theta \to 1, \theta < 1$. Therefore, the random variable $\sqrt{a(1-\theta)^{-1}}\big(P_{H,\theta,a}(S) - H(S)\big)$ does not converge to a normal distribution, as $\theta \to 1, \theta < 1$.

Remark 7.3.2 *One can use the results obtained in this section to derive some asymptotic properties of any Hadamard-differentiable functional of the $PDP(H;\theta,a)$ as $a \to \infty$. For different applications in statistics we refer the reader to van der Vaart and Wellner (1996, Section 3.9) and Lo (1987).*

Remark 7.3.3 *The approach used in this chapter can be easily applied whenever H is a multivariate cumulative distribution function, using results of Bickel and Wichura (1972).*

7.4 Glivenko-Cantelli Theorem for the Two-Parameter Poisson-Dirichlet Process

In this section, we establish an interesting analogue of the Glivenko-Cantelli theorem for the two-parameter Poisson-Dirichlet process. First we show that the strong law of large numbers continues to hold for the two-parameter Poisson-Dirichlet process.

Theorem 7.4.1 *Let $P_{H,\theta,a} \sim PDP(H;\theta,a)$. Assume that $a = nc$, for a fixed positive number c. Then as $n \to \infty$,*

$$P_{H,\theta,nc}(A) \stackrel{a.s.}{\to} H(A),$$

for any measurable subset A of $\mathfrak{X}$.

Proof: The proof is based on the first Borel-Cantelli Lemma. For any $\epsilon > 0$, we have

$$\Pr\left\{|P_{H,\theta,nc}(A) - H(A)| > \epsilon\right\} \;=\; \Pr\left\{|P_{H,\theta,nc}(A) - H(A)|^4 > \epsilon^4\right\}$$

$$\leq \frac{E\left[|P_{H,\theta,nc}(A) - H(A)|^4\right]}{\epsilon^4}.$$

By (7.3.4),

$$\begin{aligned} E\left[(P_{H,\theta,nc}(A) - H(A))^4\right] &= \frac{1}{(nc+1)(nc+2)(nc+3)}\Big[3H^2(A)(1-H(A))^2(1-\theta)^2 \\ & cn + 6H(A)^2(1-H(A))^2(\theta-1)(\theta^2-3\theta+3) \\ & +H(A)(H(A)-1)(\theta-1)(\theta^2-5\theta+6)\Big]. \end{aligned}$$

Thus,

$$\sum_{n=1}^{\infty} \Pr\left\{|P_{H,\theta,nc}(A) - H(A)| > \epsilon\right\} < \infty.$$

Therefore, by the first Borel-Cantelli Lemma, the proof follows. ∎

Theorem 7.4.2 *Let $P_{H,\theta,a} \sim PDP(H;\theta,a)$. Assume that $a = nc$, for a fixed positive number c. If the cumulative distribution function H is continuous, then*

$$\sup_{x\in\mathbb{R}} |P_{H,\theta,nc}(x) - H(x)| \overset{a.s.}{\to} 0, \tag{7.4.1}$$

as $n \to \infty$.

Proof: The proof is similar to the empirical process counterpart. See, for example, Jiang (2009, Theorem 7.2, page 218). We include the proof for the sake of completeness. To show (7.4.1) we need to show that for any $\epsilon > 0$, there is $N \geq 1$ such that the left hand side of (7.4.1) is less than 2ϵ if $n \geq N$. Since $H(x) \to 1$ as $x \to \infty$ and $H(x) \to 0$ as $x \to -\infty$, we can find $A < 0$ and $B > 0$ such that $H(A) \leq \epsilon$ and $H(B) \geq 1-\epsilon$. Because $H(x)$ is continuous over $[A, B]$, it is uniformly continuous on $[A, B]$ (Bartle and Sherbert, 2000, Theorem 5.4.3, page 138). Therefore, there are points $A = x_0 < x_1 < \cdots < x_k < x_{k+1} = B$ such that $H(x_{j+1}) - H(x_j) < \epsilon$, where $0 \leq j \leq k+1$. By Theorem 7.4.1, $P_{H,\theta,nc}(x_j) \overset{a.s.}{\to} H(x_j)$. Assume this converges holds

on the set Ω_j with $\Pr(\Omega_j) = 1$. Then for any $\omega \in \Omega_j$ there exists a number $N_j(\omega)$ such that

$$|P_{H,\theta,nc}(\omega, x_j) - H(x_j)| < \varepsilon$$

for any $n \geq N_j(\omega)$. Let $\widetilde{\Omega} = \cap_{j=0}^{k+1} \Omega_j$. Note that $\Pr(\widetilde{\Omega}) = 1$. For any $\omega \in \widetilde{\Omega}$, let $N(\omega) = \max_j N_j(\widetilde{\Omega})$. Let $\omega \in \widetilde{\Omega}$ be fixed. Using the same argument as in the proof of Polya's theorem (see Example 1.6, Jiang, 2009, page 6), it follows that

$$\sup_{x \in \mathbb{R}} |P_{H,\theta,nc}(\omega, x_j) - H(x_j)| < 2\varepsilon$$

for any $n \geq N(\omega)$. The conclusion follows. ∎

Chapter 8

Conclusions and Future Work

8.1 Conclusions

The Dirichlet process, the two-parameter Poisson-Dirichlet process, the beta process and the beta-Stacy process are important Bayesian nonparametric priors which have been used frequently in Bayesian nonparametric inference. Sampling from these processes plays a crucial role in many applications in Bayesian nonparametric inference. Since these processes consist of infinite number of weights at infinite number of points, having exact samples from these processes is impossible. The existing approximating sampling algorithms are either slow or very complex and may be difficult to apply for many users. Therefore, it is worthwhile to develop simple and efficient approximation techniques for simulating the above processes. This task was a main objective of the thesis and has been completed in Chapter 4. First, the mathematical techniques needed to construct the approximations have been discussed, then the proofs showing the convergence of the approximations to the actual processes have been provided. For instance, we have shown that our new approximation of the Dirichlet process converges almost surely to Ferguson's sum representation of the Dirichlet process. We have compared the efficiency of the new approximations to several other well-known

approximations and have demonstrated a significant improvement.

The second goal of this thesis was to calculate the Kolmogorov, Lévy and Cramér-von Mises distances between the Dirichlet process and its base measure. This goal has been achieved in Chapter 5. In Chapter 6 of the thesis, the derived expressions of each distance are used in two applications. In the first application, they are used to select proper values of the concentration parameter of the Dirichlet process. In the second application, the derived expressions for each distance have been employed to develop a Bayesian goodness of fit test for a simple and composite hypothesis for non-censored and right censored observations. For the non-censored observations, we have considered the Dirichlet process. On the other hand, the beta-Stacy process has been considered for the right censored observations. Illustrative examples and simulation results are included in this chapter.

Our last objective was describing the relationship between the frequentist and Bayesian nonparametric statistics. This objective has been accomplished in Chapter 7. We have shown that, when the concentration parameter is large, the centered and scaled two-parameter Poisson-Dirichlet process converges to a certain Brownian bridge. We have applied the functional delta method to this result to derive the limiting process for the two-parameter Poisson-Dirichlet quantile process. We also have derived the strong law of large numbers and the Glivenko-Cantelli theorem for the two-parameter Poisson-Dirichlet process.

8.2 Research Extensions

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

(1) Implementing the derived sampling methods of Chapter 4 of the thesis in some real applications.

(2) Employing the functional central limit theorem of the two-parameter Poisson-Dirichlet process in some applications in statistics, such as Bayesian bootstrap (Rubin, 1982; Lo, 1987) and testing statistical hypothesis (van der Vaart and Wellner, 1996, Section 3.9).

(3) Generalizing the functional central limit theorem of the two-parameter Poisson-Dirichlet process to the case when the base measure H is a multivariate cumulative distribution function. The result of Bickel and Wichura (1972) can be employed in the proof.

(4) Studying functional central limit theorems for processes with tractable finite dimensional distributions. An example of such processes is the normalized inverse-Gaussian process (Lijoi, Mena and Prünster, 2005).

(5) Studying the functional central limit theorem for mixtures of Dirichlet processes (Antoniak, 1974).

Appendix A

Lifetime Distributions and Product Integrals

The main objective of this section is to give a general introduction to survival analysis and product integrals. Most of the material covered in the section can be found in standard textbooks (e.g. Lawless, 2003).

Let T be a nonnegative random variable representing the lifetimes of an individual (or component) in a population. We assume first that T is continuous. Let $f(t)$ be the probability density function of T and F be the cumulative distribution function of T:

$$F(t) = \Pr\{\text{an individual fails before } T\} = \Pr\{T \leq t\} = \int_0^t f(x)dx.$$

The probability that an individual survives longer than t is given by the survival function

$$S(t) \;=\; \Pr\{T > t\} = 1 - F(t) = \int_t^\infty f(x)dx. \tag{A.0.1}$$

From (A.0.1), it follows that $S(t)$ is a monotone decreasing continuous function of t, with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$. In some cases, it is possible to

allow $S(0) < 1$ or $S(\infty) > 0$ to consider cases of an immediate failure (death), or cases where some components never fail.

An important concept for the study of lifetime distributions is the *hazard function* (or hazard rate) $h(t)$, defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T < t + \Delta t | T \geq t\}}{\Delta t} = \frac{f(t)}{S(t)}. \tag{A.0.2}$$

The hazard function specifies the instantaneous rate of death (or failure) at time t, given that the individual survives up to t; $h(t)\Delta t$ is the approximate probability of death in $[t, t + \Delta t)$, given survival up to t. Clearly, the hazard function is nonnegative.

The functions $f(t), F(t), S(t)$, and $h(t)$ are mathematically equivalent, i.e. if one of them is given, then the other three can be derived. It is easy to derive expressions for $S(t)$ and $f(t)$ in terms of $h(t)$: since $f(t) = -S'(t)$, (A.0.2) implies that

$$h(x) = -\frac{S'(x)}{S(x)} = -\frac{d}{dx} \log S(x).$$

Thus,

$$\log S(x)|_0^t = -\int_0^t h(x)dx.$$

Since $S(0) = 1$, we find that

$$S(t) = \exp\left(-\int_0^t h(x)dx\right) = \exp\left(-A(t)\right), \tag{A.0.3}$$

where

$$A(t) = \int_0^t h(x)dx$$

is the *cumulative hazard function*. The cumulative hazard function is an important quantity in survival analysis and plays a central role in this appendix. Note that, $S(0) = 1$, $A(0) = 0$, and $S(\infty) = 0$, $A(\infty) = \infty$. Sometimes, we may assume that $A(\infty) < \infty$.

Example A.0.1 Let T be an exponential random variable with mean $\lambda > 0$. The cumulative distribution function of T is $F(t) = \int_0^t e^{-x/\lambda} = 1 - e^{-t/\lambda}$. The survival function of T is $S(t) = 1 - F(t) = e^{-t/\lambda}$. The hazard function is

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\lambda}e^{-t/\lambda}}{e^{-t/\lambda}} = \frac{1}{\lambda}$$

and the cumulative hazard function is $A(t) = t/\lambda$. Note that, the hazard function is independent of t.

Next we assume that T is a discrete random variable. Suppose that T takes values $t_1, t_2, \ldots$, with $0 \leq t_1 < t_2 < \ldots$. Let $f(t_j) = \Pr\{T = t_j\} \quad j = 1, 2, \ldots$ be the probability mass function. The survival function is $S(t) = \Pr\{T > t\} = \sum_{j:t_j>t} f(t_j)$. Note that S is a left-continuous, nonincreasing step function, with $S(0) = 1$ and $S(\infty) = 0$. The discrete time hazard function is defined as:

$$\begin{aligned} h(t_j) &= \Pr\{\text{an individual fails exactly at time } t_j \text{ given that this has not occurred earlier}\} \\ &= \Pr\{T = t_j | T \geq t_j\} \\ &= \frac{f(t_j)}{S(t_j)} \quad j = 1, 2, \ldots. \end{aligned} \tag{A.0.4}$$

As in the continuous case, the probability mass function, the survival function, and the hazard function can be used to specify the distribution of T. Since $f(t_j) = S(t_j) - S(t_{j+1})$, (A.0.4) implies that

$$h(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)} \quad j = 1, 2, \ldots,$$

and thus

$$S(t) = \prod_{j:t_j<t} [1 - h(t_j)]. \tag{A.0.5}$$

The cumulative hazard function can be defined in two ways in the discrete case. One is by analogy with (A.0.3), as $A_1(t) = -\log S(t)$, where $S(t)$ is given by (A.0.5). The second is $A_2(t) = \sum_{j:t_j<t} h(t_j)$. Note that $A_1(t) \neq A_2(t)$. To see this, consider the following example.

Example A.0.2 Let T be a discrete random variable with a *Geometric*(p) distribution, i.e

$$\Pr\{T = j\} = p(1-p)^{j-1}, \quad \text{for } j = 1, 2, \ldots.$$

The hazard function is

$$h(j) = \Pr\{T = j | T \geq j\} = \frac{p(1-p)^{j-1}}{\sum_{k=j+1}^{\infty} p(1-p)^{j-1}} = p.$$

The survival function is

$$S(j) = \sum_{k=j+1}^{\infty} p(1-p)^{k-1} = \frac{p(1+p)^j}{1-(1-p)} = (1-p)^j.$$

Hence,

$$A_1(t) \;=\; -\log S(t) = -\sum_{j=1}^{t-1} \log[1-h(j)] = -\sum_{j=1}^{t-1} \log[1-p] = -(t-1)\log[1-p]$$

and

$$A_2(t) \;=\; \sum_{j=1}^{t-1} h(t_j) = \sum_{j=1}^{t-1} p = (t-1)p.$$

Clearly, $A_1(t) \neq A_2(t)$.

From the preceding discussion, it is clear that there is a need to consider a general way to handle continuous, discrete, and mixed distributions within a single framework so that the definition of $A(t)$ is unified in all cases. To achieve this, we introduce two types of integrals: the *Riemann-Stieltjes integrals* and the *product integrals.*

Let $G(u)$ be a nondecreasing, right-continuous function with left-hand limits and a finite number of discontinuities in any finite interval. Assume that $g(u) = G'(u)$ exists (except at the points of discontinuity of G) and that at a point a_j of discontinuity we have $G(a_j) - G(a_j^-) = g_j$, where $G(a^-) = \lim_{\Delta a \to 0} G(a - \Delta a)$. The *Riemann-Stieltjes integral* with respect to G over the interval $(a, b]$ is defined as

$$\int_{(a,b]} dG(u) = \int_a^b g(u)du + \sum_{j: a < a_j \leq b} g_j, \tag{A.0.6}$$

where the first integral on the right-hand side of (A.0.6) is a Riemann integral. We can think of dG as being equal to $g(u)du + G(u) - G(u^-)$.

In general, a cumulative distribution function $F(t) = \Pr\{T \leq t\}$ is a right-continuous, nondecreasing function, with jumps at points a_j for which $\Pr\{T = a_j\} = f_j > 0$. Let $f(u) = F'(u)$ at points u where F is continuous. Then, (A.0.6) gives $\Pr\{a < T \leq b\}$ as

$$F(b) - F(a) \;=\; \int_{(a,b]} dF(u) = \int_a^b f(u)du + \sum_{j:a<a_j\leq b} f_j.$$

If F is continuous, there are no jump points, and if T has a discrete distribution, then F is a step function with $f(u) = 0$ at all continuity points.

To give a general treatment of the hazard function, we introduce the product integral. Let $a = u_0^n < u_1^n < \cdots < u_{m_n}^n = b$ be partition $(a, b]$, with $\Delta u_i^n = u_i^n - u_{i-1}^n$ and $\max_{i\leq m_n}(\Delta u_i^n) \to 0$ when $n \to \infty$. The product integral with respect to a function G as defined earlier is

$$\prod_{(a,b]} \{1 + dG(u)\} = \lim_{n\to\infty} \prod_{i=1}^{m_n} \left\{1 + G(u_i^n) - G(u_{i-1}^n)\right\}. \tag{A.0.7}$$

In other words, the product integral is a product of many terms, all or almost all of them being close to 1. A close understanding of the product integral can be obtained by comparing it to the Riemann integral. As the Riemann integral is a limit of a finite sum, the product integral is a limit of finite product. For more details about product integrals, see Gill and Johansen (1990).

If G is continuous for on $(a, b]$, then $dG(u) = g(u)du$ and (A.0.7) gives

$$\begin{aligned}\prod_{(a,b]} \{1 + g(u)du\} &= \lim_{n\to\infty} \prod_{i=1}^{m_n} \{1 + g(u_i^n)\Delta u_i^n + o(\Delta u_i^n)\} \\ &= \lim_{n\to\infty} \prod_{i=1}^{m_n} \{1 + g(u_i^n)\Delta u_i^n\},\end{aligned}$$

where $o(x)$ means a function $w(x)$ such that $w(x)/x \to 0$. Observe that, for small Δu_i^n,

$$\log\{1 + g(u_i^n)\Delta u_i^n\} = g(u_i^n)\Delta u_i^n + o(\Delta u_i^n).$$

Taking the log of the product integral, we see that in the continuous case

$$\begin{aligned}
\log \prod_{(a,b]} \{1 + dG(u)\} &= \log \lim_{n\to\infty} \prod_{i=1}^{m_n} \{1 + g(u_i^n)\Delta u_i^n\} \\
&= \lim_{n\to\infty} \sum_{i=1}^{m_n} \log \{1 + g(u_i^n)\Delta u_i^n\} \\
&= \lim_{n\to\infty} \sum_{i=1}^{m_n} \{g(u_i^n)\Delta u_i^n + \Delta u_i^n\} \\
&= \lim_{n\to\infty} \sum_{i=1}^{m_n} \{g(u_i^n)\Delta u_i^n\}\,.
\end{aligned}$$

where for the last equality we used the definition of Riemann integral. We get

$$\prod_{(a,b]} \{1 + dG(u)\} = \exp\left\{\int_a^b g(u)du\right\}. \tag{A.0.8}$$

Thus, equation (A.0.8) relates the product integral and the Riemann integral.

If G has jumps at points a_j $(j = 1, 2, \ldots)$ of sizes g_j, then (A.0.7) gives

$$\prod_{(a,b]} \{1 + dG(u)\} = \prod_{(a,b]} \{1 + g(u)du\} \prod_{j:a<a_j\le b} (1 + g_j)\,. \tag{A.0.9}$$

Note that if G is a step function, then $g(u) = 0$ at all continuity points and the first term on the right side of (A.0.9) disappears.

We are now in a position to consider the hazard function. Let $h(u) = f(u)/S(u)$ be the hazard function of T at points where $F(u)$ (or $S(u)$) is continuous, and $h_j = \Pr\{T = a_j | T \ge a_j\}$ be the value of the hazard function at times a_j for which a jump in F occurs. The cumulative hazard function is then defined by a Riemann-Stieltjes integral of the form (A.0.6):

$$A(t) = \int_0^t dA(u) = \int_0^t h(u)du + \sum_{j:a_j\le t} h_j. \tag{A.0.10}$$

The cumulative hazard rate in its general form can alternatively be defined as described in (2.5.1).

Given the cumulative hazard function, we can obtain the survival function as follows: partition the interval $(0, t]$ into subintervals $0 = u_0 < u_1 < \cdots < u_m = t$. To survive from time 0 to time t, an individual needs to survive all the intermediate subintervals. Thus,

$$\begin{aligned} S(t) = \Pr\{T > t\} &= \frac{\Pr(T > u_1)}{\Pr(T > u_0)} \cdot \frac{\Pr(T > u_2)}{\Pr(T > u_1)} \cdot \cdots \cdot \frac{\Pr(T > u_m)}{\Pr(T > u_{m-1})} \\ &= \prod_{i=1}^{m} \Pr\{T > u_i | T > u_{i-1}\} \end{aligned} \tag{A.0.11}$$

Now for $\Delta u_i = u_i - u_{i-1}$ sufficiently small, $(u_{i-1}, u_i]$ contains either 0 or 1 jump points, and

$$\begin{aligned} \Pr\{T > u_i | T > u_{i-1}\} &= 1 - \frac{\Pr\{u_{i-1} < T \leq u_i\}}{\Pr\{T \geq u_{i-1}\}} \\ &= 1 - [A(u_i) - A(u_{i-1})] + o(\Delta u_i). \end{aligned}$$

Therefore, by (A.0.11) and (A.0.7),

$$S(t) = \Pr\{T > t\} = \prod_{(0,t]} [1 - dA(u)]. \tag{A.0.12}$$

Relationships (A.0.10) and (A.0.12) apply to all types of distributions. We get (A.0.3) from (A.0.12) in the case of a continuous distribution by using (A.0.8), and we get (A.0.5) for a discrete distribution by using (A.0.9).

Appendix B

Convergence of Random Measures

In this appendix, we discuss some properties of the random measures. Let E be a Polish space and $\mathcal{B}(E)$ be the Borel $\sigma-$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 $\sigma-$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 $\sigma-$algebra generated by the topology of vague convergence. If $\mu_n, \mu \in M_+(E)$, we say that $(\mu_n)_n$ converges vaguely to μ (and we write $\mu_n \xrightarrow{v} \mu$) 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 $(\Omega, \mathcal{A}, P)$ with values in $(M_+(E), \mathcal{M}_+(E))$. If ξ_n, ξ are random measures on E, we say that $(\xi_n)_n$ converges in distribution to ξ (and 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.

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

We say that $(\xi_n)_n$ converges vaguely almost surely to to ξ (and write $\xi_n \xrightarrow{a.s.} \xi$) if

there exists a set $\widetilde{\Omega} \in \mathcal{A}$ with $P(\widetilde{\Omega}) = 1$ such that $\forall \omega \in \widetilde{\Omega}$, $\xi_n(\omega, \cdot) \xrightarrow{v} \xi(\omega, \cdot)$, i.e.

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

The space $M_+(E)$ endowed with the vague topology is a complete separable metric space (Resnick, 1987, page 147). For more details about random measures see Kallenberg (1983).

Bibliography

[1] Antoniak, C. E. (1974). Mixtures of Dirichlet Processes With Applications to Nonparametric Problems. *The Annals of Statistics*, 2, 1152-1174.

[2] Argiento, R., Guglielmi, A., and Pievatolo, A. (2010). Bayesian density estimation and model selection using nonparametric hierarchical mixtures. C*omputational Statistics & Data Analysis*, 54, 816-832.

[3] Aalen, O. O. , Borgan, O., and Gjessing, H. K (2008). *Survival and Event History Analysis: A Process Point of View*. Springer.

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

[5] Balakrishnan, N., and Lai, C. (2009). *Continuous Bivariate Distributions*. Springer-Verlag.

[6] Banjevic, D., Ishwaran, H., and Zarepour, M. (2002). A recursive method for functionals of Poisson processes. *Bernoulli*, 8, 295-311.

[7] Bartle, R. G., and Sherbert, D. R. (2009). *Bartle and Sherbert*, third edition. John Wiley & Sons, Inc.

[8] Berger, J. O., and Guglielmi, A. (2001). Bayesian testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96, 174-184.

[9] Bickel, P. J., and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9, 1196-1217.

[10] Bickel, P.J. and Wichura, M.J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *The Annals of Mathematical Statistics*, 42, 1656-1670.

[11] Bickel, P. J., and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *The Annals of Mathematical Statistics*, 42, 1656-1670.

[12] Billingsley, P. (1999). *Convergence of Probability Measures*, third edition. John Wiley & Sons, Inc.

[13] Blackwell, D., and MacQueen, J. B. (1973). Ferguson Distributions via Polya Urn Schemes. *The Annals of Statistics*, 1, 353-355.

[14] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

[15] Bondesson, L., (1982). On simulation from infinitely divisible distributions. *Advances in Applied Probability*, 14, 885-869.

[16] Breiman, L. (1968). *Probability*. Addison-Wesley, Reading, Massachusettes.

[17] Brix, A., (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31, 929-953.

[18] Buntine, W., and Hutter, M. (2010). A Bayesian review of the Poisson-Dirichlet process. http://arxiv.org/abs/1007.0296.

[19] Carlton, M. A. (1999). Applications of the two-parameter Poisson-Dirichlet distribution. Unpublished Ph.D. thesis, Department of Statistics, University of California, Los Angeles.

[20] Chib, S., and Hamilton, B. H. (2002). Semiparametric Bayes Analysis of Longitudinal Data Treatment Models. *Journal of Econometrics*, 110, 67-89.

[21] Cifarelli, D. M., and Regazzini, E. (1990). Distribution functions of means of a Dirichlet process. *The Annals of Statistics*, 18, 429-442.

[22] Connor, R. J., and Mosimann, I. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64, 194-206.

[23] Cont, R., and Tankov, P. (2004). *Financial Modelling with Jump Processes*. Chapman & Hall/CRC Press.

[24] Daley, D. J., and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes, vol. II*, second edition. Springer-Verlag, New York.

[25] Damien, P., Laud, P., and Smith, A. F. M. (1995). Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *Journal of the Royal Statistical Society: Series B*, 57, 547-563.

[26] Damien, P., Laud, P., and Smith, A. F. M. (1996). Implementation of Bayesian nonparametric inference based on beta processes. *Scandinavian Journal of Statistics*, 23, 27-36.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer.

[27] De Blasi, P. (2007). Simulation of the beta-Stacy process with application to analysis of censored data. *Encyclopedia of Statistics*.

[28] De Blasi, P., Favaro, S., and Muliere, P. (2010). A class of neutral to the right priors induced by superposition of beta processes. *Journal of Statistical Planning and Inference*, 140, 1563-1575.

[29] de Haan, L., and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer, New York.

[30] Dey, J., Erickson, R .V., and Ramamoorthi, R. V. (2003). Some aspects of neutral to the right priors. *International Statistical Review*, 71, 383-401.

[31] Diaconis, P. , and Kemperman, J. B. K. (1996). Some new tools for Dirichlet priors. *In Bayesian Statistics V: Fifth Valencia International Meeting on Bayesian Statistics*, Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid , and A. F. M. Smith, 97-106.

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

[33] Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge.

[34] Dunson, D. B. (2005). Bayesian Semiparametric Isotonic Regression for Count Data. *Journal of the American Statistical Association*, 100, 618-627.

[35] Dunson, D. B., Herring, A. H., and Mulheri-Engel, S. A. (2008). Bayesian Selection and Clustering of Polymorphisms in Functionally-Related Genes. *Journal of the American Statistical Association*, 103, 534-546.

[36] Epifani, I., Lijoi, A., and Prünster, I. (2003). Exponential functionals and means of neutral-to-the-right priors. *Biometrika*, 90, 791-808.

[37] Escobar, M. D., and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90, 577-588.

[38] Favaro, S., Lijoi, A., Mena, R., and Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society Series B*, 71, 993-1008.

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

[40] Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1, 209-230.

[41] Ferguson, T. S. (1974). Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*, 2, 615-629.

[42] Ferguson, T. S. (1983). Bayesian Density Estimation by Mixtures of Normal Distributions. In *Recent Advances in Statistics*, Eds. Rizvi, H. and Rustagi. J., New York, Academic Press, 287-302.

[43] Ferguson, T. S., and Klass, M. J. (1972). A Representation of Independent Increment Processes without Gaussian Components. *The Annals of Mathematical Statistics*, 1, 209-230.

[44] Ferguson, T. S., and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, 7, 163-186.

[45] Florens, J. P., Richard, J. F., and Rolin, J. M. (1996). Bayesian Encompassing Specification Tests of a Parametric Model Against a Nonparametric Alternative. Technical Report, Universitsé Catholique de Louvain, Belgium.

[46] Gibbs, A., and Su, E. F. (2002). On choosing and Bounding Probability metrics. *International Statistical Review*, 70, 419-435.

[47] Gill, R. D., and Johansen, S. (1990). A survey of product integration with a view towards survival analysis. *The Annals of Statistics*, 18, 1501-1555.

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

[49] Grandell, J. (1977). Point processes and random measures. *Advances in Applied Probability*, 9, 502-526.

[50] Griffin, J. E., and Walker, S. G. (2011). Posterior simulation of Normalised Random Measure mixtures. *Journal of Computational and Graphical Statistics*, 20, 241-259.

[51] Hamada, M. S., Wilson, A. G., Reese, C., and Martz, H. F. (2008). *Bayesian Reliability*. Springer, New York.

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

[53] Huber, P. (1981). *Robust Statistics*. Wiley, New York.

[54] Ishwaran, H., and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96, 161-173.

[55] Ishwaran, H., and Zarepour, M. (2002). Exact and Approximate Sum Representations for the Dirichlet Process. *The Canadian Journal of Statistics*, 30, 269-283.

[56] Ishwaran, H., and Zarepour, M. (2009). Series Representations for Multivariate Generalized Gamma Processes via a Scale Invariance Principle. *Statistica Sinica*, 19, 1665-1682.

[57] James, L.F., Lijoi, A., and Prünster, I. (2010). On the posterior distribution of classes of random means. *Bernoulli*, 1, 155-180.

[58] Jiang, J. (2009). *Large Sample Techniques for Statistics*. Springer.

[59] Jang, G. H., Lee, J., and Lee, S. (2010). Posterior consistency of species sampling priors. *Statistica Sinica*, 20, 581-593.

[60] Kacperczyk, M., Damien, P., and Walker, S. G. (2003). A new class of Bayesian semiparametric models with applications to option pricing. Technical Report, University of Michigan Business School.

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

[62] Kim, N., and Bickel, P. (1987). The limit distribution of a test statistic for bivariate normality. *Statistica Sinica*, 13, 327-349.

[63] Kottas, A., Branco, M. D., and Gelfand, A. E. (2002). A Nonparametric Bayesian Modeling Approach for Cytogenetic Dosimetry. *Biometrics*, 58, 593-600.

[64] Kottas, A., and Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96, 1458-1468.

[65] Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 20, 1222-1235.

[66] Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, second edition. Wiley & Sons, Inc., Hoboken, New Jersey.

[67] Lee, J. (2007). Sampling methods for neutral to the right processes. *Journal of Computational and Graphical Statistics*, 16, 656-671.

[68] Lee, E. T., and Wang, J. W. (2003). *Statistical methods for survival data analysis*, third edition. Wiley-Interscience, Hoboken.

[69] Lee, J., and Kim, Y. (2004). A new algorithm to generate beta processes. *Computational Statistics and Data Analysis*, 25, 401-405.

[70] Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *Journal of the American Statistical Association*, 100, 1278-1291.

[71] Lijoi, A., Regazzini, E. (2004). Means of a Dirichlet process and multiple hypergeometric functions. *The Annals of probability*, 32, 1469-1495.

[72] Lijoi, A., and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, Eds. Hjort, N. L., Holmes, C. C. Müller, P., Walker, S. G., Cambridge University Press, Cambridge, 80-136.

[73] Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12, 351-357.

[74] Lo , A. Y. (1987). A large sample study of the Bayesian bootstrap. *The Annals of Statistics*, 15, 360-375.

[75] MacEachern, S. N., and Müller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, 7, 223-238.

[76] Muliere, P., and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet prior. *The Canadian Journal of Statistics*, 26, 283-297.

[77] Navarrete, C., Quintana, F. A., and Müller, P. (2008). Some issues on nonparametric Bayesian modeling using species sampling models. S*tatistical Modelling*, 8, 3-21.

[78] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 249-265.

[79] Papaspiliopoulos, O., and Roberts, G. O. (2008). Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika*, 95, 169-186.

[80] Pestman, W. R. (2009). *Mathematical Statistics*, second edition. Walter de Gruyter, Berlin.

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

[82] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

[83] Regazzini, E., Guglielmi, A., Di Nunno, G. (2002). Theory and numerical analysis for exact distribution of functionals of a Dirichlet process. *The Annals of Statistics*, 30, 1376-1411.

[84] Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York.

[85] Resnick, S. I. (1998). *A Probability Path*. Birkhauser.

[86] Rosiński, J. (2001). Series representations of Lévy processes from the perspective of point processes. In *Lévy Processes - Theory and Applications*, Eds. Barndorff-Nielsen, O. E., Mikosch, T., and Resnick. S. I., Birkhauser, Boston, 401-415.

[87] Robert, C. P., and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer.

[88] Rubin, D. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9, 130-134.

[89] Sato, K. I.(1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, Cambridge.

[90] Sethuraman, J., and Tiwari, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. *Statistical Decision Theory and Related Topics III*, 2, 305-315.

[91] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

[92] Sinha, D. (1997). Time discrete Beta-process model for interval censored survival data. *Canadian Journal of Statistics*, 25, 445-56.

[93] Susarla, V., and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71, 897-902.

[94] Swartz, T. B. (1999). Nonparametric goodness-of-fit. *Communications in Statistics: Theory and Methods*, 28, 2821-2841.

[95] Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 985-992.

[96] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[97] van der Vaart, A.W., and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New York.

[98] Walker, S., and Damien, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic process. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Eds. Dey, D., Müller, P., and Sinha, D., Springer-Verlage, 243-254.

[99] Walker, S. G., and Damien, P. (2000). Representation of Lévy processes without Gaussian components. *Biometrika*, 87, 477-483.

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

[101] Wilkes, S. S. (1963). *Mathematical Statistics*. John Wiley & Sons, Inc.

[102] Wolpert, R. L., and Ickstadt, K. (1998). Simulation of Lévy random fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics*. Eds. Day, D., Nuller, P., and Sinha, D., Springer, 2237-242.

[103] Wrench, J. W. Jr. (1968). Concerning two series for the gamma function. *Mathematics of Computation*, 22, 616-626.

[104] Zarepour, M., and Al Labadi, L. (2012). On a Rapid Simulation of the Dirichlet Process. *Statistics and Probability Letters*, 82, 916-924.

[105] Zarepour, M., Bedard, T., and Dabrowski, A. (2008). A Return and Value at Risk Using the Dirichlet Process. *Applied Mathematical Finance*, 3, 205-218.

[106] Zolotarev, V. M. (1997). *Modern Theory of Summation of Random Variables*. VSP, Utrecht, The Netherlands.