Eswar G. Phadia

Prior Processes and Their Applications

Nonparametric Bayesian Estimation

Springer

Prior Processes and Their Applications

Eswar G. Phadia

Prior Processes and Their Applications

Nonparametric Bayesian Estimation

Springer

Eswar G. Phadia
Department of Mathematics
William Paterson University of New Jersey
Wayne, NJ, USA

ISBN 978-3-642-39279-5 ISBN 978-3-642-39280-1 (eBook)
DOI 10.1007/978-3-642-39280-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013945285

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The foundation of the subject of nonparametric Bayesian inference was laid in two technical reports: a 1969 UCLA report by Thomas S. Ferguson (later published in 1973 as a paper in the *Annals of Statistics*) entitled "A Bayesian analysis of some nonparametric problems"; and a 1970 report by Kjell Doksum (later published in 1974 as a paper in the *Annals of Probability*) entitled "Tailfree and neutral random probabilities and their posterior distributions". In view of simplicity with which the posterior distributions were calculated (by updating the parameters), the Dirichlet process became an instant hit and generated quite an enthusiastic response. During the decades of 1970s and 1980s, hundreds of papers were published in developing nonparametric Bayesian procedures to handle many inferential problems. These publications may be considered as "pioneers" in championing the Bayesian methods and opening a vast unexplored area in solving nonparametric problems. A review article (Ferguson et al. 1992) summarized the progress of the two decades. However, the paper was not meant to provide details but just an overview. Moreover, since then several new prior processes and their applications have appeared in technical publications. Also in the last decade there has been a renewed interest in the applications of variants of the Dirichlet process in modeling large scale data (see for example the recent papers by Chung and Dunson 2011, and Rodriguez et al. 2010 and references cited therein; and a volume of essays "Bayesian Nonparametric" edited by Hjort et al. 2010). For these reasons there seems to be a need for a single source of the material published on this topic during the earlier decades. This is a prime motivator for undertaking the present task.

The objective of this monograph is to assemble and consolidate the scattered material on various prior processes, their properties and their numerous applications, in solving Bayesian inferential problems based on data that may possibly be right censored, sequential or quantal response data. Emphasis is placed on the Dirichlet process as well as other prior processes that have been discovered through 1990s and their applications. We anticipate that it would serve as a one-stop resource for future researchers. In that spirit, first various processes are introduced and their properties are stated. Thereafter, the focus is to present various applications in estimation of distribution and survival functions, estimation of density functions and hazard rates,

empirical Bayes, hypothesis testing, covariate analysis, and many other applications. A major requirement of Bayesian analysis is its analytical tractability. Since the Dirichlet process possesses the conjugacy property, it has simplicity and ability to get results in a closed form. Therefore, most of the applications that were published soon after Ferguson's paper, are based on the Dirichlet process. Unlike the trend in recent years where computational procedures are developed to handle large and complex data sets, the earlier procedures relied mostly on developing procedures in closed forms.

In addition, several new and interesting processes, such as, the Chinese restaurant process, Indian buffet process, and hierarchical processes have been introduced in the last decade with an eye toward applications in the fields outside mainstream statistics, such as machine learning, ecology, document classification, etc. Similarly, dependent and spatial Dirichlet processes are proposed to incorporate covariates and handle random effects models. They have roots in the Ferguson-Sethuraman infinite sum representation of the Dirichlet process and shed new light on the robustness of this approach. They are included here without going into much details but a long list of references is included for the reader to explore relevant areas of interest further.

This material is an outgrowth of my lecture notes developed during the week long lectures I gave at Zhongshen University in China in 2007 on this topic, followed by lectures at universities in India, Singapore and Jordan. Obviously, the choice of material included and the style of presentation solely reflects my preferences. This manuscript is not expected to include all the applications, but references are given, wherever possible for additional applications. The mathematical rigor is limited as it has already been dealt with in the theoretical book by Ghosh and Ramamoorthi (2003). Therefore, many theorems and results are stated without proofs and the questions regarding existence, consistency and convergences are skipped. To conserve space, numerical examples are not included but referred to the papers originating those specific topics. For these reasons, the notations of the originating papers are preserved so that the reader may find it easy to migrate to the original publications as needed.

Computational procedures that make nonparametric Bayesian analysis feasible when closed forms of solutions are impossible or complex, are becoming increasingly popular in view of the availability of inexpensive and fast computation power. In fact they are indispensable tools in modeling large scale and high dimensional data. There are numerous papers published in the last two decades that discuss them in great details and algorithms are developed to simulate the posterior distributions so that the Bayesian analysis can proceed. These aspects are covered extensively in books by Ibrahim et al. (2001) and Dey et al. (1998). To avoid duplication, they are not discussed here. Some newer applications are also discussed in the book of essays edited by Hjort et al. (2010). We refer the reader to the these books. The papers by Chung and Dunson (2011) and Rodriguez et al. (2010) and references cited therein, should also prove useful in this regard.

Since this book discusses various prior processes, their properties and inferential procedures in solving problems encountered in practice, it is ideal to serve as

a comprehensive introduction to the subject of nonparametric Bayesian inference. It is to be considered as a complement to the book authored by Ghosh and Ramamoorthi (2003) but at a less rigorous level. It may be viewed as something in between their theoretical book and the books by Ibrahim et al. (2001) and Dey et al. (1998).

The first chapter is devoted to introducing various prior processes, their formulation and their properties. The sequencing of these priors reflects mostly the order in which they were developed. The Dirichlet process and its immediate generalizations are presented first. The neutral to the right processes and the processes with independent increments, which form the basis for other processes are discussed next. They are key in the development of processes that include beta, gamma and extended gamma processes, which are proposed primarily to address specific applications in the reliability theory. Beta-Stacy process which generalizes the Dirichlet process is discussed thereafter. Following that, tailfree and Polya tree processes are presented which are especially convenient for estimating density functions, and to place greater weights, where it is deemed appropriate, by selecting suitable partitions in developing the prior. Lijoi and Prünster's (2010) recent paper tie many of these processes in presenting a general unifying framework in terms of the completely random measures (Kingman 1967). Finally, some additional processes that have been discovered in recent years (mostly variants of existing processes) and found to be useful in practice are mentioned. They have origin in the Ferguson-Sethuraman infinite sum representation in which the weights are constructed by a stick-breaking construction. They are collectively called here as *Ferguson-Sethuraman processes* and include dependent and spatial Dirichlet processes, Pitman-Yor process, Chinese restaurant and Indian buffet processes, etc.

The second chapter contains various applications that cover multitudes of fields such as, estimation, hypothesis testing, empirical Bayes, density estimation, bioassay, etc. They are grouped according to the inferential task they signify. Since, a major part of efforts have been devoted to the estimation of the distribution function and its functional, they receive significant attention. This is followed by confidence bands, two-sample problems and other applications.

The third chapter is devoted to presenting inferential procedures based on censored data. Heavy emphasis is given to the estimation of survival function since it plays an important role in the survival data analysis. Estimation procedures based on different priors and under various sampling schemes are also included. This is followed by other examples which include estimation procedures in certain stochastic process models, Markov Chains, and competing risks models. Finally, estimation of the survival function in the presence of covariates is presented.

Since this book avoids deeper technical details, it should therefore be accessible to first time researchers and graduate students venturing into this interesting, fertile and promising field. As evident by the recent increased interest in using nonparametric Bayesian methods in modeling data, the field is wide open for new entrants. As such, it is my hope that this attempt will serve the purpose it was intended for, namely, to make such techniques readily available via this comprehensive but sim-

ple monograph. At the least, the reader will gain familiarity with many successful attempts in solving nonparametric problems from a Bayesian point of view in wide ranging areas of applications.

Wayne, USA Eswar G. Phadia

Acknowledgements

Such tasks as writing a book takes a lot of patience and hard work. My undertaking was no exception. However, I was fortunate to receive lot of encouragement, advice and support on the way.

I had the privilege of support, collaboration and blessing of Tom Ferguson, the architect of nonparametric Bayesian statistics, which inspired me to explore this area during the early years of my career. Recent flurry of activity in this area renewed my interest and prompted me to undertake this task. I am greatly indebted to him. Jagdish Rustagi brought to my attention in 1970 a pre-publication copy of Ferguson's seminal 1973 paper which led to my doctoral dissertation at the Ohio State University. I am eternally grateful to him for his advice and support in shaping my research interests which stayed on track with me for the last 40 years except for a 10-year stint in administration.

The initial template of the manuscript was developed as lecture notes for presentation at Zhongshen University in China at the behest of Qiqing Yu of Binghamton University. I thank him and thank Zhongshen University faculty and staff for their hospitality. The final shape of the manuscript took place during my sabbatical at the University of Pennsylvania's Wharton School of Business. I gratefully thank Edward George and Larry Brown of the Department of Statistics for their kindness in providing me the necessary facilities and intellectual environment (and not to forget complimentary lattes) which enabled me to advance my endeavor substantially. I also take pleasure in thanking Bill Strawderman, for his friendship of over 30 years, sound advice and useful discussions during my earlier sabbatical and frequent visits to Rutgers University campus. My sincere thanks to anonymous reviewers for their valuable comments and suggestions which proved useful and improved the manuscript by updating it to reflect the current level of activity in nonparametric Bayesian field. I must have exchanged scores of emails and had countless conversations with Dr. Eva Hiripi, Associate Editor of Springer during the last year. Her patience, understanding and helpful suggestions were instrumental in shaping the final product in the present form. My heartfelt thanks to her. The production staff at Springer including Ulrike Stricker, and at VTeX including Edita Baronaitė did

a fantastic job in detecting missing references and producing the final product. They deserve my thanks.

This task could not have been accomplished without the support of my institution in terms of ART awards over a period of number of years, and cooperation of my colleagues. In particular, I thank my colleague Jyoti Champanerker for creating the flow chart of Chap. 1. Finally, I owe thanks to my wife and companion Jyotsna, my daughter Sonia, and my granddaughter Alexis, who at her tender age, provided me happiness and stimulus to keep going when early retirement would have been a preferred option.

Contents

Chapter 1
Prior Processes

1.1 Prior Processes—An Overview

1.1.1 Introduction

In this section we give an overview of the various processes that have been developed to serve as prior distributions in the treatment of nonparametric problems from a Bayesian point of view. We indicate their relationship with each other, discuss circumstances in which they are appropriate to use and their relative merits and drawbacks in solving inferential problems. In subsequent sections we provide more details on many of them and state their properties. To preserve the historical perspective, they are arranged in the order of their discovery and development.

In the Bayesian approach, the unknown distribution function from which the sample arises is itself considered as a parameter. Thus, we need to construct prior distributions on the space of all distribution functions, to be denoted by $\mathcal{F}(\chi)$, defined on a sample space χ, or on all probability measures, Π defined on certain probability space, $(\mathfrak{X}, \mathcal{A})$, where $\mathcal{A}$ is σ-field of subsets of $\mathfrak{X}$.

Consider for example the Bernoulli distribution which assigns mass p to 0 and $1-p$ to $1, 0 < p < 1$. In this case the sample space is $\chi = \{0, 1\}$ and the space of all distributions consists of distributions taking jumps of size p at 0 and $1-p$ at 1 or $\mathcal{F} = \{F : F(t) = pI[t \geq 0] + (1-p)I[t \geq 1]\}$, where $I[A]$ is an indicator function of the set A. Here the random distribution is characterized by treating p as random. In this case, a prior on $\mathcal{F}(\chi)$ may then be specified by simply assigning a prior distribution to p on Π, say uniform, $U(0, 1)$ or a beta distribution, $Be(a, b)$ with parameters $a > 0$, and $b > 0$. A prior distribution on $\mathcal{F}(\chi)$ or Π will be denoted by $\mathfrak{P}$ whenever needed.

As a second example, consider the multinomial experiment with the sample space, $\chi = \{1, 2, \ldots, k\}$. In this case, $\mathcal{F}(\chi)$ is the space of all distribution functions corresponding to a $(k-1)$-dimensional probability simplex $S_k = \{(p_1, p_2, \ldots, p_k) : 0 \leq p_i \leq 1, \sum_{i=1}^{k} p_i = 1\}$ of probabilities. Then a prior distribution $\mathfrak{P}$ can be specified on $\mathcal{F}(\chi)$ by defining a measure on S_k which yields the joint distribution of

E.G. Phadia, *Prior Processes and Their Applications*,
DOI 10.1007/978-3-642-39280-1_1, © Springer-Verlag Berlin Heidelberg 2013

$(p_1, p_2, \ldots, p_k)$, say, the Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, where $\alpha_i \geq 0$ for $i = 1, 2, \ldots, k$. However, we will mostly be dealing with $\chi = \mathbb{N}$ or R.

While the distribution function is the parameter of primary interest in nonparametric Bayesian analysis, at times it is more convenient to discuss the prior process in terms of a probability measure P instead of the corresponding distribution function. The Dirichlet process is defined in this way. However, many of the applications are given in terms of the distribution function or its functional.

Defining a prior for an unknown F on $\mathcal{F}$ or for a P on Π give rise to some theoretical difficulties (see for example, Ferguson 1973). The challenge therefore is how to circumvent these difficulties and define viable priors. The priors so defined should have, according to Ferguson (1973), two desirable properties: The support should be large enough to accommodate all shades of belief; and the posterior distribution, given a sample should be analytically tractable so that the Bayesian analysis can proceed. The second desirable property has led to a search of priors which are conjugate, i.e. the posterior has the same structure except for the parameters. This would facilitate posterior analysis since one needs only to update the parameters of the prior. However, it could also be construed as a limitation in choice of priors. A balance between the two would be preferable. (Antoniak 1974, adds some more desirable properties.) In addition, since the Bayesian approach involves incorporating prior information to make inferential procedures more efficient, it may be considered as an extension of the classical maximum likelihood approach. Therefore, it is natural to expect that the results of the procedures so developed should reduce to those obtained through the classical methods when the prior information, reflected in parameters of the priors, tends to nil. It will be seen that this is mostly true, especially in the case of Dirichlet and neutral to the right processes.

Prior to 1973, the subject area of nonparametric Bayesian inference was nonexistent. Earlier attempts in defining such priors on $\mathcal{F}$ can be traced to Dubins and Freedman (1963) whose methods to construct a random distribution function resulted in a singular continuous distribution, with probability one. In dealing with a bioassay problem, Kraft and van Eeden (1964) constructs a prior in terms of the joint distribution of the ordinates of F at certain fixed points of a countable dense subset of the real line. In Kraft (1964), the author describes a procedure of choosing a distribution function on the interval [0, 1] which is absolutely continuous with probability one. Freedman (1963) introduced the notion of *tailfree* distributions on a countable space and Fabius (1964) extended the notion to the interval [0, 1]. But all these attempts had limited success because either the base was not sufficiently large or the solutions were analytically or computationally intractable.

Ferguson's landmark paper was the first successful attempt in defining a prior which met the above requirements. Encouraged by his success, several new prior processes have been proposed in the literature since then to meet specific needs. We review them briefly in this section and present them formally in subsequent sections.

1.1.2 Methods of Construction

During the earlier period of development, the method of placing a prior on $\mathcal{F}$ or Π can broadly be classified as based essentially on four different approaches. First one is through the joint distribution of random probabilities, and next two are based on different independence properties, and the last one is based on generating a sequence of exchangeable random variable using the generalized Polya urn scheme. The first three approaches are closely related to different properties of the Dirichlet distribution (see Basu and Tiwari 1982 for extensive discussion of these properties). However, in the last decade or so, several new processes have been developed which can be constructed via the countable mixture representation of a random probability, also known as the *stick-breaking* construction. These are described here informally without going into the underlying technicalities.

The first method is defined by Ferguson (1973) in terms of the joint distribution of probabilities of sets of a measurable partition of an arbitrary set. For any positive integer k, let $A_1, \ldots, A_k$ be a measurable partition of $\mathfrak{X}$ and let α be a nonnegative finite measure on $(\mathfrak{X}, \mathcal{A})$. A random probability measure P defined on $(\mathfrak{X}, \mathcal{A})$ is said to be a *Dirichlet process with parameter* α if the distribution of the vector $(P(A_1), \ldots, P(A_k))$ is Dirichlet distribution, $D(\alpha(A_1), \ldots, \alpha(A_k))$. In symbols it will be denoted as $P \in \mathcal{D}(\alpha)$. (In our presentation, we will ignore the distinction between a random probability P being a Dirichlet process and the Dirichlet process being a prior distribution for a random probability P on the space Π.) This approach was used in two immediate generalizations: one by Antoniak (1974) who defined the *mixtures of Dirichlet processes*, and the other by Dalal (1979a) who defined the *Dirichlet Invariant process. Kernel mixtures* (Lo 1984) and *Hierarchical Dirichlet processes* (Teh et al. 2006) are also outgrowth of this approach.

The second method is based on the property of independence of successive normalized increments of a distribution function F defined on the real line R. It is based on the Connor and Mosimann (1969) concept of neutrality for k-dimensional random vectors. For $m = 1, 2, \ldots$ consider the sequence of real numbers $-\infty < t_1 < t_2 < \ldots < t_m < \infty$. Doksum (1974) defines a random distribution function F as *neutral to the right* if for all m, the successive normalized increments $F(t_1), (F(t_2) - F(t_1))/(1 - F(t_1)), \ldots,$ are independent. Since a distribution function can be represented as $F(t) = 1 - \exp(-Y_t)$, where Y_t is a process with independent nonnegative increments, the neutral to the right processes can also be viewed in terms of the processes with independent nonnegative increments. Since the latter processes are well known, they became the main tool in defining a class of specific processes tailored to suit particular applications. Kalbfleisch (1978) defined a *gamma process*, Dykstra and Laud (1981) proposed an *extended gamma process*, Hjort (1990) developed a *beta process*, Thibaux and Jordan (2007) defined a *Hierarchical beta process*, and Walker and Muliere (1997a) introduced the *beta-Stacy process*.

The third method is based on a different independence property which corresponds to the tailfree property of the Dirichlet distribution. Let $\{\pi_n\}$ be a sequence of nested partitions of R such that π_{n+1} is a refinement of π_n, for $n = 1, 2, \ldots$.

Let $\{B_{m1}, \ldots, B_{mk_m}\}$ denote the partition π_m. Since the partitions are nested, then for $s < m$, there is one set in π_s that contains the set B_{mi} of π_m. This set will be denoted by $B_{s(mi)}$. A random probability P is said to be *tailfree* if the families $\{P(B_{1j}|B_{0(1j)}) : j = 1, \ldots, k_1\}, \ldots, \{P(B_{m+1\,j}|B_{m(mj)}) : j = 1, \ldots, k_{m+1}\}$ are independent, where $B_{0(1j)} = R$. That is, a random probability P is said to be *tailfree* if the sets of random variables $\{P(B|A) : A \in \pi_n \text{ and } B \in \pi_{n+1}\}$ for $n = 1, 2, \ldots$ are independent. Here $\pi_0 = R$. The random probability P is defined via the joint distribution of all the random variables $P(B|A)$. The origin of this process goes back to Freedman (1963) and Fabius (1964), but Doksum (1974) clarified the notion of tailfree and Ferguson (1974) gave a concrete example, thus formalizing the discussion in the context of a prior. *Tailfree* is a misnomer since the definition does not depend on the tails (Doksum 1974, attributes it to Fabius for pointing out this distinction). Doksum used the term F-neutral. However, we will use the term *tailfree* as it has become a common practice. The *Polya tree* processes developed more formally by Lavine (1992, 1994) and Mauldin et al. (1992), are a special case of tailfree processes in which *all* random variables are assumed to be independent.

As a fourth approach, Blackwell and MacQueen (1973) showed that a prior process can also be constructed by constructing a sequence of exchangeable random variables via the Polya urn scheme and then applying a theorem of de Finetti. In particular, they showed that the Dirichlet process can also be constructed in this way. The Polya urn scheme may be described as follows. Let $\chi = \{1, 2, \ldots, k\}$. We start with an urn containing α_i balls of color i, $i = 1, 2, \ldots, k$ (later extended to continuum of colors). Draw a ball at random of color i and define the random variable X_1 so that $P(X_1 = i) = \overline{\alpha}_i$, where $\overline{\alpha}_i = \alpha_i/(\sum_{i=1}^k \alpha_i)$. Now replace the ball with two balls of the same color and draw a second ball. Define the random variable X_2 so that $P(X_2 = j \mid X_1 = i) = (\alpha_j + \delta_j)/(\sum_{i=1}^k \alpha_i + 1)$, where $\delta_j = 1$ if $j = i$, 0 otherwise. This is a conditional predictive probability of a future observation. Repeat this process to obtain a sequence of exchangeable random variables $X_1, X_2, \ldots$ taking values in χ. The sample distribution of $X_1, X_2, \ldots$ converges almost surely to a random vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ which has the Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_k)$. Also, given $\boldsymbol{\theta}$, X_i's are independent with $\mathcal{P}(X_i = j) = \theta_j$ for $j = 1, \ldots, k$ and $i \geq 1$. Then a theorem of de Finetti assures that there exists a probability measure μ such that the marginal finite dimensional joint probability distributions under this measure is same for any permutation of the variables. This mixing measure is treated as a prior distribution.

Blackwell and MacQueen generalize the Polya urn scheme by taking a continuum of colors α. Since the sequence so obtained is exchangeable, they have shown that the sequence $\alpha_n(\cdot)/\alpha_n(\mathfrak{X})$, where $\alpha_n(\cdot) = \alpha(\cdot) + \sum_{i=1}^n \delta_i(\cdot)$ converges with probability one as $n \to \infty$ to a limiting discrete measure P and that P is the Dirichlet process with parameter α. It is shown later on that this method leads to characterizations of different prior processes, since once the sequence is constructed by a predictive distribution, the existence of the prior measure is assured. However the identification of that prior measure is troublesome. This approach was adopted by Mauldin et al. (1992) who use a generalized Polya urn scheme to generate sequences of exchangeable random variables and based upon them, they defined a Polya tree process. It is also used in constructing other prior processes.

In addition to the above four methods, the countable mixture representation of a random probability measure has been found to be a useful tool in developing recently several new processes. Note that Ferguson's primary definition of the Dirichlet process with parameter α was in terms of a stochastic process indexed by the elements of $\mathcal{A}$. His alternative definition was constructive and described the Dirichlet process as a random probability measure with a countable sum $\sum_{i=1}^{\infty} p_i \delta_{\xi_i}$ representation, which is a mixture of unit masses placed at random points ξ_i's, chosen independently and identically with distribution $F_0 = \alpha(\cdot)/\alpha(\mathfrak{X})$, and the random weights p_i's are such that $0 \le p_i \le 1$ and $\sum_{i=1}^{\infty} p_i = 1$. His weights were constructed using a gamma distribution. Because of the infinite sum involved in these weights it did not, with some exceptions, garner much interest in earlier applications. Sethuraman (1994) (see also Sethuraman and Tiwari 1982) remedied this problem by using beta random variables and the interest was renewed. In fact a second wave of generalization in the recent years got boost from this alternative Sethuraman representation and served as an important tool leading to a dramatic increase in the development of new priors. By varying the ingredients of this infinite sum representation, several new processes are developed, which we call *Ferguson-Sethuraman* processes. They include discrete random distributions, a beta two-parameter process, a Dirichlet Dependent process, the Chinese Restaurant and Indian buffet processes, etc.

The remarkable feature of the Dirichlet process is that it serves as a 'base' prior and is the main source for generalizations in many different directions (see Fig. 1.1). Antoniak (1974) treated the parameter α itself as random index by u, u having a certain distribution H and proposed the mixture of Dirichlet processes, i.e. $P \in \int \mathcal{D}(\alpha_u) dH(u)$. Dalal (1979a) treated the measure α as invariant under a finite group of transformations and proposed the Dirichlet Invariant process over a class of invariant distributions which included, symmetric distributions around a location ξ, or distributions having a median at 0. By writing $f(x) = \int K(x,u) dG(u)$ with a known kernel K, and taking $G \in \mathcal{D}(\alpha)$, Lo (1981) was able to place priors on the space of density functions. By taking $\alpha(\mathfrak{X})$ as a positive function instead of a constant, Walker and Muliere (1997a) were able to generalize the Dirichlet process so that the support included absolutely continuous distribution functions as well. They named it as a *beta-Stacy* prior. Teh et al. (2006) use it as a mixing distribution which leads to hierarchical models where the parameters of the prior distributions themselves are assigned priors with hyper parameters. They discuss hierarchical Dirichlet processes and indicate their extensions to other priors.

If the infinite sum $\sum_{i=1}^{\infty} p_i \delta_{\xi_i}$ is truncated at a fixed or random $N < \infty$, it generates a class of discrete distribution priors studied by Ongaro and Cattaneo (2004). In Sethuraman's representation, the weights are defined as $p_1 = V_1$ and $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$, $i = 1, 2, \ldots$, and $V_i \overset{iid}{\sim} Be(1, \alpha(\mathfrak{X}))$. By replacing $Be(1, \alpha(\mathfrak{X}))$ with $Be(a_i, b_i)$, a second group of priors are defined (see Ishwaran and James 2001). A third group of priors are developed to accommodate covariates, by indexing ξ_i with a covariate $\mathbf{x} = (x_1, \ldots, x_k)$, denoted as $\xi_{i\mathbf{x}}$ (MacEachern 1999). A further generalization is proposed by replacing the degenerate probability measure δ by a nondegenerate positive probability measure G (Dunson and Park 2008). The Sethuraman representation as well as the predictive distribution based on a generalized

Polya urn scheme proposed by Blackwell and MacQueen (1973) have been found useful in the development of new processes, some of them popularly known as the Chinese restaurant and Indian buffet processes. They have applications in nontraditional fields such as word documentation, machine learning and mixture models.

All of the above mentioned generalizations were based on the Dirichlet process. An alternative line of generalizations is based on reparametrization of F via the representation $F(t) = 1 - \exp(-Y_t)$, where Y_t is a process with independent nonnegative increments. Kalbfleisch (1978) assumed the increments to be distributed according to a gamma distribution which led to the development of the *gamma process* prior for F. Dykstra and Laud (1981) defined a weighted hazard function $r(t) = \int_{[0,t]} h(s)dZ(s)$ for any positive real valued function h, and Z, a gamma process, and thus placed priors on the space of hazard functions. By treating the increments as approximately beta random variables, Hjort (1990) was able define a *beta process* which places a prior on the space of cumulative hazard functions.

A brief exposé of these processes follows. Details are discussed in subsequent sections.

A recently published chapter by Lijoi and Prünster (2010) provides a unified framework for various priors processes in terms of the completely random measures studied by Kingman (1967). This formulation is elegant. However, we will stick with the original approach in which the priors have been constructed by suitable modifications of Lévy measures of the processes with independent nonnegative increments. The rationale being that it provides a historical perspective of the development of these processes, and perhaps easy to understand. It also reveals how these measures came about, for example in the development of the beta and beta-Stacy processes, which is not evident by the completely random measures approach.

1.1.3 Prior Processes

Ferguson's *Dirichlet process* essentially met the two basic requirements of a prior process. It is simple, defined on an arbitrary probability space and belonged to a conjugate family of priors. Lijoi and Prünster (2010) identifies conjugacy as of two types: structural and parametric. In the first one, the posterior distribution has the same structure as the prior, where as in the second case, the posterior distribution is same as the prior but the parameters are changed. Neutral to the right process is an example of the first kind and the Dirichlet process is an example of the second. While the conjugacy offers mathematical tractability, it may also be construed as limiting the family of posterior distributions.

The Dirichlet process has 'one' parameter which is interpretable. If we have a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from P and $P \in \mathcal{D}(\alpha)$, then Ferguson (1973) proved that the posterior distribution, given the sample is again a Dirichlet process with parameter $\alpha + \sum_{i=1}^{n} \delta_{x_i}$, i.e. $P|\mathbf{X} \in \mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{x_i})$. Thus it is easy to compute the posterior distribution, by simply updating the parameter of the prior process. This important property made it possible to derive nonparametric Bayesian

estimators of various functions of P, such as the distribution function, the mean, median, and a number of other quantities, by simply updating α. In fact the parameter α may be considered as representing two parameters: $F_0(\cdot) = \overline{\alpha}(\cdot) = \alpha(\cdot)/\alpha(\mathfrak{X})$ and $M = \alpha(\mathfrak{X})$. F_0 is interpreted as prior guess at random F, or prior mean, and M as prior sample size or precision parameter indicating how concentrated the F's are around F_0. Doss (1985a, 1985b) accentuates this point by constructing a prior on the space of distribution functions in the neighborhood of F_0. The posterior mean of F is shown to be a convex combination of the prior guess F_0 and the empirical distribution function. If $M \to 0$, it reduces to the classical maximum likelihood estimator (MLE) of F. On the other hand, if $M \to \infty$, it reduces to the prior guess F_0. This phenomena is shown to be true in many estimation problems.

Ferguson (1973) proved various properties and showed their applicability in solving nonparametric inference problems from a Bayesian point of view by giving several illustrative examples. His initiative set the tone and created a surge in the activity and numerous papers were published thereafter describing its utility. These applications include, sequential estimation, empirical Bayes estimation, confidence bands, hypothesis testing, and survival data analysis, to name a few. Dirichlet process is also neutral to the right process, and is essentially the only process that is tailfree with respect to every sequence of partitions. It is also the only prior process such that the distribution of $P(A)$ depends only upon the number of observations falling in the set A and not on where they fall. This may be considered as a weakness of the prior. A major deficiency is that it's support is confined to discrete probability measures only. However, several recent applications in the fields of machine learning, document classification, etc. have proved that this deficiency is after all not as serious as previously thought, and on the contrary is useful in modeling such data. It's popularity has remained unabated.

While the Dirichlet process has many desirable features and is popular, it was inadequate in treating certain problems encountered in practice, such as density estimation, bioassay, problems in reliability theory, etc. Similarly, it is inadequate in modeling hazard rates and cumulative hazard rates. Therefore several new, and in some cases an extension, are proposed in the literature as mentioned above. They are outlined next.

In dealing with the estimation of dose-response curve or estimation based on the right censored data, if the Dirichlet process prior was assumed, it was found that the posterior distribution was not a Dirichlet process, but a mixture of Dirichlet processes. This led to the development of *mixtures of Dirichlet processes* (Antoniak 1974). Roughly speaking, the parameter α of the Dirichlet process is treated as random indexed by U, U having a distribution, say, H. Thus P is said to have a mixture of Dirichlet processes (MDP) prior, if $P \in \int \mathcal{D}(\alpha_u) dH(u)$. It has some attractive properties and is flexible enough to handle purely parametric or semiparametric models. This has led to the development of mixtures models. In fact, its applications in modeling high dimensional and complex data have exploded in recent years (Dunson and Park 2008). Clearly, the Dirichlet process is a special case of MDP.

Like the Dirichlet process, MDP also has the conjugacy property. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be a sample of size n from P, $P \in \int_U D(\alpha_u) dH(u)$, then $P|\boldsymbol{\theta} \in$

$\int_U D(\alpha_u + \sum_{i=1}^n \delta_{\theta_i}) dH_{\boldsymbol{\theta}}(u)$, where $H_{\boldsymbol{\theta}}$ is the conditional distribution of u given $\boldsymbol{\theta}$. An important result proved by Antoniak is that if we have a sample from a mixture of Dirichlet processes and the sample is subjected to a random error, then the posterior distribution is still a mixture of Dirichlet processes. In applications to survival data, if the prior is assumed to be a Dirichlet process prior, then the posterior distribution given the right censored observations turns out to be a MDP. MDP is shown to be useful in treating estimation problems in bioassay. However, because of the multiplicities of observations that we expect in the posterior distribution, explicit expressions for the posterior distribution are difficult to obtain. Nevertheless, with the development of computational procedures, this limitation has practically dissipated.

The Dirichlet process is nonparametric in the sense that it has a broad support. In certain situation however Dalal (1979a) saw the need that the prior should account for some inherent structure present, such as symmetry, in the case of estimation of a location parameter, or some invariance property. This led him to define a process which is invariant, with respect to a finite group of measurable transformations $\mathcal{G} = \{g_1, \ldots, g_k\}$, $g_i : \mathfrak{X} \to \mathfrak{X}$, $i = 1, \ldots, k$, and which selects an invariant distribution function with probability one. He calls it a *Dirichlet Invariant process* with parameter α, a positive finite measure, and denotes by $\mathcal{DGI}(\alpha)$. The Dirichlet process is a special case with the group consisting of a single element, the identity transformation. The conjugacy property also holds true for the Dirichlet invariant process. That is, if $P \in \mathcal{DGI}(\alpha)$, and $X_1, \ldots, X_n$ is a sample of size n from P, then the posterior distribution of P given $X_1, \ldots, X_n$ is $\mathcal{DGI}(\alpha + \sum_{i=1}^n \delta^g_{X_i})$, where $\delta^g_{X_i} = (1/k) \sum_{i=1}^k \delta_{gX_i}$. It is found to be useful in solving some estimation problems regarding location and symmetry.

The Dirichlet process had only one parameter and it was easy to carry out the Bayesian analysis. However, Doksum (1974) saw it as a limitation and discovered that if the random P is defined on the real line R, it is possible to define a more flexible prior. He introduced a *neutral to the right process* which is based on independence of successive normalized increments of F and represents unfolding of F sequentially. That is, for any partition of the real line, $-\infty < t_1 < t_2 < \ldots < t_m < \infty$, for $m = 1, 2, \ldots$, the successive normalized increments $F(t_1), (F(t_2) - F(t_1))/(1 - F(t_1)), \ldots$ are independent. In other words, F is said to be neutral to the right, if there exists independent random variables $V_1, \ldots, V_m$ such that the distribution of the vector $(1 - F(t_1), 1 - F(t_2), \ldots, 1 - F(t_m))$ is same as the distribution of $(V_1, V_1V_2, \ldots, \prod_1^m V_i)$. Thus the prior can be described in terms of several quantities providing more flexibility. Furthermore the Dirichlet process defined on the real line is a neutral to the right process. Doksum proved the conjugacy property with respect to the data which may include right censored observations, i.e. if the prior is neutral to the right, so is the posterior. However, the expressions for the posterior distribution are complicated. Ferguson (1974) showed that it is possible to describe the posterior distribution in simple terms. The neutral to the right process is found to be especially useful in treating problems in survival data analysis but has it's own weaknesses. It's parameters are difficult to interpret and like the Dirichlet process, it also concentrates on discrete distribution functions