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 simple 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, $\varPi$ 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 $\varPi$, 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 $\varPi$ will be denoted by $\mathfrak{P}$ whenever needed.

As a second example, consider the multinomial experiment with the sample space, $\chi=\{1,2,\dots,k\}$. In this case, $\mathcal{F}(\chi)$ is the space of all distribution functions corresponding to a $(k-1)$-dimensional probability simplex $S_k=\{(p_1,p_2,\dots,p_k): 0\leq p_i\leq 1, \sum_{i=1}^{k} p_i=1\}$ 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 $\{P(B_{1j}|B_{0(1j)}) : j = 1, \ldots, k_1\}, \ldots, \{P(B_{m+1j}|B_{m(mj)}) : j = 1, \ldots, k_{m+1}\}$ 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 $\mathcal{DGI}(\alpha + \sum_{i=1}^n \delta_{X_i}^g)$, where $\delta_{X_i}^g = (1/k) \sum_{i=1}^k \delta_{gX_i}$. 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 $(V_1, V_1 V_2, \ldots, \prod_1^m V_i)$. 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 only. However, some specific neutral to the right type processes, such as beta and beta-Stacy have been since developed which soften the deficiency. These processes provide a compromise between the Dirichlet process and processes neutral to the right. They alleviate the drawbacks, and at the same time, are more manageable, parameters are interpretable and they are also conjugate with respect to the right censored data.

The neutral to the right process may also be viewed in terms of a process with independent non-negative increments (Doksum 1974; Ferguson 1974) via the reparametrization $F(t) = 1 - e^{-Y_t}$, where Y_t is a process with independent nonnegative increments. Thus a prior on $\mathcal{F}$ can be placed by using such processes. This reparametrization is key to the development of a class of neutral to the right or like neutral to the right processes to suit the needs of different applications by appropriate modification of the Lévy measure involved. They are constructed by selecting a specific independent increment process, such as, gamma, extended gamma, beta, and log-beta processes. The log-beta process leads to a beta-Stacy process prior on $\mathcal{F}$ which is a neutral to the right process. The advantage in some cases is that a posterior distribution could be described explicitly having the same structure as the prior, while in other cases only the parameters needed to be updated. This was demonstrated in Doksum (1974), Ferguson (1974) and Ferguson and Phadia (1979), and subsequently in other papers (Wild and Kalbfleisch 1981; Hjort 1990; Walker and Muliere 1997a) and was especially shown to be convenient in dealing with right censored and covariate data.

While the processes with independent increments mentioned above may be used to define priors on the space of all distribution functions, Kalbfleisch (1978), Dykstra and Laud (1981) and Hjort (1990) saw the need to define priors on the space of hazard rates and cumulative hazard rates. In view of the above reparametrization, F may also be viewed in terms of a random cumulative hazard function. In the discrete case, for an arbitrary partition of the real line, $-\infty < t_1 < t_2 < \ldots < t_m < \infty$, let q_j denote the hazard contribution of F for the interval $[t_{j-1}, t_j)$ i.e. $q_j = (F(t_j) - F(t_{j-1}))/(1 - F(t_{j-1}))$. Then the cumulative hazard function $Y(t)$ is the sum of hazard rates r_j's, $Y(t) = \sum_{t_j \le t} - \log(1 - q_j) = \sum_{t_j \le t} r_j$, and as such $Y(t)$ is identified as the *cumulative hazard rate*. Therefore, in covariate analysis of survival data, Kalbfleisch assumed r_j to be independently distributed as gamma distribution and thus was able to define a gamma process prior on the space of cumulative hazard rates, which led him to obtain the Bayes estimator for the survival function, although this was not his primary interest. In fact he was treating the baseline survival function as a nuisance parameter in dealing with covariate data under the Cox model and wanted to eliminate it.

Dykstra and Laud also notes this relationship. However, their interest being in hazard rates, they define the hazard rate in a more generalized form, $r(t) = \int_0^t \beta(s) dZ(s)$, $\beta(s) > 0$. By taking Z to be a gamma process, they place a prior on the space of all hazard rates and call it an *extended gamma process*. It can also be used to deal with a distribution function. It's parameters are interpretable. They show it to be conjugate with respect to the right censored data. But in the case of exact observations, the posterior turns out to be a mixture of extended gamma processes and the evaluation of resulting integrals become difficult.

Hjort (1990) introduced a different prior process to handle the cumulative hazard function. Like Kalbfleisch, he also defines the cumulative hazard rate as the sum of hazard rates in the discrete case (integral in the continuous case). It is clear that $Y = -\log(1 - F)$, and if F is absolutely continuous, then Y is the cumulative hazard function. To allow the case when F may not have a density, he defines a new general form of function H such that $F(t) = 1 - \prod_0^t \{1 - dH(t)\}$, where $\prod$ is the product integral. This creates a problem in defining a suitable prior on the space of all H's. Still, he attempts to model it as an independent increment process and takes the increments to be distributed approximately as beta distributions. Since the beta distribution lacks the necessary convolution properties, he had to get around by defining it in terms of 'infinitesimal' increments being beta distributed. Hjort uses this relationship to define a prior on the space of all cumulative hazard rates and consequently, on the distribution functions as it generates a proper CDF. He calls the resulting process a *beta process*. The beta process is shown to be conjugate with respect to the data, which may include right censored observations, and its posterior distribution is easy to compute by updating the parameters. It covers a broad class of models in dealing with life history data, including Markov Chain and regression models, and its parameters are accessible to meaningful interpretation.

By taking Y to be a log-beta process, Walker and Muliere (1997a) proposed a new prior process on the space of all distribution functions defined on $[0, \infty)$, and called it a *beta-Stacy process*. The process uses a generalized beta distribution and in that sense can be considered as a generalization of the beta process. Its parameters were defined in terms of the parameters of the log-beta process. By taking these parameters in more general forms they are able to construct a process whose support includes absolutely continuous distribution functions, thereby extending the Dirichlet process. It has some additional pluses as well. It's parameters have reasonable interpretation; it is a neutral to the right process; it generalizes the Dirichlet process in the sense that it offers more flexibility and unlike the Dirichlet process, it is conjugate to the right censored data. It emerges as a posterior distribution with respect to the right censored data when the prior is assumed to be a Dirichlet process. This way they are able to generalize Susarla and Van Ryzin (1978b, 1980) results in obtaining the posterior expectation of the survival function. They also introduce a generalization of Polya Urn scheme to cover the discrete time version of the process.

The tailfree and Polya tree processes are defined on the real line based on sequences of nested partitions of the real line and the property of independence of variables between partitions. Their support includes absolutely continuous distributions. They are flexible and are particularly useful when it is desired to give greater weights to the regions where it is deemed appropriate, by selecting suitable partitions. They possess the conjugacy property. However, unlike the case of the Dirichlet and other processes, the Bayesian results based on these priors are strongly influenced by the partitions chosen. Furthermore, it is difficult to derive expressions in close forms and lack adequate interpretation of the parameters involved. The Dirichlet process is essentially the only process which is tailfree with respect to every sequence of partitions.

Lavine (1992, 1994) specializes the tailfree process in which all variables involved, not just variables between partitions, are assumed to be independent each having a beta distribution. This way the expressions are manageable. He names the resulting process as a *Polya Tree process.* He shows that this process preserves the conjugacy property and for the posterior distribution, one has only to update the parameters of the beta distributions. The predictive distribution of a future observation under the Polya tree prior has a simple form and is easily computable. Under certain constraints on the parameters, the Polya tree prior reduces to the Dirichlet process. Mixtures of Polya trees are found to be useful in density estimation.

Mauldin et al. (1992) propose a different method of constructing priors, via *Polya trees*, which is slightly a generalization of Lavine's approach. Their approach is to generate sequences of exchangeable random variables based on a generalized Polya urn scheme. By a de Finetti theorem each such sequence is a mixture of iid random variables. The mixing measure is viewed as a prior on distribution functions. It is shown that this class of priors also form a conjugate family which includes the Dirichlet process and can assign probability one to continuous distributions. A thorough study of such an approach is carried out in their paper. However the approach is complicated, and from the practical point of view it is not clear if it provides any advantage over Lavine's *Polya Tree process*, and therefore will not be pursued here.

Teh et al. (2006) propose several hierarchical models where the parameters of the prior distributions themselves are assigned priors with hyper parameters. Their interest stems from a need to model group data where each group is associated with a mixture model and the desire is to link them together. In such cases, the discreteness of the Dirichlet process turns out to be an asset. They are referred to as *Hierarchical Dirichlet processes and mixture models.*

A broad and useful review of Polya tree processes with discussion may be found in Walker et al. (1999).

As mentioned earlier, a large number of priors have been proposed in the literature in recent years. They are offshoots of the Dirichlet process and are based on the constructive definition of the Dirichlet process. Ferguson's alternative definition was constructive and described the Dirichlet process as a random probability measure with a countable mixture $\sum_{i=1}^{\infty} p_i \delta_{\xi_i}$ of point masses where the weights were constructed using a gamma distribution. However, because the weights involved calculation of an infinite sum, it was impractical. On the other hand, Sethuraman's (1994) (also see, Sethuraman and Tiwari 1982) definition was also in terms of a countable mixture of point masses, but his weights involved beta random variables with a single parameter and were easy to calculate. His representation of the Dirichlet process with parameter α is

$$P(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}(\cdot), \tag{1.1}$$

where $\delta_{\xi_i}(\cdot)$ denotes a discrete measure of unit mass concentrated at ξ_i, ξ_i's are independent and identically distributed according to the distribution $\alpha(\cdot)/\alpha(\mathfrak{X})$; and

p_i (known as random weights) are chosen independently of ξ_i such that $0 \le p_i \le 1$ and $\sum_{i=1}^{\infty} p_i = 1$ a.s., $p_1 = V_1$ and $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$, $i = 1, 2, \ldots$, and $V_i \overset{iid}{\sim} Be(1, \alpha(\mathfrak{X}))$. Based on this representation, possibilities for developing several new prior processes seem natural. Ishwaran and James (2001) call them collectively as *stick-breaking* priors although they should rightly be called *Ferguson-Sethuraman* priors, and we do so, especially in view of the recent discovery of several processes based on this formulation and their newer applications in the fields such as machine learning, population genetics and ecology.

By truncating the sum to a positive integer N, a class of discrete prior processes can be generated (see Ongaro and Cattaneo 2004). Obviously, if $N = \infty$, it yields the Dirichlet process. A second generalization is achieved by assuming independent V_i's each as $V_i \sim Be(a_i, b_i)$, which results in some additional priors. If $a_i = a$ and $b_i = b$, then we have a process known as *beta two-parameter process* (Ishwaran and Zarepour 2000). On the other hand, if $a_i = 1 - a$ and $b_i = b + ia$, with $0 \le a < 1$ and $b > -1$, then it identifies the *two-parameter Poisson-Dirichlet* process described by Pitman and Yor (1997), which itself is a two parameter generalization of the *Poisson-Dirichlet* process derived by Kingman (1975) as a limiting distribution of decreasing ordered probabilities of a Dirichlet distribution. Obviously, Dirichlet process is a special case of this process when $a = 1$ and $b = \alpha(\chi)$. They all are special cases of a class of models proposed by Pitman (1996a) and called *species sampling models*,

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot) + \left(1 - \sum_{j=1}^{\infty} p_j\right) Q(\cdot), \tag{1.2}$$

where Q is the probability measure corresponding to a continuous distribution G, $\xi_j \overset{iid}{\sim} G$, and weights p_j's are constrained by the condition, $\sum_{j=1}^{\infty} p_j \le 1$.

Teh et al. (2006) discuss hierarchical Dirichlet processes and indicate their extensions to other priors. Their interest stems from a need to model group data where each group is associated with a mixture model and the desire is to link them together. In such cases, the discreteness of the Dirichlet process turns out to be an asset. To accommodate covariates, a different line of generalizations have been proposed where the locations ξ_i and/or the weights (via V_i's) are also made to depend on vectors of covariates. This class of priors include processes such as, the *dependent Dirichlet* (MacEachern 1999), *spatial Dirichlet* (Gelfand et al. 2005), *generalized spatial Dirichlet* (Duan et al. 2007), *order-based Dependent Dirichlet* (Griffin and Steel 2006), *kernel stick-breaking* (Dunson and Park 2008), *latent stick-breaking* (Rodriguez et al. 2010) and *local Dirichlet processes* (Chung and Dunson 2011). These priors are primarily used in modeling large and complex data and the analysis rely on simulations methods. For this purpose, fast running algorithms are developed. These priors are not covered in this manuscript but briefly mentioned in the Sect. 1.12. The development on this line has seen a tremendous growth and scores of papers have been published in recent years. It would be unfair to cite some at the cost of others. However, Rodriguez et al. (2010) and Chung and Dunson (2011) are a good source of references as a starting point for the reader to dwell further in detail if interested.

In addition, some interesting prior processes are developed in recent years. They include one- and two-Poisson-Dirichlet, Chinese restaurant and Indian buffet processes, etc. which may be constructed using the stick-breaking construction. Also, attempts have been made to develop bivariate prior processes. They are discussed in Sects. 1.13 to 1.16.

In summary, from the above description of various processes, it is clear that no one process is expected, nor emerged, as an ideal prior to carry out the Bayesian analysis. They all have certain advantages as well as certain limitations. They have been developed, in some cases, to address specific needs. Thus, when to use which prior process very much depends upon what is our objective and what kind of data we have on hand. Practical considerations such as incorporation of prior information, mathematical convenience, and interpretation of parameters involved and the results obtained, play a critical role in the decision. Judging from the various applications that are published in the literature and presented in the applications sections, it appears that the Dirichlet process and mixtures of Dirichlet processes have a substantial edge over the others not withstanding their limitations.

Figure 1.1 depicts the connection among the various processes.

The development of the above mentioned processes made the nonparametric Bayesian analysis feasible, but had limitations due to complexities involved in deriving explicit formulae. Therefore, the attention was focused in the past on those applications where the expressions could be obtained in closed forms and the obvious choice was the Dirichlet process. However in recent years a tremendous progress is made in developing computational methods such as, MCMC, Gibbs sampler, importance sampling, etc. for simulating the posterior distributions for implementation of Bayesian analysis, and thus made it possible to handle more complex models. And in view of the explosion that as occurred in cheap computing power, the previous limitations have almost dissipated. In fact, the mixtures of Dirichlet processes and mixtures of Polya trees have been found to be hugely popular in modeling high dimensional data encountered in practice. For example, in addition to the analysis of survival data, now it is possible to implement full Bayesian analysis in treating covariate analysis, random effect models, hierarchical procedures, wavelet models, etc. The present trend has been to combine parametric and semiparametric models in modeling such data. Books authored by Dey et al. (1998), and Ibrahim et al. (2001) contains numerous examples and applications, and is a good source if one wants to explore further from the application angle. See also Chung and Dunson (2011) and Rodriguez et al. (2010) for an extended list of references.

1.2 Dirichlet Process

The Dirichlet process is the most popular and extensively used prior in the nonparametric Bayesian analysis. In fact it is recognized that with its discovery, Ferguson's path breaking paper laid the foundation of the field of nonparametric Bayesian statistics. It's success can be traced to its mathematical tractability. Dirichlet process and

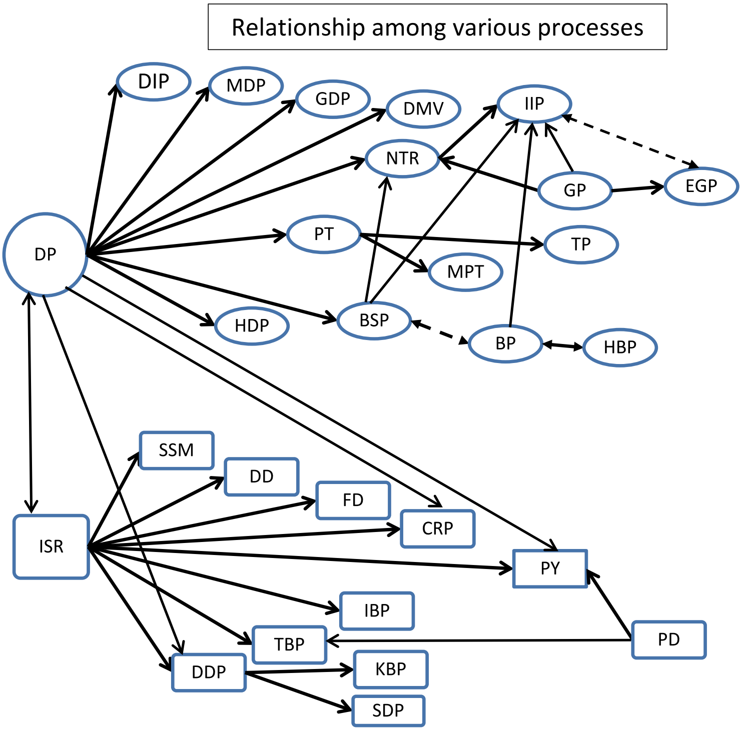

Fig. 1.1 An arrow A → B signifies either B generalizes A; or B originates from A; or A can be viewed as a particular case of B. Some relations need not be quite precise. A - -→ B suggests B can be reached from A via a transformation. Processes in rectangles are identified as Ferguson-Sethuraman processes. *Legends*: *BP*: Beta Process, *BSP*: Beta-Stacy Process, *CRP*: Chinese Restaurant Process, *DD*: Discrete Distributions, *DP*: Dirichlet Process, *DIP*: Dirichlet Invariant Process, *MDP*: Mixtures of Dirichlet Processes, *DDP*: Dependent Dirichlet Process, *DMV*: Dirichlet Multivariate Process, *EGP*: Extended Gamma Process, *FD*: Finite Dimensional Process, *GDP*: Generalized Dirichlet Process, *GP*: Gamma Process, *HBP*: Hierarchical Beta Process, *HDP*: Hierarchical Dirichlet Process, *IBP*: Indian Buffet Process, *IIP*: Independent Increments Process, *ISR*: Infinite Sum Representation, *KBP*: Kernel Based Process, *MPT*: Mixtures of Polya trees, *NTR*: Neutral to the right Process, *PD*: Poisson-Dirichlet Process, *PT*: Polya Tree Process, *PY*: Pitman-Yor (Two-parameter Poisson-Dirichlet Process), *SDP*: Spatial dependent Dirichlet Process, *SSM*: Species sampling Model, *TBP*: Two-parameter Beta Process, *TP*: Tailfree Process

its offshoots, such as mixtures of Dirichlet processes, hierarchical Dirichlet process, Dependent and Spatial Dirichlet processes, etc., are most important and widely used priors in modeling high dimensional covariate data. This is made possible due to the development of computational techniques that make full Bayesian analysis feasible.

A Dirichlet process prior with parameter α for a random distribution function F is a probability measure on the space of all distribution functions F and is governed by two parameters: a baseline distribution function F_0 that defines the "location" or "center" of the prior, and a positive scalar precision parameter M which governs how concentrated the prior is around the prior 'guess' or baseline distribution

F_0. The latter therefore measures the strength of belief in the prior guess. For large values of M, a sampled F is likely to be closed to F_0. For small values of M, it is likely to put most of its mass on just a few atoms. Ferguson defines the Dirichlet process more broadly in terms of a random probability measure. However, the Dirichlet process is concentrated on discrete probability distributions.

1.2.1 Definition

Let P be a probability measure defined on a measurable space $(\mathfrak{X}, \mathcal{A})$, where $\mathfrak{X}$ is a separable metric space and $\mathcal{A} = \sigma(\mathfrak{X})$ is the corresponding σ-field of subsets of $\mathfrak{X}$, and Π be a set of all probability measures on $(\mathfrak{X}, \mathcal{A})$. In our context P is considered to be a parameter and $(\Pi, \sigma(\Pi))$ serves as the parameter space. Thus P may be viewed as a stochastic process indexed by sets $A \in \mathcal{A}$ and is a mapping from Π into $[0, 1]$. F is a cumulative distribution function corresponding to P and let $\mathcal{F}$ denote the space of all distribution functions.

In his fundamental paper, Ferguson (1973) developed a prior process which he called the *Dirichlet process* (Blackwell 1973 and Blackwell and MacQueen 1973 named it as *Ferguson prior*). It is especially convenient since it satisfies the two desirable properties mentioned in the earlier section. Because of its simplicity and analytical tractability, the Dirichlet Process has been widely used despite its main problem that it gives positive probability to discrete distributions only. We now give a formal definition. His paper is the main source of the definition and some of the properties described here.

Let $D(\gamma_1, \ldots, \gamma_k)$ denote a $(k-1)$-dimensional Dirichlet distribution with density function given by

$$f(x_1, \ldots, x_{k-1}) = \frac{\Gamma(\gamma_1 + \ldots + \gamma_k)}{\Gamma(\gamma_1) \ldots \Gamma(\gamma_k)} \prod_{i=1}^{k-1} x_i^{\gamma_i - 1} \left(1 - \sum_{i=1}^{k-1} x_i\right)^{\gamma_k - 1}, \tag{1.3}$$

over the simplex: $S : \{(x_1, \ldots, x_{k-1}) : x_i \geq 0, i = 1, \ldots, k-1, \sum_{i=1}^{k-1} x_i \leq 1\}$ where all γ_i are positive real numbers. The Dirichlet distribution has many interesting properties which lead to the corresponding properties of the Dirichlet process. For example, its representation in terms of gamma random variables leads to the alternative definition of the Dirichlet process. It's tailfree property shows that the Dirichlet process is a tailfree process. Also, it's neutral to the right property shows that the Dirichlet process is a neutral to the right process. These and many other properties are discussed in great details by Basu and Tiwari (1982).

We say P is a *random probability measure* on $(\mathfrak{X}, \mathcal{A})$ (i.e. a measurable map from some probability space $(\Omega, \mathcal{F}, Q)$ into $(\Pi, \sigma(\Pi))$) if for any $A \in \mathcal{A}$, $P(A)$ is random taking values in $[0, 1]$, $P(\mathfrak{X}) = 1$ a.s. and P is finitely additive in distribution. Essentially it is a stochastic process index by set $A \in \mathcal{A}$.

The Dirichlet process with parameter α, to be denoted as $\mathcal{D}(\alpha)$, is defined as follows.

Definition 1.1 (Ferguson) Let α be a non-null non-negative finite measure on $(\mathfrak{X}, \mathcal{A})$. A random probability P is said to be a Dirichlet process on $(\mathfrak{X}, \mathcal{A})$ with parameter α if for any positive integer k, and measurable partition $(A_1, \ldots, A_k)$ of $\mathfrak{X}$, the distribution of the vector $(P(A_1), \ldots, P(A_k))$ is Dirichlet distribution, $D(\alpha(A_1), \ldots, \alpha(A_k))$.

By verifying the Kolmogorov consistency criterion, he showed the existence of a probability measure $\boldsymbol{P}$ on the space of all functions from $\mathcal{A}$ into [0, 1] with σ-field generated by the cylinder sets, such that the finite dimensional joint distribution of probabilities of sets $A_1, \ldots, A_k$ is Dirichlet distribution. $\mathcal{D}(\alpha)$ may thus be considered as a prior distribution on the space Π.

The parameter α, can in fact be represented by two values. The total mass $M = \alpha(\mathfrak{X})$, and the normalized function $\overline{\alpha}(\cdot) = \alpha(\cdot)/\alpha(\mathfrak{X})$ which may be identified with F_0, the prior guess at F mentioned earlier in the section. This fact is used in defining later the *generalized Dirichlet* process, where M is replaced by a positive function. For the sake of brevity, we will write $\alpha(a, b)$ for $\alpha((a, b))$, the α measure of the set (a, b).

Several equivalent definitions of the Dirichlet process have been proposed in the literature which are described next.

The above definition was given in terms of a stochastic process indexed by the elements $A \in \mathcal{A}$. Ferguson also gave an alternative definition in terms of a countable mixture $\sum_{j=1}^{\infty} P_j \delta_{\xi_j}$ of point masses at random points with mixing weights derived from a gamma process. In doing so, he was motivated by the fact that as the Dirichlet distribution is definable by taking the joint distribution of gamma variables divided by their sum, so should be the Dirichlet process be definable as a gamma process with increments divided by their sum. Let $\mathcal{P}$ denote the probability of an event. Let $J_1, J_2, \ldots$ be a sequence of random variables with the distribution, $\mathcal{P}(J_1 \leq x_1) = \exp\{N(x_1)\}$ for $x_1 > 0$, and for $j = 2, 3, \ldots$, $\mathcal{P}(J_j \leq x_j | J_{j-1} = x_{j-1}, \ldots, J_1 = x_1) = \exp\{N(x_j) - N(x_{j-1})\}$ for $0 < x_j < x_{j-1}$, where $N(x) = -\alpha(\mathfrak{X}) \int_x^{\infty} u^{-1} e^{-u} du$ for $0 < x < \infty$. Then the sum $\sum_{j=1}^{\infty} J_j$ converges with probability one and is a gamma variate with parameters $\alpha(\mathfrak{X})$ and 1, $G(\alpha(\mathfrak{X}), 1)$. Define $P_j = J_j / \sum_{j=1}^{\infty} J_j$, then $P_j \geq 0$ and $\sum_{j=1}^{\infty} P_j = 1$ with probability one. Let ξ_j's be iid $\mathfrak{X}$-valued random variables with common distribution $\overline{\alpha}(\cdot)$ and independent of $P_1, P_2, \ldots$. Then the random probability measure defined by $P(A) = \sum_{j=1}^{\infty} P_j \delta_{\xi_j}(A)$, $A \in \mathcal{A}$ is a Dirichlet process on $(\mathfrak{X}, \mathcal{A})$.

Ishwaran and James (2001) describe an alternate formulation of the weights P_j's. Let $\gamma_k = \varepsilon_1 + \ldots + \varepsilon_{k,}$ where each ε_i is distributed as exponential distribution with parameter 1, that is $\varepsilon_i \overset{iid}{\sim} \exp(1)$. Let N^{-1} stand for the inverse of the Lévy measure of a gamma distribution with parameter $\alpha(\mathfrak{X})$, where $N(x) = \alpha(\mathfrak{X}) \int_x^{\infty} u^{-1} e^{-u} du$, for $x > 0$. Then

$$P(\cdot) = \sum_{j=1}^{\infty} \frac{N^{-1}(\gamma_j)}{\sum_{j=1}^{\infty} N^{-1}(\gamma_j)} \delta_{\xi_j}(\cdot). \tag{1.4}$$

This form relies on the random weights constructed using infinitely divisible random variables. The weights are $N^{-1}(\gamma_j)$ normalized by their sum $\sum_{j=1}^{\infty} N^{-1}(\gamma_j)$, a gamma $(\alpha(\mathfrak{X}))$ variable. They are in descending order and the distribution of these weights is the Poisson-Dirichlet developed by Kingman (1975).

Another constructive definition is given by Sethuraman and Tiwari (1982) and Sethuraman (1994) in which the weights are derived using a beta distribution with parameters 1 and $\alpha(\mathfrak{X})$, $Be(1, \alpha(\mathfrak{X}))$. Their representation of a random probability measure P having a Dirichlet prior $\mathcal{D}(\alpha)$ is

$$P(A) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(A), \quad A \in \mathcal{A}, \tag{1.5}$$

where ξ_j's are as above, $p_1 = V_1$, and for $j \geq 2$, $p_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$ with $V_j \overset{iid}{\sim} Be(1, \alpha(\mathfrak{X}))$ and independent of $\xi_1, \xi_2, \ldots$. That is the weights are generated by the so called *stick-breaking* construction. Break off a part of a stick of unit length randomly according to $Be(1, \alpha(\mathfrak{X}))$ and assign the length of the break part V_1 to p_1. Next, from the remaining part $1 - V_1$ of the stick, again break off randomly a V_2 fraction according to $Be(1, \alpha(\mathfrak{X}))$ and set $p_2 = V_2(1 - V_1)$. Continue this process. At the j-th step break off a V_j fraction according to $Be(1, \alpha(\mathfrak{X}))$ of the remaining length of stick $\prod_{l=1}^{j-1}(1 - V_l)$ and set $p_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$. This results in a sequence of weights $p_{1,} p_2, \ldots$ of Sethuraman representation. In contrast to Ferguson's weights, these weights need not be ordered. Ferguson's weights are equivalent to these weights rearranged in decreasing order. That is, Ferguson weights are order statistics of these weights, $p_{(1)}, p_{(2)}, \ldots$. As will be seen later, the stick-breaking construction has been found useful in representing other priors as well.

This representation reveals that a random probability P is a countable mixture of point masses at random locations $\xi_1, \xi_2, \ldots$, chosen according to the distribution $\overline{\alpha}(\cdot)$. The mixing probabilities $p_1, p_2, \ldots$ may be viewed as a discrete distribution defined on the set of nonnegative integers with $V_1, V_2, \ldots$ representing independent discrete failure rates of this discrete distribution and each distributed as $Be(1, \alpha(\mathfrak{X}))$. It immediately shows that the Dirichlet process chooses a discrete distribution with probability one, a fact proved by various authors. This representation is more amenable to proving various useful properties of the Dirichlet process and, as will be seen later, it forms the basis for generating several generalization of the Dirichlet process. Sethuraman (1994) also established an interesting distributional equality

$$P(A) = p_1 \delta_{\xi_1}(A) + \sum_{j=2}^{\infty} p_j \delta_{\xi_j}(A), \quad A \in \mathcal{A}. \tag{1.6}$$

The stick-breaking construction is related to the Poisson process (Ishwaran and Zarepour 2002) as follows.

$$P(\cdot) \overset{D}{=} \sum_{j=1}^{\infty} \left(e^{-\gamma_{j-1}/\alpha} - e^{-\gamma_j/\alpha}\right) \delta_{\xi_j}(\cdot), \tag{1.7}$$

with $\gamma_0 = 0$, γ's as defined above. This can be seen by observing that $p_j = e^{-\gamma_{j-1}/\alpha} - e^{-\gamma_j/\alpha} = e^{-\varepsilon_1/\alpha} \ldots e^{-\varepsilon_{j-1}/\alpha}(1 - e^{-\varepsilon_j/\alpha}) \overset{D}{=} \prod_{l=1}^{j-1}(1 - V_l)V_j$, since $e^{-\varepsilon_1/\alpha} \sim Be(\alpha, 1)$ and $1 - e^{-\varepsilon_1/\alpha} \sim Be(1, \alpha)$. Their paper (Ishwaran and Zarepour 2002) gives an interesting discussion of different representations of the weights.

Blackwell and MacQueen (1973) also provide an alternative definition. Let α be a measure as before. Define a sequence $\{X_n : n \geq 1\}$ of random variables taking values in $\mathfrak{X}$ as follows. For every $A \in \mathcal{A}$, let $\mathcal{P}(X_1 \in A) = \alpha(A)/\alpha(\mathfrak{X})$ and

$$\mathcal{P}(X_{n+1} \in A | X_1, \ldots, X_n) = \frac{\alpha_n(A)}{\alpha_n(\mathfrak{X})} = \frac{\alpha(A) + \sum_{i=1}^{n} \delta_{x_i}(A)}{\alpha(\mathfrak{X}) + n}, \tag{1.8}$$

where $\alpha_n = \alpha + \sum_{i=1}^{n} \delta_{x_i}$. This sequence, called a *Polya sequence with parameter* α, may be viewed as the results of successive draws of balls from a Polya urn containing $\alpha(x)$ balls of color $x \in \mathfrak{X}$ in which at every stage a ball is drawn at random, it's color is noted and is replaced by two balls of the same color. Then Blackwell and MacQueen have shown that the sequence $\alpha_n(\cdot)/\alpha_n(\mathfrak{X})$ converges with probability one as $n \to \infty$ to a limiting discrete measure P and that P is the Dirichlet process with parameter α. The predictive distribution in (1.8) can also be expressed as

$$\mathcal{P}(X_{n+1} \in \cdot | X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \frac{1}{\alpha(\mathfrak{X}) + n} \delta_{x_i}(\cdot) + \frac{\alpha(\mathfrak{X})}{\alpha(\mathfrak{X}) + n} \overline{\alpha}(\cdot). \tag{1.9}$$

The sequence $\{X_n : n \geq 1\}$ so constructed is exchangeable and therefore, applying a theorem of de Finetti it can be shown (Ghosh and Ramamoorthi 2003) that the mixing distribution turns out to be the Dirichlet process. Basu and Tiwari (1982) have studied Blackwell-MacQueen definition in detail. Their paper clears up the measure theoretical details. Also see an interesting commentary by Sethuraman ("Commentary on a note on the Dirichlet process" in *Selected Works of Debabrata Basu*, Ed.: Anirban DasGupta, Springer, 2011).

A sample from P is defined as follows.

Definition 1.2 (Ferguson) Let P be a random probability measure on $(\mathfrak{X}, \mathcal{A})$. $X_1, \ldots, X_n$ is said to be a sample from P, if for any positive integer m and measurable sets $A_1, \ldots, A_m, C_1, \ldots, C_n$ of $\mathfrak{X}$,

$$\begin{aligned} &P\left\{X_1 \in C_1, \ldots, X_n \in C_n | P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_n)\right\} \\ &\quad = \prod_{j=1}^{n} P(C_j) \quad \text{a.s.} \end{aligned} \tag{1.10}$$

1.2.2 Properties

Ferguson proved several properties of the Dirichlet process and gave a few applications. Following this, various authors have proved additional properties. For proofs, the reader is referred to the original source. In all these properties, P is assumed to be a Dirichlet process on $(\mathfrak{X}, \mathcal{A})$, denoted as $P \in \mathcal{D}(\alpha)$ and given P, let $X_1, \ldots, X_n$ be a sample of size n from P. Also let $M = \alpha(\mathfrak{X})$.

1. Let Z be measurable real valued functions defined on $(\mathfrak{X}, \mathcal{A})$. If $\int |Z| d\alpha < \infty$, then $\int |Z| dP < \infty$ with probability one and $\mathcal{E} \int Z dP = \int Z d\mathcal{E}(P) = \int Z d\overline{\alpha}$.
2. Let Z_1 and Z_2 be two measurable real valued functions defined on $(\mathfrak{X}, \mathcal{A})$. If $\int |Z_1| d\alpha < \infty$, $\int |Z_2| d\alpha < \infty$ and $\int |Z_1 Z_2| d\alpha < \infty$, then $\mathcal{E} \int Z_1 dP \int Z_2 dP = \sigma_{12}/(M+1) + \mu_1\mu_2$, where $\mu_i = \int Z_i d\overline{\alpha}$, $i = 1, 2$ and $\sigma_{12} = \int Z_1 Z_2 d\overline{\alpha} - \mu_1\mu_2$. If further $\int |Z_1|^2 d\alpha < \infty$ and $\int |Z_2|^2 d\alpha < \infty$, then $\mathrm{Cov}(\int Z_1 dP, \int Z_2 dP) = \sigma_{12}/(M+1)$, and from this we get $\mathrm{Var}(\int Z_1 dP) = \sigma_0^2/(M+1)$, where $\sigma_0^2 = \int Z_1^2 d\overline{\alpha} - \mu_1^2$.
3. Let $\mu = \int Z dP$ and $\mu_0 = \mathcal{E}(\mu) = \int Z d\overline{\alpha}$. If $\int Z^4 d\alpha < \infty$, then $\mathcal{E}(\mu - \mu_0)^3 = 2\mu_3/[(M+1)(M+2)]$.

   Also
   $$\mathcal{E}(\mu - \mu_0)^4 = \frac{6\mu_4 + 3M\sigma^4}{(M+1)(M+2)(M+3)}, \tag{1.11}$$
   where $\sigma^2 = \int (Z - \mu_0)^2 d\overline{\alpha}$, $\mu_3 = \int (Z - \mu_0)^3 d\overline{\alpha}$, and $\mu_4 = \int (Z - \mu_0)^4 d\overline{\alpha}$.
4. Let X be a sample of size one from P. Then $\mathcal{P}(X \in A) = \alpha(A)/\alpha(\mathfrak{X})$ for $A \in \mathcal{A}$.
5. Let $X \sim P$. Then $\alpha(\mathfrak{X}) = \mathcal{E}[\mathrm{Var}(X|P)]/\mathrm{Var}[\mathcal{E}(X|P)]$. This property provides an interpretation of Bayes rule for $\mu = \int x dP$ as a Gauss-Markov estimator of $(1/n)\alpha(\mathfrak{X})$ representing the relative precision of the prior mean $\mathcal{E}(\mu)$ to the sample mean $\overline{X}$, and provides additional support for the interpretation of $\alpha(\mathfrak{X})$ as the prior sample size.
6. If α is σ-additive, then so is P the Dirichlet process.

Two probability measures P_1 and P_2 are said to be mutually singular, in symbol $P_1 \perp P_2$, if there exists an event A such that $P_1(A) = 1$ and $P_2(A) = 0$. Similarly, by saying that two Dirichlet processes $\mathcal{D}(\alpha_1)$ and $\mathcal{D}(\alpha_2)$ are mutually singular it is meant that given one sample process P from either $\mathcal{D}(\alpha_1)$ or $\mathcal{D}(\alpha_2)$, it is possible to identify with probability 1 to which distribution it belongs.

7. Let α_1 and α_2 be two non-atomic, non-null finite measures on $(\mathfrak{X}, \mathcal{A})$. If $\alpha_1 \neq \alpha_2$, then $\mathcal{D}(\alpha_1) \perp \mathcal{D}(\alpha_2)$.
8. The Dirichlet process chooses a discrete random probability measure with probability one. This is true even when α is assumed to be continuous. Therefore, it cannot serve as prior in density estimation (however kernel mixture of Dirichlet processes can be used. See Sect. 2.5.4)

Let $g(x_1, \dots, x_k)$ be a measurable real-valued function defined on the k-fold product space $(\mathfrak{X}^k, \mathcal{A}^k)$ of $(\mathfrak{X}, \mathcal{A})$ and symmetric in $x_1, \dots, x_k$. Assume that

$$\int_{\mathfrak{X}^m} \left| g(x_1, \dots, x_1, x_2, \dots, x_2, \dots, x_m, \dots, x_m) \right| d\overline{\alpha}(x_1) \dots d\overline{\alpha}(x_m) < \infty, \tag{1.12}$$

for all possible combinations of arguments $(x_1, \dots, x_1, x_2, \dots, x_2, \dots, x_m, \dots, x_m)$ from all of x_i's distinct ($m = k$) to all identical ($m = 1$). Note that the function g vanishes whenever any two coordinates are equal, and condition (1.12) reduces to the simple condition

$$\int_{\mathfrak{X}^k} \left| g(x_1, \dots, x_k) \right| d\overline{\alpha}(x_1) \dots d\overline{\alpha}(x_k) < \infty. \tag{1.13}$$

An important property that has been used widely in solving some non-parametric Bayesian estimation problems is stated next. Yamato (1977a, 1977b, 1984) and Tiwari (1981, 1988) have derived the following:

9. Under the assumption (1.12),

$$\int_{\mathfrak{X}^k} \left| g(x_1, \dots, x_k) \right| dP(x_1) \dots dP(x_k) < \infty \quad \text{with probability one,} \tag{1.14}$$

and

$$\mathcal{E} \int_{\mathfrak{X}^k} g(x_1, \dots, x_k) dP(x_1) \dots dP(x_k)$$

$$= \sum_{C(\sum i m_i = k)} \frac{k! [\alpha(\mathfrak{X})]^{\sum m_i}}{\prod_{i=1}^k [i^{m_i} (m_i)!] \alpha(\mathfrak{X})^{(k)}} \int_{\mathfrak{X}^{\sum m_i}} \Psi(\mathbf{x}) \prod_{i=1}^{k} \prod_{j=1}^{m_i} d\overline{\alpha}(x_{ij}), \tag{1.15}$$

where $\Psi(\mathbf{x}) = g(x_{11}, \dots, x_{1m_1}, x_{21}, \dots, x_{2m_2}, \dots, x_{km_k}, \dots, x_{km_k})$ and the summation $\sum_{C(\sum i m_i = k)}$ extends over all non-negative integers $m_1, \dots, m_k$ such that $\sum i m_i = k$. Taking g to be the indicator function, one can derive the marginal distribution of a sample $(x_1, \dots, x_k)$ from P. Yamato (1977a, 1977b) gives examples of various estimable functions as special cases of g. For example letting $g(x_1, \dots, x_k) = x_1 x_2 \dots x_k$, he derived the k-th moment of the Dirichlet process $\mathcal{D}(\alpha)$.

Again, using the alternative definition of the Dirichlet process given by Ferguson (1973) mentioned before, namely, $P(A) = \sum_{j=1}^{\infty} P_j \delta_{\xi_j}(A)$, $A \in \mathcal{A}$, Yamato (1977a, 1977b, 1984) proved the following result:

10. For any positive integers $m, r_1, \dots, r_m$,

$$\mathcal{E} \Big(\sum_{j_1 \neq j_2 \neq \dots \neq j_m} \sum \dots \sum P_{j_1}^{r_1} \dots P_{j_m}^{r_m} \Big) = \frac{(r_1 - 1)! \dots (r_m - 1)! \cdot M^m}{M^{(k)}}, \tag{1.16}$$

where $M = \alpha(\mathfrak{X})$, $k = \sum_{i=1}^{m} r_i$. As special cases we have $\mathcal{E}(\sum_{j=1}^{\infty} P_j^2) = 1/(M+1)$, $\mathcal{E}(\sum_{i \neq j} P_i P_j) = M/(M+1)$, $\mathcal{E}(\sum_{j=1}^{\infty} P_j^3) = 2/[(M+1)(M+2)]$, $\mathcal{E}(\sum_{i \neq j} P_i^2 P_j) = M/[(M+1)(M+2)]$, $\mathcal{E}(\sum_{i \neq j \neq k} P_i P_j P_k) = M^2/[(M+1)(M+2)]$, $\mathcal{E}(\sum_{j=1}^{\infty} P_j^2)^2 = M(M+6)/M^{(4)}$, $\mathcal{E}(\sum_{j=1}^{\infty} P_j^2)^3 = M(M^2 + 18M + 120)/M^{(6)}$, etc.

11. The distribution of $\mu(P)$, the mean of the process, can also be obtained from Hannum et al. (1981) who have shown that $\mathcal{P}\{\int g(t)dP(t) \leq x\} = \mathcal{P}\{T^x \leq 0\}$, where $-\infty < x < \infty$ and T^x is a random variable with characteristic function $\exp\{-\int_R \log[1 - it\{g(t) - x\}]d\alpha(t)\}$. By the use of this result they have shown that when g is odd and α is symmetric about 0, then the distribution of $\int g(t)dP(t)$ is symmetric about 0; and that if P and P_n are random probability measures on $(R, \mathcal{B})$ with priors $\mathcal{D}(\alpha)$ and $\mathcal{D}(\alpha_n)$, $n = 1, 2, \ldots$, and if $\alpha_n \overset{w}{\to} \alpha$ as $n \to \infty$, then under some mild regularity conditions $\int g dP_n$ converges in distribution to $\int g dP$.
12. Let $P \in \mathcal{D}(\alpha)$ and let $A \in \mathcal{A}$. Then Antoniak (1974) has shown that given $P(A) = c$, the conditional distribution of $(1/c)P$ restricted to $(A, \mathcal{A} \cap A)$ is a Dirichlet process on $(A, \mathcal{B} \cap A)$ with parameter α restricted to A. That is, for any measurable partition $(A_1, \ldots, A_k)$ of A, the distribution of the vector $(P(A_1)/c, \ldots, P(A_k)/c)$ is Dirichlet, $D(\alpha(A_1), \ldots, \alpha(A_k))$.
13. Let $\{\pi_m; m = 1, 2, \ldots\}$ be a nested tree of measurable partitions of $(R, \mathcal{B})$; that is $\pi_1, \pi_2, \ldots$ be a nested sequence of measurable partitions such that π_{m+1} is a refinement of π_m for each m and $\bigcup_0^{\infty} \pi_m$ generates $\mathcal{B}$. Then the Dirichlet process is tailfree with respect to every tree of partitions.
14. Dirichlet process is neutral to the right with respect to every sequence of nested, measurable ordered partitions.
15. Let $P \in \mathcal{D}(\alpha)$ and given $P = P$, let X be a random sample of size one from P. Let $A \in \mathcal{A}$. Then the conditional distribution of P given $P(A)$ and $X \in A$, is same as the conditional distribution of P given $P(A)$ (Antoniak 1974).

In view of property 8, not all n observations of a sample drawn from P may be distinct. Therefore, let $k(n)$ be the number of distinct observations denoted by $X_1^*, \ldots, X_{k(n)}^*$. Assuming α to be non-atomic and unknown, Korwar and Hollander (1973) established a very interesting result:

16. (i) $k(n)/\log n \to \alpha(\mathfrak{X})$ a.s. and (ii) $X_1^*, \ldots, X_{k(n)}^*$ are independent and identically distributed as $\alpha(\cdot)/\alpha(\mathfrak{X})$. Thus $\alpha(\mathfrak{X})$ can be estimated consistently by the quantity $k(n)/\log n$ and the second result leads to a strong law of large numbers for the mean $\sum_{i=1}^{k(n)} X_i^*/k(n)$. This supports Ferguson's remark that $\alpha(\mathfrak{X})$ may be interpreted as the sample size.

Since the observations repeat themselves with probability one, let m_i stand for the number of observations in the sample $\mathbf{X} = (X_1, \ldots, X_n)$ which repeat exactly i times, $i = 1, 2, \ldots, n$. Then $\sum_{i=1}^{n} i m_i = n$. Let $(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n)$ be the event that in the sample exactly m_1 observations occur only once and they are the first $m_1 X$'s, exactly m_2 observations occur in pairs and they are the next $2m_2 X$'s,

$\dots$, exactly m_n observations occur each n times and they are the last nm_nX's. For example, $m_1 = n$ and $m_i = 0$ for $i > 1$ means all observations are distinct. If $m_1 = \dots = m_{n-1} = 0$, $m_n = 1$ means all observations are identical. Then Antoniak (1974) proved the following result.

17. Let α be non-atomic. Then

$$\mathcal{P}\big((X_1, \dots, X_n) \in C(m_1, \dots, m_n)\big)$$
$$= \frac{n!}{\prod_{i=1}^{n} i^{m_i}(m_i!)} \frac{\alpha(\mathfrak{X})^{\sum_{i=1}^{n} m_i}}{\alpha(\mathfrak{X})^{(n)}}, \tag{1.17}$$

where $s^{(n)} = s(s+1)\dots(s+n-1)$, and $\mathcal{P}$ denotes the probability of an event.

This formula is discovered independently by Ewen (1972). See (1.121).

It is easier to compute the moments of p_j's using the Sethuraman (1994) representation.

18. Let m_i's be as before. Then, for any combination $(m_1, \dots, m_n)$ of $\{1, 2, \dots, n\}$ Tiwari (1981, 1988) proved

$$\mathcal{E}\left(\sum^{*} p_{11} \dots p_{1m_1} p_{21}^2 \dots p_{2m_2}^2 \dots p_{n1}^n \dots p_{nm_n}^n\right)$$
$$= \frac{\prod_{i=1}^{n} [(i-1)!\alpha(\mathfrak{X})]^{m_i}}{\alpha(\mathfrak{X})^{(n)}}, \tag{1.18}$$

where the summation $\sum^*$ is over each p_{ij} $(j = 1, \dots, m_i, i = 1, \dots, n)$ taking all mutually distinct values of $p_1, p_2, \dots$.

Yamato (1977a, 1977b) also derived similar results using Ferguson's (1973) alternative definition. They both use Antoniak's result of the previous Property. For $n = 2$, the above result was established in Ferguson (1973).

19. The marginal (predictive) distribution of $X_{n+1}|\alpha, X_1, \dots, X_n$ is a rescaled version of the updated $\overline{\alpha}$ measure namely $\overline{\alpha}_n = (\alpha + 1)/(\alpha(\mathfrak{X}) + n)$. It can be written as a mixture with $X_{n+1} \sim \overline{\alpha}$ with probability $M/(M+n)$ and X_{n+1} set equal to X_i, $i = 1, \dots, n$, each with probability $1/(M+n)$ and $M = \alpha(\mathfrak{X})$. Since $X_1, \dots, X_n|P$ are iid, they are exchangeable. Therefore one can rewrite the conditional distribution of $X_i|\alpha, X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ as the mixture with $X_i \sim \overline{\alpha}$ with probability $M/(M+n-1)$ and $X_i = X_j$, $j = 1, \dots, i-1, i+1, \dots, n$, each with probability $1/(M+n-1)$. Because of the discreteness of the Dirichlet process, we expect some observations to be repeated. The predictive distribution of a future observation may thus be written as

$$X_{n+1}|X_1, X_2, \dots, X_n \sim \frac{M}{M+n}\overline{\alpha} + \frac{n}{M+n} \cdot \frac{1}{n} \sum_{j=1}^{K} n_j \delta_{X_j^*}, \tag{1.19}$$

where X_j^* are $K \leq n$ distinct observations among n with frequency n_j. This implies that X_{n+1} will be a new observation with probability $M/(M+n)$ and will coincide with X_j^* with probability $n_j/(M+n)$, $j = 1, \ldots, K$. As $n \to \infty$, the probability increases that it will coincide with one of the previous observations. As noted by Lijoi and Prünster (2010) that since these probabilities do not depend upon K, the number of distinct observations or on frequency n_j, important information is un-utilized. This is remedied by the generalization of the Dirichlet process, namely, the two-parameter Poisson-Dirichlet distribution (Pitman and Yor 1997. See Sect. 1.13.2).

Sethuraman and Tiwari (1982) have investigated the limits of prior distributions as the parameter α tends to various values. Using their representation (1.1), they have proved the following results.

20. Let $\{\alpha_r\}$ be a sequence of non-null σ-additive finite measures on $(\mathfrak{X}, \mathcal{A})$ such that

$$\alpha_r(\mathfrak{X}) \to 0 \quad \text{and} \quad \operatorname*{Sup}_A \left|\overline{\alpha}_r(A) - \overline{\alpha}_0(A)\right| \to 0 \quad \text{as } r \to \infty, A \in \mathcal{A} \tag{1.20}$$

where $\overline{\alpha}_r$ is a probability measure in Π. Then $\mathcal{D}(\alpha_r) \xrightarrow{w} \delta_{Y_0}$ and $\mathcal{D}(\alpha_r + \sum_{i=1}^n \delta_{x_i}) \xrightarrow{w} \mathcal{D}(\sum_{i=1}^n \delta_{x_i})$ as $r \to \infty$, where Y_0 has the distribution $\overline{\alpha}_0$. This means that if the total mass $\alpha(\mathfrak{X})$ converges to zero, the Dirichlet process reduces to a degenerate mass at a point Y_0 selected according to $\overline{\alpha}_0$ distribution.

For various other convergence properties of the Dirichlet process, the reader is referred to Ghosh and Ramamoorthi (2003). James (2006), and references cited in his paper, study the distribution of a linear functional of the Dirichlet process. Cifarelli and Regazzini (1979a, 1979b) and Hannum et al. (1981) report in their papers methods of finding the distribution of the mean of a Dirichlet process.

Characterization Let $\mathcal{C}$ denote the class of all random probability measures P such that (1) P is degenerate at a given probability distribution P_0; (2) P concentrates at a random point; or (3) P concentrates at two random points. Doksum (1974) proved the following characterizations.

1. If $P \notin \mathcal{C}$ is tailfree with respect to all sequences of nested, measurable partitions, then P is a Dirichlet process.
2. If $P \notin \mathcal{C}$ is neutral to the right with respect to all sequences of nested, measurable, ordered partitions, then P is a Dirichlet process.
3. The Dirichlet process is the only process not in $\mathcal{C}$ such that for each $A \in \mathcal{A}$, the posterior distribution of $P(A)$ given a sample $X_1, \ldots, X_n$ from P, depends only on the number of X's that fall in A and not where they fall. That is, if the posterior distribution of the random probability P given a sample $X_1, \ldots, X_n$ from P depends on the sample only through the empirical distribution function, then P is a Dirichlet random probability.

4. Lo (1991) gives a different characterization of the Dirichlet process. If the posterior mean of the random probability P, given a sample $X_1, \ldots, X_n$ from P, is linear in the empirical distribution function, then P is a Dirichlet random probability. That is, P is a Dirichlet random probability on $(\mathfrak{X}, \mathcal{A})$ if and only if for each $n = 1, 2, \ldots$, the posterior mean of P given a sample $X_1, \ldots, X_n$ is given by $(1 - a_n)P_n + a_n(1/n)\sum_{i=1}^{n} \delta_{x_i}$ for some $a_n \in (0, 1)$ and some probability P_n on $(\mathfrak{X}, \mathcal{A})$.

This characterization can also be expressed in terms of the predictive probability based on a sequence of exchangeable random variables $X_1, \ldots, X_n$.

$$\begin{aligned} \mathcal{P}(X_{n+1} \in A | X_1, \ldots, X_n) &= \frac{\alpha(A) + \sum_{i=1}^{n} \delta_{x_i}(A)}{\alpha(\mathfrak{X}) + n} \\ &= p_n \frac{\alpha(A)}{\alpha(\mathfrak{X})} + (1 - p_n) \frac{\sum_{i=1}^{n} \delta_{x_i}(A)}{n}, \end{aligned} \tag{1.21}$$

which is a linear combination of the prior measure α and the empirical distribution with $p_n = \alpha(\mathfrak{X})/(\alpha(\mathfrak{X}) + n)$.

This predictive distribution can be interpreted as X_{n+1} will be a new observation with probability p_n and will coincide with one of the previous observations $X_1, \ldots, X_n$ with probability $1 - p_n$. Note that not all $X_1, \ldots, X_n$ need be distinct.

Posterior Distribution All of the above properties are derived for functions of random probability P having a Dirichlet process prior. Given P, if we have a random sample from P, the following important property was proved in Ferguson (1973). This property is fundamental in solving non-parametric Bayesian problems—the motivator for developing the Dirichlet process in the first place. It forms the basis of various applications reported in Chap. 2 and Chap. 3.

The Dirichlet process is conjugate with respect to exact (uncensored) observations.

Theorem 1.1 (Ferguson) *Let $P \in \mathcal{D}(\alpha)$ and given $P = P$, let X be a random sample of size one from P, then the marginal distribution of X is $\overline{\alpha} = \alpha/\alpha(\mathfrak{X})$, the normalized measure corresponding to α. Also, the posterior distribution of P given $X = x$ is $\mathcal{D}(\alpha + \delta_x)$, the Dirichlet process prior with updated parameter $\alpha + \delta_x$, where δ_x is the degenerate probability measure at x. If $X_1, \ldots, X_n$ is sample of size n from P, then the posterior distribution of P given $X_1, \ldots, X_n$ is $\mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{x_i})$.*

Remark 1.1 However, the posterior distribution with respect to right censored observations is no longer a Dirichlet process but is a mixture of Dirichlet processes, a beta-Stacy or a neutral to the right process.

The implication of this theorem is that in obtaining the posterior distribution, all one has to do is to update the parameter α. Lijoi and Prünster (2010) make an interesting observation. They distinguish between the two types of conjugacy: parametric and structural. For the parametric, the distribution of the posterior is same as of prior except that the parameters get updated. The Dirichlet process is an example of this. Whereas for the second type, the posterior distribution has the same structure as the prior in the sense that they both belong to the same general class of processes. Neutral to the right process is an example of this. The first imply the second, but not the other way around.

Sampling a Dirichlet Process As indicated earlier, the Dirichlet process is a stochastic process index by the sets of $\mathcal{A}$. Therefore, to generate a sample from the Dirichlet process with parameter α, we need to assign probability $P(A)$ for each set $A \in \mathcal{A}$. In using the Sethuraman representation $P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot)$ (1.5), we need two things: the realization of weights p_j's such that $p_j \in [0, 1]$ and $\sum_{j=1}^{\infty} p_j = 1$; and independent of them, locations ξ_j's such that $\xi_j \overset{iid}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$. Since p_i's are not independent, they are generated via V_i's. The generated sequences $p_1, p_2, \ldots$ and $\xi_1, \xi_2, \ldots$ are then used in producing a realization of P via $P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot)$. There are two approaches to generate these infinite sequences. One is based on the stick-breaking construction described earlier. The other is the extended Polya urn scheme (Blackwell and MacQueen 1973) which is known outside the statistical community as the Chinese Restaurant process. It involves the process of generating an exchangeable sequence of random variables.

In the first approach, the sequence $\{p_j\}$ is generated by the stick-breaking procedure, taking each $V_j \overset{iid}{\sim} Be(1, \alpha(\mathfrak{X}))$ and then setting $p_1 = V_1$ and $p_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$ for $j \geq 2$. Also, ξ_j's are generated as $\xi_j \overset{iid}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$. Thus a realization of the random probability P will be a map from $([0, 1] \times \mathfrak{X})^{\infty}$ to Π.

For the second approach recall the Polya urn scheme of Sect. 1.1. Suppose we have an infinite number of balls of different colors, denoted by $c_1, c_2, \ldots$ The colors are distributed according to $\overline{\alpha} = \alpha(\cdot)/\alpha(\mathfrak{X})$. At the first step, a ball X_1 is drawn at random from this set according to the distribution $\overline{\alpha}$, and its color is noted. The ball is replaced along with an additional ball of the same color. At the $(n + 1)$-th step, either a ball which is of one of the observed colors is picked with probability $n/(\alpha(\mathfrak{X}) + n)$ or a ball of new color is picked with probability $\alpha(\mathfrak{X})/(\alpha(\mathfrak{X}) + n)$. In both cases, the ball is replaced along with another ball of the same color, and the step is repeated. Thus a sequence $X_1, X_2, \ldots,$ of random variables is generated where X_i is a random color from the set of colors $\{c_1, c_2, \ldots\}$ (note that we do not need to know completely the set $\{c_k\}$ ahead of the time) with $X_1 \sim \alpha/\alpha(\mathfrak{X})$ and $X_{n+1}|X_1, X_2, \ldots, X_n \sim (\alpha + \sum_{i=1}^{n} \delta_{x_i})/(\alpha(\mathcal{X}) + n)$, which can be written equivalently as

$$X_{n+1}|X_1, X_2, \ldots, X_n \sim \sum_{i=1}^{n} \frac{1}{\alpha(\mathfrak{X}) + n} \delta_{x_i} + \frac{1}{\alpha(\mathfrak{X}) + n} \alpha. \tag{1.22}$$

Some colors, say $K \leq n$ will be repeated in n draws. Denote the distinct colors among them by $X_1^*, \ldots, X_K^*$ and let n_k be the number of times the color X_k^* is repeated, $k = 1, 2, \ldots, K$, $n_1 + \ldots + n_K = n$. Then the above expression can be re-written in terms of K, as

$$X_{n+1}|X_1, X_2, \ldots, X_n, K \sim \sum_{k=1}^{K} \frac{n_k}{\alpha(\mathfrak{X}) + n} \delta_{X_k^*} + \frac{1}{\alpha(\mathfrak{X}) + n} \alpha, \tag{1.23}$$

This process is continued indefinitely. It has been interpreted in practical terms and popularized in culinary metaphor by the catchy name, Chinese Restaurant process (attributed to Jim Pitman and Lester Dubins by Griffiths and Ghahramani 2006). The correspondence is as follows.

The sequence $X_1, X_2, \ldots,$ of draws of balls represents incoming patrons at a Chinese restaurant, distinct colors of balls represent tables with different dishes (one dish per table), each of unlimited sitting capacity (that is there are infinite many balls of each color). Each customer sits at a table. The first customer sits at a table and orders randomly a dish for the table according to the distribution $\alpha/\alpha(\mathfrak{X})$. The $(n + 1)$-th customer chooses to join previous customers with probability $n/(\alpha(\mathfrak{X}) + n)$ or chooses a new table with probability $\alpha(\mathfrak{X})/(\alpha(\mathfrak{X}) + n)$ and orders a dish. If he joins previous customers and there are already K tables occupied, then he joins the k-th table X_k^* (or orders dish X_k^*) with probability $n_k/(\alpha(\mathfrak{X}) + n)$, where n_k is the number of customers already occupying the table (or enjoying the dish) X_k^*, $k = 1, 2, \ldots, K$ i.e. $X_{n+1} = X_k^*$. If he chooses a new table, he orders a random dish distributed according to $\alpha/\alpha(\mathfrak{X})$. This results in the above two expressions.

Patrons are exchangeable as are the random variables X_i's in the Polya urn sequence. The probability of a particular sitting arrangement depends only on n_k which is a function of n. Thus a realization of p_k is obtained by

$$p_k = \lim_{n \to \infty} \frac{n_k(n)}{\alpha(\mathfrak{X}) + n} \tag{1.24}$$

and $\xi_j \overset{iid}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$. The CRP is obtained by integrating out the Dirichlet process and thus it describes the marginal distributions in terms of random partitions determined by K tables in a restaurant.

The distinction between the two methods is that in the stick-breaking method the weights generated are exact, where as in Polya sequence process they are approximate. However, in both the methods to sample a P, we need to continue the process for an infinitely long period which is impossible. Therefore termination at some suitable stage is employed. Alternately, one approximate method is to generate a sample from a finite dimensional symmetric Dirichlet distribution with parameter $\alpha(\mathfrak{X})/N$ and use them as weights. That is let $(q_1, \ldots, q_N) \sim \mathrm{D}(\alpha(\mathfrak{X})/N, \ldots, \alpha(\mathfrak{X})/N)$ and then define $P_N(\cdot) = \sum_{j=1}^{N} q_j \delta_{\xi_j}(\cdot)$ with $\xi_j \overset{iid}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$. It can be shown that $P_N \to P$, as $N \to \infty$, in distribution.

There is an extensive literature on sampling methods. For example, see Ishwaran and James (2001), who discuss Gibbs sampling methods, and the papers in the book edited by Dey et al. (1998).

Dirichlet Process Mixtures In solving some Bayesian estimation problems, Antoniak (1974) saw the need to define mixtures of Dirichlet processes by indexing the parameter α of the Dirichlet process by θ and θ having certain parametric distribution. That is, $\theta \sim H(\theta)$, $P|\theta \sim \mathcal{D}(\alpha_\theta)$. Here the Dirichlet processes are mixed with a parametric mixing distribution. On the other hand, Lo (1984) recognizing that the Dirichlet process is inadequate in dealing with density functions, he considered a different type of mixture—mixtures where the mixing distribution is taken to be random with a Dirichlet prior—to serve as priors for density functions. He modeled a random density function on R by $f(x) = \int K(x,s)dG(s)$, where $K(x,s)$ is a known kernel on $R \times \mathfrak{X}$, and proceeded to derive Bayes estimate of $f(x)$ by taking G to be a Dirichlet process. The difference between the two being that the mixing components and mixing distributions are interchanged. To distinguish it from the Mixtures of Dirichlet processes proposed by Antoniak, we will call this mixture as *kernel mixture of Dirichlet processes* (KMDP). Lo considers in his treatment a broad class of kernels, such as histogram, normal density with location and/or scale parameters, symmetric and unimodal densities, decreasing densities, etc. (see Sect. 2.5.4). In defining a random hazard rate, Dykstra and Laud (1981) used a mixture with respect to a gamma process (see Sect. 1.7.1).

Ferguson (1983) considered a countably infinite mixture of normal densities in formulating the density function differently: $f(x) = \sum_{i=1}^{\infty} p_i h(x|\mu_i, \sigma_i)$ where $h(x|\mu,\sigma)$ is the normal density with mean μ and variance σ^2. This formulation has countably infinite number of parameters, $(p_1, p_2, \ldots, \mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$. It can be written as $f(x) = \int h(x|\mu,\sigma)dG(\mu,\sigma)$, where G is the probability measure on the half plane $\{(\mu,\sigma) : \sigma > 0\}$ that gives weight p_i to the point $\xi_i = (\mu_i, \sigma_i)$, $i = 1, 2, \ldots$. While Lo assumes a Dirichlet process prior for the unknown G, Ferguson defines a prior via Sethuraman representation. He defines a prior distribution for the parameter vector $(p_1, p_2, \ldots, \mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$ as follows: vectors $(p_1, p_2, \ldots)$ and $(\mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$ are independent; $p_1, p_2, \ldots$ are the weights with parameter M in Sethuraman representation; and (μ_i, σ_i) are iid with common gamma-normal conjugate prior for the two-parameter normal distribution. This shows that G is a Dirichlet process with parameter $\alpha = MG_0$, where $G_0 = \mathcal{E}(G)$ is the conjugate prior for (μ, σ^2), and its infinite sum representation is $G = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}$ where as usual $(p_1, p_2, \ldots)$ and $(\xi_1, \xi_2, \ldots)$ are independent and $\xi_i \overset{iid}{\sim} G_0$.

In both cases, given a sample $x_1, \ldots, x_n$ of size n from a distribution with density $f(x)$, the posterior distribution of G has been obtained by Antoniak (1974) as a mixture of Dirichlet processes. The Bayesian estimate of density function f is pursued in Sect. 2.5.4.

Normal mixtures also turn up in Escobar (1994) and Escobar and West (1995). Escobar's set up is as follows. Let $Y_i|\mu_i \sim N(\mu_i, 1)$, $\mu_i|G \overset{iid}{\sim} G$, μ_i and G are unknown. His objective, in contrast to those of Ferguson's and Lo's, is to estimate μ_i's (with the variance being known) based on observed Y_i's. He proceeds by assuming a Dirichlet process prior for G.

Escobar and West (1995) describe a normal mixture model for density estimation, similar to Ferguson's (1983), but in terms of the predictive distribution of a future observation. For their model, given (μ_i, σ_i^2), we have independent observations, say $Y_1, \ldots, Y_n$, such that $Y_i|(\mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ and $\nu_i = (\mu_i, \sigma_i^2)$ are drawn from some prior distribution G on $R \times R^+$. Having observed $Y_1, \ldots, Y_n$, the objective is to find the predictive distribution of next observation Y_{n+1} which is a mixture of normals, $Y_{n+1}|Y_1, \ldots, Y_n \sim N(\mu_{n+1}, \sigma_{n+1}^2)$. A usual practice is to put a parametric prior on the vector $\boldsymbol{\nu} = (\mu_1, \ldots, \mu_n, \sigma_1^2, \ldots, \sigma_n^2)$. However, in a particular case of $(\mu_i, \sigma_i^2) = (\mu_i, \sigma^2)$ studied, among others by West (1992), the distribution of μ_i's is modeled via the Dirichlet process with a normal base measure.

In the present case the authors assume $G \sim \mathcal{D}(MG_0)$, where G_0 is the prior guess taken to be a bivariate distribution on $R \times R^+$. In view of the discreteness of Dirichlet process prior which induces multiplicities of observations, $\nu_{n+1}|\nu_1, \ldots, \nu_n$ will have distribution of the form given in property 19 above. Then they proceed on the line of Ferguson, derive the conditional distribution of $Y_{n+1}|\nu_1, \ldots, \nu_n$ which is a mixture of a Student's t-distribution and n normals $N(\mu_i, \sigma_i^2)$. Then it is shown that the unconditional predictive distribution is given by $Y_{n+1}|Y_1, \ldots, Y_n \sim \int \mathcal{P}(Y_{n+1}|\boldsymbol{\nu})d\mathcal{P}(\boldsymbol{\nu}|Y_1, \ldots, Y_n)$.

A third type of mixture is the one which leads to hierarchical models where the parameters of the prior distributions themselves are assigned priors with hyper parameters. It has a long history of applications in parametric and semiparametric set ups. It's adaptation to the case of nonparametric (infinite-dimensional parameters) was undertaken in Teh et al. (2006) where they discuss hierarchical Dirichlet processes and indicate their extensions to other priors. Their interest stems from a need to model group data where each group is associated with a mixture model and the desire is to link them together. This is done for example, by assuming the baseline distribution G_0 of a Dirichlet process itself a Dirichlet process with hyper parameters M^* and G^*, and such models are referred to as *hierarchical Dirichlet models.* In such cases, the discreteness of the Dirichlet process turns out to be an asset (see Sect. 1.15)

Generalizations of the Dirichlet Process The remarkable feature of the Dirichlet process, as the chart in Fig. 1.1 indicates, is that it has led to the development of many generalizations and/or the Dirichlet process is a particular case of these processes. It may rightly be considered as a 'base' prior process giving rise to other priors. It is related to the various processes, discussed in later sections, as follows.

The Dirichlet process is obviously a particular case of the Dirichlet Invariant and mixtures of Dirichlet processes introduced in Sects. 1.3 and 1.4, respectively. When defined on the real line, the Dirichlet process is also a neutral to the right process discussed in Sect. 1.5. A certain transformation of the beta process of Sect. 1.8 yields the Dirichlet process and the Dirichlet process is also a particular case of the beta-Stacy process presented in Sect. 1.9. It is a tailfree process of Sect. 1.10 with respect to every tree of partitions, and when the parameters of the Polya tree process of

Sect. 1.11 are subjected to certain constraints, it yields the Dirichlet process. When the discount parameter of the two-parameter Poisson-Dirichlet process of Sect. 1.13 is set to zero, the process reduces to the Dirichlet process. When the parameters of prior distributions are modeled as Dirichlet process priors, they give rise to different forms of hierarchical models mentioned in Sect. 1.15. Some other generalizations are also listed in that section.

The Sethuraman infinite mixture representation of the Dirichlet process was originally used mainly for proving various properties of the Dirichlet process. However in recent years, as pointed out in Sect. 1.1 and further detailed in Sects. 1.12 and 1.15, it's use as an instrument in developing several related processes, has exploded. This representation has four ingredients and by varying them, a number of new processes have been developed for carrying out Bayesian analysis and modeling large and complex data sets. 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 (Ongaro and Cattaneo 2004). If the weights defined by one parameter beta distribution $Be(1, \alpha(\mathfrak{X}))$ is replaced by two-parameter beta distribution $Be(a_i, b_i)$, a second group of priors emerged (Ishwaran and James 2001; Pitman and Yor 1997). 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 $\xi_{i\mathbf{x}}$. This approach is generalized further in different directions and the resulting priors include processes such as, the *dependent Dirichlet* (MacEachern 1999), *spatial Dirichlet* (Gelfand et al. 2005), *generalized spatial Dirichlet* (Duan et al. 2007), *multivariate spatial Dirichlet* (Reich and Fuentes 2007), *order-based Dirichlet* (Griffin and Steel 2006), *Latent stick-breaking processes* (Rodriguez et al. 2010) etc. to name a few.

For the fourth group, a different type of extension is proposed in which the degenerate probability measure δ is replaced by a nondegenerate positive probability measure G, called the *kernel stick-breaking Dirichlet* process (Dunson and Park 2008). There are no convenient expressions for the posterior distributions available and hence the Bayesian analyzes have to rely on simulation methods. For this purpose, fast running algorithms are developed.

The volume of literature on these processes is vast and it is impossible to do a reasonable justice to them in this book. They could form an excellent fodder for a follow up volume. Therefore they are just mentioned briefly in Sect. 1.12. Chung and Dunson (2011) and Rodriguez et al. (2010) are a good source of references for the reader to dwell further in detail if interested.

The Sethuraman representation, as well as the predictive distribution based on a generalized Polya urn scheme proposed by Blackwell and MacQueen (1973) have also been found useful in the development of new processes, some of them popularly known as the Chinese restaurant and Indian buffet processes discussed in Sect. 1.14. They have applications in nontraditional fields such as word documentation, machine learning and mixture models.

As mentioned in Sect. 1.1, the above processes may be considered as special cases of a class of models proposed by Pitman (1996a) and called *species sampling* *models*,

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot) + \left(1 - \sum_{j=1}^{\infty} p_j\right) Q(\cdot), \tag{1.25}$$

where Q is the probability measure corresponding to a continuous distribution G, $\xi_j \overset{iid}{\sim} G$, and weights p_j's are constrained by the condition, $\sum_{j=1}^{\infty} p_j \leq 1$.

1.3 Dirichlet Invariant Process

The support of the Dirichlet process is sufficiently broad to accommodate any prior belief. However, in treating some nonparametric Bayesian inference problems, Dalal (1979a) realized that the prior information is more structured. For example, without knowing the specific form of the underlying distribution, it may be of interest to estimate the median of a symmetric distribution or test for independence under a permutation symmetry. In such situations, it is evident that a subset of the space of all probability measures or distribution functions would be more efficient to use. Thus there is a justification to consider priors defined on a subset of all distribution functions possessing certain inherent characteristics such as symmetry about a point, exchangeability or invariance under a finite group of transformations. Dalal (1979a) initiated a study of such priors.

1.3.1 Definition

Let $\mathcal{G} = \{g_1, \ldots, g_k\}$ be any finite group of measurable transformations from a p-dimensional Euclidean space $\mathcal{X} \to \mathcal{X}$. A set $B \in \mathcal{X}$ is called $\mathcal{G}$-invariant if $B = gB$ for all $g \in \mathcal{G}$, and a finite non-null measure γ is said to be $\mathcal{G}$-invariant if $\gamma(A) = \gamma(gA)$ for all $g \in \mathcal{G}$ and all $A \in \mathcal{X}$. A measurable partition $(A_1, \ldots, A_m)$ of $\mathcal{X}$ is said to be $\mathcal{G}$-invariant if $A_j = gA_j$ for all $g \in \mathcal{G}$ and $j = 1, \ldots, m$.

Definition 1.3 (Dalal) A $\mathcal{G}$-invariant random probability measure P is said to be a Dirichlet $\mathcal{G}$-invariant process if there exists a $\mathcal{G}$-invariant measure α on $(\mathcal{X}, \sigma(\mathcal{X}))$ such that for every $\mathcal{G}$-invariant measurable partition $(A_1, \ldots, A_m)$ of $\mathcal{X}$, the joint distribution of $(P(A_1), \ldots, P(A_m))$ is $D(\alpha(A_1), \ldots, \alpha(A_m))$. Symbolically $P \in \mathcal{DGI}(\alpha)$.

Using the alternative constructive definition of Ferguson (1973), Dalal proves the existence of such a process.

As is expected, if $\mathcal{G}$ consists of only one element, the identity transformation, the Dirichlet Invariant process corresponds to the Ferguson's Dirichlet process. Dalal derives several properties and applies them to estimate a distribution function known to be symmetric at a known point θ, which will be discussed later in Chap. 2.

Tiwari (1981, 1988) extends the Sethuraman's (1994) alternative representation of the Dirichlet process to the Dirichlet Invariant process. Let α be a $\mathcal{G}$-invariant measure on $(\mathcal{X}, \sigma(\mathcal{X}))$. Let $(p_1, p_2, \ldots)$ and $(\xi_1, \xi_2, \ldots)$ be two independent sequences of iid random variables, such that $\xi_i \sim \overline{\alpha}(\cdot) = \alpha(\cdot)/\alpha(\chi)$; $p_1 = V_1$, and for $j \geq 2$, $p_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$ and $V_j \sim Be(1, \alpha(\chi))$. Then the random probability measure P given by

$$P(A) = \sum_{j=1}^{\infty} p_j \frac{1}{k} \sum_{i=1}^{k} \delta_{g_i \xi_j}(A), \quad A \in \sigma(\mathcal{X}) \tag{1.26}$$

is a Dirichlet Invariant process with parameter α.

1.3.2 Properties

Dalal established the following properties that were shown earlier to hold for the Dirichlet process.

1. Let $P \in \mathcal{DGI}(\alpha)$, and let $A \in \sigma(\mathcal{X})$. Then $P(A) = 0$ with probability one if and only if $\alpha(A) = 0$.
2. Let $P \in \mathcal{DGI}(\alpha)$, and let Z be a real valued measurable function defined on $(\mathcal{X}, \sigma(\mathcal{X}))$. If $\int |Z| d\alpha < \infty$, then $\int |Z| dP < \infty$ with probability one and $\mathcal{E} \int Z dP = \int Z d\alpha / \alpha(\mathcal{X})$.

Samples from the Dirichlet Invariant process are defined in the same way as for the Dirichlet process.

3. Let $P \in \mathcal{DGI}(\alpha)$ and let X be a sample of size 1 from P. Then for $A \in \sigma(\mathcal{X})$

$$\mathcal{P}(X \in A) = \mathcal{P}(X \in gA) = \alpha(A)/\alpha(\mathcal{X}) \quad \text{for any } g \in \mathcal{G}. \tag{1.27}$$

4. Let $P \in \mathcal{DGI}(\alpha)$ and let X be a sample of size 1 from P. Let $B_1, \ldots, B_m$ be a $\mathcal{G}$-invariant measurable partition of $\mathcal{X}$, and $A \in \sigma(\mathcal{X})$. Then

$$\mathcal{P}\big(X \in A, P(B_1) \leq s_1, \ldots, P(B_m) \leq s_m\big) = \sum_{i=1}^{m} \left[\frac{\alpha(A \cap B_i)}{\alpha(\mathcal{X})} D^*(s_1, \ldots, s_m) \right], \tag{1.28}$$

where $D^*(s_1, \ldots, s_m) = D(s_1, \ldots, s_m | \alpha(B_1), \ldots, \alpha(B_i) + 1, \ldots, \alpha(B_m))$.
5. If $P \in \mathcal{DGI}(\alpha)$, then like the Dirichlet process, P is also discrete with probability one.
6. The main property of the Dirichlet Invariant process is the following conjugacy property.

Theorem 1.2 (Dalal) *Let* $P \in \mathcal{DGI}(\alpha)$, *and* $X_1, \ldots, X_n$ *be sample of size n from* P. *Then the posterior distribution of* P *given* $X_1, \ldots, X_n$ *is* $\mathcal{DGI}(\alpha + \sum_{i=1}^{n} \delta_{X_i}^{g})$, *where* $\delta_{X_i}^{g} = (1/k) \sum_{j=1}^{k} \delta_{g_j X_i}$, $i = 1, \ldots, n$.

Using the alternative definition of Tiwari (1981, 1988), Yamato (1986, 1987) and Tiwari prove properties of an estimable parameter φ that are similar to the properties 8 and 9 stated for the Dirichlet process, and extend a weak convergence result (Property 20) of the Dirichlet measures to the Dirichlet Invariant measures as well. When $\mathcal{G}$ is generated by $g(x) = -x$, $\mathcal{DGI}$ gives probability one to the distributions that are symmetric about zero.

1.3.3 Symmetrized Dirichlet Process

Doss (1984) provides an alternative formulation of the symmetrized Dirichlet process on the real line R. Let α_- and α_+ denote the restriction of α to $(-\infty, 0)$ and $(0, \infty)$, respectively and let $F_- \in \mathcal{D}(\alpha_-)$ and $F_+ \in \mathcal{D}(\alpha_+)$. Then $F(t) = \frac{1}{2}F_+(t) + \frac{1}{2}(1 - F_+(-t^-))$ has a symmetrized Dirichlet process prior and F is symmetric about 0. In view of 1–1 correspondence, the construction of a random distribution on R symmetric about 0 is equivalent to the construction of a random distribution function on $[0, \infty)$. If instead we use $F(t) = \frac{1}{2}F_+(t) + \frac{1}{2}F_-(t)$, then F will not be symmetric but $F(0) = \frac{1}{2}$. Denote its prior as $\mathcal{D}^*(\alpha)$. The prior $\mathcal{D}^*(\alpha)$ also has many properties similar to the Dirichlet process in terms of having a large support and meaningful interpretation of parameters.

While Dalal provides a general framework for the invariant Dirichlet process, Doss (1984) provides a deeper extension of the above theory in the case of distributions having a median θ. He extends the construction of symmetric Dirichlet priors to symmetric neutral to the right priors. Let F_1 and F_2 be two independent neutral to the right distribution functions on $[0, \infty)$. He construct a random F as $F(t) = \frac{1}{2}F_1(t) + \frac{1}{2}(1 - F_2(-t^-))$, $t \in R$, a mixture of two neutral to the right distribution functions. Such an F is labeled by Doss as a random distribution function of "the neutral to the right type" and belongs to the set of all CDFs with median 0. Note that this can be expressed in terms of two nonnegative independent increments processes $Y_1(t)$ and $Y_2(t)$ via the representation $F_i(t) = 1 - \exp(-Y_i(t))$, $i = 1, 2$, $t \geq 0$. In the estimation of median θ, F is considered as a nuisance parameter. He uses this representation in deriving the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$ when F is a random distribution neutral to the right type, and the sample $\mathbf{X}$ is obtained from $F(x - \theta)$, and F and θ are assumed to be independent.

1.4 Mixtures of Dirichlet Processes

There are situations where the Dirichlet process is inadequate. For example, consider the following bioassay problem (Antoniak 1974; Ferguson 1974).

Let $F(t)$ denote the probability of a positive response of an experimental animal to a certain drug administered at level $t \geq 0$. We assume that $F(0) = 0$ and that $F(t)$ is non-decreasing with $\lim_{t\to\infty} F(t) = 1$. To learn certain properties of F, we treat n experimental animals at levels $t_1, t_2, \ldots, t_n$, and observe independent random variables $Y_1, \ldots, Y_n$, where for $i = 1, \ldots, n$, Y_i is equal to one if the animal given the dose at level t_i shows a positive response, and Y_i is equal to zero otherwise. We may treat this problem from a non-parametric Bayesian approach by choosing a Dirichlet process prior with parameter α for F, $F \in \mathcal{D}(\alpha)$. In solving the problem, we need to find the posterior distribution of F given the data, which is not the Dirichlet process.

Another example is when the data observed is censored on the right. By this we mean that we do not observe exact observation X, but instead observe the pair $Z = \min(X, Y)$ and $\delta = I[X \leq Y]$, where Y is a censoring variable and δ is an indicator whether we observe X or Y. The problem is to estimate the unknown distribution function F from which X is sampled and $F \in \mathcal{D}(\alpha)$.

For these and other closely related problems, the posterior distribution of F given the data turns out to be a mixture of Dirichlet processes which provides a rationale to study the same. A mixture of Dirichlet processes, roughly speaking, is a Dirichlet process where the parameter α is itself treated as random having a certain distribution. In this sense it is a parametric mixture of the Dirichlet processes. The study of mixtures of Dirichlet processes was initiated by Antoniak (1974), where the bioassay and several other problems are discussed. An important result is contained in Dalal and Hall (1980). They show that a parametric Bayes model can be approximated by a nonparametric Bayes model with mixtures of Dirichlet models so that the prior assigns most of its weight to neighborhoods of parametric model, and that any parametric or nonparametric prior may be approximated arbitrarily closely by a prior which is a mixture of Dirichlet processes. Of late they are extensively used in modeling large scale high-dimensional data. It allows one to proceed in a parametric Bayesian fashion. This approach essentially is considered as a compromise between purely a parametric and purely a nonparametric models in applications.

The simplest mixture of Dirichlet processes would be the one that chooses P from $\mathcal{D}(\alpha_1)$ with probability π, and chooses P from $\mathcal{D}(\alpha_2)$ with probability $1 - \pi$.

The mixture of Dirichlet processes (MDP) should not be confused with the other type of mixtures mentioned in Sect. 1.13. For example consider $f(x) = \int K(x, u) dG(u)$, where $K(x, u)$ is a known kernel and $G \in \mathcal{D}(\alpha)$. Here the parametric functions are mixed with respect to a nonparametric mixing distribution.

1.4.1 Definition

First we need the following definition of a transition measure.

Definition 1.4 (Antoniak) Let $(\Theta, \sigma(\Theta))$ and $(U, \sigma(U))$ be two measurable spaces. A transition measure is a mapping of $U \times \sigma(\Theta)$ into $[0, \infty)$ such that

(a) For every $u \in U$, $\alpha(u, \cdot)$ is a finite, non-negative, non-null measure on $(\Theta, \sigma(\Theta))$.
(b) For every $A \in \sigma(\Theta)$, $\alpha(\cdot, A)$ is measurable on $(U, \sigma(U))$.

As pointed out by Antoniak, this differs from the usual definition of a transition probability since $\alpha(u, \Theta)$ need not be identically equal to one. It is needed so that $\alpha(u, \cdot)$ may serve as a parameter for the Dirichlet process. Also, instead of $\alpha(u, \cdot)$ it is convenient to use the notation $\alpha_u(\cdot)$.

Definition 1.5 (Antoniak) We say that P is a mixture of Dirichlet processes on $(\Theta, \sigma(\Theta))$ with mixing distribution H on $(U, \sigma(U))$ and transition measure α_u, if for all $k = 1, 2, \ldots$ and any measurable partition $A_1, \ldots, A_k$ of Θ, we have the random vector $(P(A_1), \ldots, P(A_k))$ distributed as the mixture

$$\int_U D\big(\alpha_u(A_1), \ldots, \alpha_u(A_k)\big) dH(u), \tag{1.29}$$

i.e.

$$\begin{aligned} &\mathcal{P}\big(P(A_1) \le y_1, \ldots, P(A_k) \le y_k\big) \\ &\quad = \int D\big(y_1, \ldots, y_k \mid \alpha_u(A_1), \ldots, \alpha_u(A_k)\big) dH(u) \end{aligned} \tag{1.30}$$

where $D(\alpha_1, \ldots, \alpha_k)$ denotes the k-dimensional Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$. Concisely, $P \in \int \mathcal{D}(\alpha_u(\cdot)) dH(u)$.

We may consider the index u as a random variable with distribution H, and conditional upon given u, P is a Dirichlet process with parameter α_u, or symbolically, $u \sim H$, $P|u \in \mathcal{D}(\alpha_u)$. The resulting marginal distribution of P is $\int_U \mathcal{D}(\alpha_u) dH(u)$.

A random sample from a mixture of Dirichlet processes is defined in the same way as for the Dirichlet process.

Definition 1.6 (Antoniak) Let P be a mixture of Dirichlet processes on $(\Theta, \sigma(\Theta))$ with transition measure α_u and mixing distribution H. $\theta_1, \ldots, \theta_n$ is said to be a sample from P, if for any positive integer m and measurable sets $A_1, \ldots, A_m$, $C_1, \ldots, C_n$,

$$\begin{aligned} &P\big\{\theta_1 \in C_1, \ldots, \theta_n \in C_n | u, P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_n)\big\} \\ &\quad = \prod_{j=1}^{n} P(C_j) \quad \text{a.s.} \end{aligned} \tag{1.31}$$

1.4.2 Properties

The following properties were established by Antoniak.

1. Let $P \in \mathcal{D}(\alpha)$ and H be a fixed probability measure on $(\Theta, \sigma(\Theta))$. If $\alpha_u(A) = \alpha(A) + \delta_u(A)$, then the process P^* which chooses u according to H, and P from a Dirichlet process with parameter α_u, is a mixture of Dirichlet processes. Also, if $A_1, \ldots, A_k$ is any partition of Θ, then
$$\big(P(A_1), \ldots, P(A_k)\big) \sim \sum_{i=1}^{k} H(A_i) D\big(\alpha_u(A_1), \ldots, \alpha_u(A_i) + 1, \ldots, \alpha_u(A_k)\big). \tag{1.32}$$
2. If $P \in \int \mathcal{D}(\alpha_u(\cdot)) dH(u)$ and θ is a sample of size one from P, then for any measurable set A,
$$\mathcal{P}(\theta \in A) = \int_U \frac{\alpha_u(A)}{\alpha_u(\Theta)} dH(u) \tag{1.33}$$
3. Let $P \in \mathcal{D}(\alpha)$ and θ be a sample of size one from P, and let A be a measurable set such that $\alpha(A) > 0$. Then the conditional distribution of P given $\theta \in A$ is a mixture of Dirichlet processes with transition measure $\alpha_u = \alpha + \delta_u$ for $u \in A$ and mixing distribution $H_A(\cdot) = \alpha(\cdot)/\alpha(A)$. If $A = \Theta$, then it reduces to the Dirichlet process. In symbols, $P|\theta \in A \in \int_A \mathcal{D}(\alpha + \delta_u) dH_A(u)$.
4. The mixtures of Dirichlet processes satisfy the conjugacy property.

Theorem 1.3 (Antoniak) *Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be a sample of size n from P, $P \in \int_U \mathcal{D}(\alpha_u(\cdot)) dH(u)$. Then $P|\boldsymbol{\theta} \in \int_U \mathcal{D}(\alpha_u + \sum_{i=1}^{n} \delta_{\theta_i}) dH_{\boldsymbol{\theta}}(u)$, where $H_{\boldsymbol{\theta}}$ is the conditional distribution of u given $\boldsymbol{\theta}$.*

For further discussion on the term $dH_{\boldsymbol{\theta}}(u)$, you are referred to his paper.

5. An important theorem of his 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. Symbolically,

Theorem 1.4 (Antoniak) *Let θ be a sample of size one from P, $P \in \int_U \mathcal{D}(\alpha_u) dH(u)$, and $X|P, \theta, u \sim F(\theta, \cdot)$, then $P|X = x \in \int_{\Theta \times U} \mathcal{D}(\alpha_u + \delta_\theta) dH_x(\theta, u)$, where H_x is the conditional distribution of (θ, u) given $X = x$.*

6. From the applications point of view, the most important property is that if we have a sample from the Dirichlet process which is distorted by random error, then the posterior distribution of the process given the distorted sample is a mixture of Dirichlet processes. This corollary is a special case of the theorem in property 5.

Corollary 1.1 (Antoniak) *Let $P \in \mathcal{D}(\alpha)$ on $(\Theta, \sigma(\Theta))$, and θ be a sample from P. If X is a random variable such that the conditional distribution of X given P and*

θ is $F(\theta,\cdot)$, then the conditional distribution of P given $X=x$ is a mixture of Dirichlet processes with mixing distribution H, which is the conditional distribution of θ given $X=x$, and transition measure $\alpha_u(\theta,\cdot)=\alpha(\cdot)+\delta_\theta(\cdot)$. In notation, if $P\in\mathcal{D}(\alpha)$, $\theta\sim P$, and $X|P,\theta\sim F(\theta,\cdot)$, then $P|X\in\int_\Theta\mathcal{D}(\alpha+\delta_\theta)dH_x(\theta)$.

In the last two properties computing the posterior distributions H_θ and H_x faces the same kind of difficulties that will be seen with respect to the neutral to the right priors. Antoniak gives a partial solution. However as noted earlier, in recent years a great deal of progress has been made in simulating posterior distributions using computational algorithms. This development has somewhat mitigated the difficulty.

7. Property 3 was proved for a single observation. This was extended by Blum and Susarla (1977) under certain restrictions. This extension is useful in deriving the posterior distribution given the right censored data.

Theorem 1.5 (Blum and Susarla) *Let $P\in\mathcal{D}(\alpha)$, and $\theta_1,\ldots,\theta_k$ be a sample from P. The conditional distribution of P given $\theta_i\in A_i$ for $i=1,\ldots,k$ and $A_1\subseteq A_2\subseteq\ldots\subseteq A_k\in\sigma(\Theta)$ and $\alpha(A_1)>0$ is a mixture of Dirichlet processes with transition measure $\alpha_k(u,A)=\alpha(A)+\sum_{i=1}^k\mu_i(A)+\delta_u(A)$ for $(u,A)\in\Theta\times\sigma(\Theta)$ and with mixing measure μ_k, where $\mu_1,\ldots,\mu_k$ are defined by $\mu_1(A)=\alpha(A\cap A_1)/\alpha(A_1)$ for $A\in\sigma(\Theta)$, and*

$$\mu_l(A)=\frac{\alpha(A\cap A_l\cap A_{l-1}^c)}{\alpha(A_l)+l-1}+\sum_{j=1}^{l-1}\frac{\alpha(A\cap A_j\cap A_{j-1}^c)}{\alpha(A_j)+j-1}\prod_{i=j}^{l-1}\frac{\alpha(A_i)+i}{\alpha(A_{i+1})+i}, \quad (1.34)$$

for $l=2,\ldots,k$, where A^c stands for the complement of A and $A_0=\emptyset$.

1.5 Processes Neutral to the Right

The Dirichlet process was defined on an arbitrary space $\mathfrak{X}$ and involved only one parameter, α. Thus, with specification of α, the prior is completely identified and the Bayesian analysis can proceed. However, it can also be construed as a limitation. Doksum (1974) introduced a significant generalization of the Dirichlet process on the real line, R and called it a *neutral to the right* process. The Dirichlet process was defined in terms of the joint distribution of probabilities of sets which formed a measurable partition of $\mathfrak{X}$. The neutral to the right process is based on the independence of successive normalized increments of a distribution function F. It is conjugate with respect to the data that may include right censored observations. It can be defined in terms of an independent increment processes. In fact F is neutral to the right if and only if the process $Y_t=-\log(1-F(t))$ is a non-decreasing process with independent increments. The neutral to the right process has an advantage over the tailfree processes (to be formally defined later on) in that it may be defined independently of any partition points. The process allows a large amount of freedom to the statistician in choosing a prior. On the other hand, it is difficult to interpret these parameters in terms of prior belief. However, as will be seen later, there are specific processes in the class of neutral to the right processes which are found to be useful in specific applications. Doksum's paper is the main source of the following material.

1.5.1 Definition

Let P be a random probability measure on $(R, \mathcal{B})$, where $\mathcal{B} = \sigma(R)$, and let $F(t) = P((-\infty, t])$ denote the corresponding random distribution function. P and F (as well as their distributions) are said to be neutral to the right if for all partitions $-\infty = t_0 < t_1 < \ldots < t_k < t_{k+1} = \infty$ of R, and k a positive integer, the normalized increments of F,

$$F(t_1), \big[F(t_2) - F(t_1)\big]/\big[1 - F(t_1)\big], \ldots, \big[F(t_k) - F(t_{k-1})\big]/\big[1 - F(t_{k-1})\big] \quad (1.35)$$

are independent. In other words, a random distribution function $F(t)$ on the real line is neutral to the right if for every t_1 and t_2 with $t_1 < t_2$,

$$\frac{1 - F(t_2)}{1 - F(t_1)} \quad \text{and} \quad \big\{F(t) : t < t_1\big\} \quad \text{are independent.} \qquad (1.36)$$

It can be interpreted as the proportion of mass F assigns to the subinterval (t_2, ∞) of the interval (t_1, ∞) is independent of what $F(t)$ does to the left of t_1. The fractions in (1.35) are the hazard contributions of F of respective intervals and are known as *residual fractions*. Further discussion on these fractions is given in Sect. 1.12.

Similarly, F is said to be neutral to the left if the ratios

$$F(t_k), \big[F(t_k) - F(t_{k-1})\big]/F(t_k), \ldots, \big[F(t_2) - F(t_1)\big]/F(t_2) \qquad (1.37)$$

are independent. Because in (1.35) the denominator may be zero with positive probability, the following formal definition is preferred.

Definition (Doksum) A random distribution function F on $(R, \mathcal{B})$ is said to be neutral to the right if for all $m = 1, 2, \ldots,$ and all sequences $t_1 < t_2 < \ldots < t_m$ of real numbers, there exists independent non-negative random variables $V_{1,} V_2, \ldots, V_m$ such that the distribution of the vector $(F(t_1), \ldots, F(t_m))$ is same as the distribution of $(V_1, 1 - (1 - V_1)(1 - V_2), \ldots, 1 - \prod_{i=1}^{m}(1 - V_i))$. Or equivalently, if there exist independent non-negative random variables $V_1, V_2, \ldots, V_m$ such that the distribution of $(1 - F(t_1), 1 - F(t_2), \ldots, 1 - F(t_m))$ is the same as the distribution of $(V_1, V_1 V_2, \ldots, \prod_1^m V_i)$.

By solving the set of equations

$$1 - F(t_i) = \prod_1^i V_i \quad \text{for } i = 1, \ldots, m, \qquad (1.38)$$

for V_i, we find (defining $t_0 = -\infty$ so that $F(t_0) = 0$)

$$V_i = \frac{1 - F(t_i)}{1 - F(t_{i-1})} \quad \text{for } i = 1, \ldots, m. \tag{1.39}$$

Thus, if the difficulties entailed in dividing zero by zero are ignored, a process neutral to the right may be defined to be one for which the ratios (1.35) are stochastically independent.

The definition may also be stated in terms of probability measures as was done with the Dirichlet process. If P is a probability measure on the real line, then for any partition $B_1, B_2, \ldots, B_m, m \geq 1$ with $B_i = (t_{i-1}, t_i], i = 1, \ldots, m-1, t_0 = -\infty$ and $B_m = (t_{m-1}, \infty)$, the vector $(P(B_1), \ldots, P(B_m))$ has the same joint distribution as $(V_1, V_2(1 - V_1), \ldots, V_m \prod_{i=1}^{m-1}(1 - V_i))$. If each V_i is taken to be a beta random variable with parameter (α_i, β_i) and $\beta_i = \sum_{j=i+1}^{m} \alpha_j$, then it yields the Dirichlet process (see property 5 below). Thus the process neutral to the right generalizes the Dirichlet process and provides much more flexibility, and a rich class of models can be developed by specifying different distributions for the variables V_i's. However it should be noted that it is defined on the real line as opposed to the Dirichlet process which is defined on an arbitrary measurable space.

Doksum defines a sample from a random distribution function F in the same way as Ferguson (1973) did (see Sect. 1.2). When we say F is neutral to the right, we also mean that a prior distribution defined on the space $(\mathcal{F}, \sigma(\mathcal{F}))$ is neutral to the right.

Viewing it differently and writing $1 - F(t) = e^{-H(t)}$, a prior can also be specified on the space $\{H(t)\}$ of cumulative hazard rates (as well as $\mathcal{F}$), by assigning finite dimensional distributions to the hazard contributions $q_1, \ldots, q_k$ of F, where for any partition $-\infty < t_1 < \ldots < t_k < \infty$ of the real line, $q_j = (F(t_j) - F(t_{j-1}))/(1 - F(t_{j-1})) = 1 - V_j$, is the hazard contribution of the j-th interval (Kalbfleisch 1978). Assigning independent prior distributions to q_j's, subject to some consistency requirement, results in the process neutral to the right. In the case of Dirichlet process, q_j's have independent beta distributions.

By associating a variable $x \in \chi$, James (2006) extends the definition of neutral to the right processes to a class of random probability measures on more general spaces $[0, \infty) \times \chi$ and calls them *spatial neutral to the right processes.* Essentially a spatial neutral to the right process is a random probability measure where $x \in \chi$ is associated with the cumulative hazard function H in the same way as in the Dependent (and Spatial) Dirichlet process the covariate x (or random field) was attached to the random distribution function F. He gives a description of the marginal distributions of this class of models. This extension facilitates the implementation of the neutral to the right mixture models. Details may be found in his paper.

1.5.2 Properties

Various properties of the neutral to the right process were shown to hold by Doksum.

1. F is neutral to the right is equivalent to the following. For all $t_1 < \ldots < t_k$, $k \geq 1$ there exists independent random variables $V_1, \ldots, V_k$ such that $(F(t_1), F(t_2) - F(t_1), \ldots, F(t_k) - F(t_{k-1}))$ and $(V_1, V_2(1-V_1), \ldots, V_k \prod_1^{k-1}(1-V_i))$ have the same distribution.
2. The following connection is seen between a random distribution function and a particular stochastic process which is exploited in developing several prior processes.

Theorem 1.6 (Doksum) *$F(t)$ is a random distribution function neutral to the right if and only if it has the same distribution as of process $1 - e^{-Y(t)}$ for some a.s. non-decreasing, a.s. right continuous, independent increments process with $\lim_{t\to-\infty} Y(t) = 0$ a.s. and $\lim_{t\to\infty} Y(t) = \infty$ a.s.*

As noted by Doksum, this theorem shows that the Dirichlet random distribution function F is not the only random distribution function that is neutral to the right. In fact, different random distribution functions which are neutral to the right may be constructed corresponding to the processes with independent nonnegative increments and their Laplace transforms. He gives some examples in his paper and other examples, presented later, include gamma, beta and beta-Stacy processes.

In view of this theorem, one can compute $\mathcal{E}(F(t))$ by using $\mathcal{E}(F(t)) = 1 - \mathcal{E}(e^{-Y(t)})$.

3. If a random distribution function $F(t)$ is neutral to the right and if $-\log(1 - F(t))$ has no nonrandom part, then F is discrete with probability one.
4. If F is neutral to the right and neutral to the left, then F is either a Dirichlet process (on R) or a limit of Dirichlet processes or processes concentrated on two nonrandom points.
5. If V_i's in the above definition of a neutral to the right process, are chosen to be $Be(\alpha_i, \beta_i)$ such that $\beta_i = \sum_{j\geq i+1} \alpha_i$, then it reduces to the Dirichlet process as noted before.
6. Dirichlet process is neutral to the right process with respect to every sequence of nested, measurable ordered partitions. This can be seen as follows. For each $m = 1, 2, \ldots$ consider the sequence of nested partitions, $\{\pi_m\} = \{A_{m1}, \ldots, A_{mk_m}\}$ denoting the ordered partition π_m of R. We need to show that for each m, there exists independent family of random variables $V_{m1}, V_{m2}, \ldots, V_{mk_m}$ such that the joint distribution of the vector $(P(A_{m1}), \ldots, P(A_{mk_m}))$ has the same distribution as $(V_{m1}, V_{m2}(1 - V_{m1}), \ldots, V_{mk_m} \prod_{j=1}^{k_m-1}(1 - V_{mj}))$, namely, the Dirichlet distribution with parameter α. For this it would be sufficient to show that $P(A_{mi})$ and $V_{mi} \prod_{j=1}^{i-1}(1 - V_{mj})$ have the same distribution, namely, $Be(\alpha(A_{mi}), \alpha(R) - \sum_{j=1}^{i} \alpha(A_{mj}))$, for some i, $1 \leq i \leq k_m$. To see this define $V_{m1} = P(A_{m1})$, $V_{m2} = P(A_{m2}|A_{m1}^c)$, $V_{m3} = P(A_{m3}|A_{m1}^c \cap A_{m2}^c)$, $\ldots$ and so on. Now take each of them distributed independently as beta distribution with parameters $(\alpha(A_{m1}), \alpha(R) - \alpha(A_{m1}))$, $(\alpha(A_{m2}), \alpha(R) - \alpha(A_{m1}) - \alpha(A_{m2}))$, $(\alpha(A_{m3}), \alpha(R) - \sum_{j=1}^{3} \alpha(A_{mj}))$, etc., respectively. Continuing in this way, it can be seen that $P(A_{mi}) = V_{mi} \prod_{j=1}^{i-1}(1 - V_{mj})$, for $i = 2, \ldots, k_m$. Now using

the properties of the beta distribution, it can be seen that $V_{mi}\prod_{j=1}^{i-1}(1-V_{mj}) \sim Be(\alpha(A_{mi}), \alpha(R)-\sum_{j=1}^{i}\alpha(A_{mj}))$.

7. Neutral to the right processes satisfy the structural conjugacy property:

Theorem 1.7 (Doksum) *Let $X_1, \ldots, X_n$ be a sample from F which may include right censored observations. If F is a random distribution function neutral to the right, then the posterior distribution of F given the data is also neutral to the right.*

It should be noted that the posterior will have discontinuities irrespective of whether the prior had them or not. Doksum proved the conjugacy property in a complicated way in terms of finding the posterior distributions of variables V_j's given the sample, and then extending to F. Recall that the posterior distribution of the Dirichlet process was simple—all you had to do was to update its parameter, α. Since the neutral to the right process has many more parameters, it would be difficult to find the posterior in a similar way. Doksum gives the description of it, but it is complicated. Ferguson (1974) provided an alternative description of the posterior which is much simpler. This description was further extended to cover the right censored data in Ferguson and Phadia (1979). It turns out that the neutral to the right process is particularly well suited and easier to use as a prior for the right censored data than for exact observations.

8. Description of the posterior is given later under the heading of "Posterior distribution".
9. *Characterization*: A characterization of the process is described later as it utilizes the Lévy process representation.

Property 2 shows a close connection between the neutral to the right process and a process with independent nonnegative increments. This connection plays an important role in developing other prior processes presented later. Therefore, we digress a bit to consider it in slightly more detail.

1.5.3 Non-decreasing Processes with Independent Increments[1]

Non-decreasing processes with independent increments (also called positive Lévy processes and subordinators) play an important role in nonparametric Bayesian analysis. Ferguson's (1973, 1974) alternative definition of the Dirichlet process, which is contrary to one's intuition, is described through an independent increment process and the corresponding Lévy measure. As will be seen later on that several other

[1] Part of the material of this and the next two subsections is based on Ferguson (1974), Ferguson and Phadia (1979) and Ferguson's unpublished notes which clarify and provide further insight into the description of the posterior processes neutral to the right. I am grateful to Tom Ferguson for passing on his notes to me which helped in developing these sections.

prior processes also emerge from non-decreasing processes with independent increments (and their Lévy measures). In fact all that is needed is an infinitely divisible random variable and a baseline distribution function (Doksum 1974). Besides the processes neutral to the right, they include beta, beta-Stacey and gamma processes that are to be discussed in later sections. Another important advantage is that it can be used in describing the posterior distributions and in treating inference problems involving right censored data. In view of this, we first briefly discuss the stochastic processes with independent increments (rigorous treatment of which may be found in any standard textbook on probability) and connect them with the above mentioned processes.

A stochastic process X_t with $t \in R$ has independent increments if for every positive integer n and for every real numbers $t_0 < t_1 < \ldots < t_n$ the increments $X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}}$ are stochastically independent. We consider below only processes with independent increments for which the sample paths X_t are non-decreasing with probability one. Thus each increment $X_t - X_s$ for $t > s$ is non-negative with probability one. The process may have at most countable number of fixed points of discontinuities. Let S_t be the height of the jump at t which is non-negative with probability one and may have any distribution without disturbing the independence of the increments. Thus the processes with independent increments have two components: one component is the process in which positive increments (jumps) and the location points are both random, where as in the second component, only the jumps are random but the locations are fixed. This may happen a priori as a part of the prior process or as a result of a sampled observation.

However, for a non-decreasing process with independent increments without fixed points of discontinuity, the distributions of the increments are known to be restricted to the class of infinitely divisible laws, which may be worthwhile to review briefly first.

Infinitely Divisible Laws A random variable X is said to be infinitely divisible, and its law is said to be an infinitely divisible distribution, if for every positive integer n it can be represented as a sum

$$X = X_{n1} + X_{n2} + \ldots + X_{nn},$$

where $X_{n1}, X_{n2}, \ldots, X_{nn}$ are independent and identically distributed random variables. Here we are mainly concerned with the case where the infinitely divisible random variable is non-negative. In such a case, the one-sided Laplace transform of the distribution exists and has a very simple form. (See, for example, Chap. XIII.7 in Feller 1966.) It's logarithm form for $\theta \geq 0$ is,

$$\psi(\theta) = \log \mathcal{E} e^{-\theta X} = -\theta b + \int_0^\infty \left(e^{-\theta z} - 1\right) dN(z), \tag{1.40}$$

where the location parameter $b \geq 0$, and N is a measure, called the Lévy measure, on the open interval $(0, \infty)$ such that

$$\int_0^1 z dN(z) < \infty \quad \text{and} \quad \int_1^\infty dN(z) < \infty. \tag{1.41}$$

Some simple well known examples of a non-negative infinitely divisible random variable are:

1. The *Poisson* random variable X with parameter λ defined on $0, c, 2c, \ldots, c > 0$, and having a probability mass function $P(X = nc) = e^{-\lambda}\lambda^n/n!$, $n = 0, 1, 2, \ldots,$ has the log Laplace transform $\psi(\theta) = \lambda(e^{-\theta c} - 1)$ which is (1.40) with $b = 0$ and Lévy measure that assigns mass λ to the point c.
2. The *gamma* random variable X defined on $(0, \infty)$ with density
$$f(x) = \Gamma(\alpha)^{-1}\beta^{-\alpha}e^{-x/\beta}x^{\alpha-1}I_{(0,\infty)}(x), \tag{1.42}$$
where $\alpha > 0$ and $\beta > 0$, has the log Laplace transform $\psi(\theta) = -\alpha \log(1 + \beta\theta)$ which may be represented in the form of Eq. (1.40) as
$$\psi(\theta) = \alpha \int_0^\infty \left(e^{-\theta z} - 1\right)e^{-z/\beta}z^{-1}dz \tag{1.43}$$
with $b = 0$ and the Lévy measure as $dN(z) = \alpha e^{-z/\beta}z^{-1}dz$. This measure gives infinite mass to the positive axis but does satisfy conditions (1.41).
3. The random variable $X = -\log Y$, where $Y \sim Be(\alpha, \beta)$, has the *Log Beta distribution* with density
$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}e^{\alpha x}\left(1 - e^x\right)^{\beta-1}I_{(0,\infty)}(x), \tag{1.44}$$
and the log Laplace transform of X,
$$\psi(\theta) = \log\Gamma(\alpha+\beta) + \log\Gamma(\alpha+\theta) - \log\Gamma(\alpha+\beta+\theta) - \log\Gamma(u), \tag{1.45}$$
which may be represented in the form of Eq. (1.40) as
$$\psi(\theta) = \int_0^\infty \left(e^{-\theta z} - 1\right)\frac{e^{-\alpha z}(1 - e^{-\beta z})}{(1 - e^{-z})z}dz \tag{1.46}$$
with $b = 0$ (see Lemma 1 in Ferguson 1974). Again the corresponding Lévy measure
$$dN(z) = \frac{e^{-\alpha z}(1 - e^{-\beta z})}{(1 - e^{-z})z}dz \tag{1.47}$$
is not finite yet satisfies conditions (1.41).

Other examples include the negative binomial distribution, and the completely asymmetric stable distributions with characteristic exponent less than unity.

Non-decreasing Infinitely Divisible Processes Let Y_t be an almost surely non-decreasing process with independent increments and no fixed points of discontinuity. As noted above, it is known that all increments $Y_t - Y_s$, $s < t$ are infinitely divisible. Therefore, we may write the log Laplace transform of the increment $Y_t - Y_0$ as

$$\psi_t(\theta) = -b(t)\theta + \int_0^\infty \left(e^{-\theta z} - 1\right) dN_t(z), \tag{1.48}$$

where $N_t(z)$ is a Lévy measure depending on t. (If N_t does not depend on t, then the process is known as homogeneous.) It may also be written as $dN_t(z) = \nu(dt, dz)$, ν being a finite measure.

Here $b(t)$ is easily seen to be continuous and non-decreasing.

Lévy Measures of Different Processes As mentioned before, the Lévy measure $N_t(z)$ plays a critical role and identifies different processes, discussed in subsequent sections, as follows.

i. For the gamma process with shape parameter $\gamma(t)$ and intensity parameter τ, the Lévy measure is given by $dN_t(s) = \gamma(t)e^{-\tau s} ds/s$.

ii. For the simple homogeneous process (Ferguson and Phadia 1979), the Lévy measure turns out to be $dN_t(s) = \gamma(t)dN(s)$, where N is any measure on the interval $(0, \infty)$ such that $\int_0^\infty \frac{z}{1+z} dN(z) < \infty$.

iii. The log-beta process (Walker and Muliere 1997a) with parameters $\alpha(\cdot)$, a right continuous measure, and $\beta(\cdot)$, a positive function, both defined on the interval $[0, \infty)$, has the Lévy measure

$$dN_t(s) = \int_0^t \frac{\exp(-s(\beta(u) + \alpha\{u\}))d\alpha_c(u)}{(1 - e^{-s})} ds, \tag{1.49}$$

where $\alpha_c(\cdot)$ is the continuous part of $\alpha(\cdot)$.

iv. The Lévy measure of the beta process $A(t)$ (Hjort 1990) with parameters $A_0(t)$ and $c(\cdot)$ is given by

$$dN_t(s) = \int_0^t \frac{c(z)(1 - s)^{c(z)-1}}{s} dA_{0c}(z) ds, \tag{1.50}$$

where $A_{0c}(z)$ is the continuous part of $A_0(t)$ and $c(\cdot)$ is a piece wise continuous nonnegative function on $[0, \infty)$. Unlike the others mentioned above, this Lévy measure is concentrated on the interval $[0, 1]$.

v. The Lévy measure

$$dN_t(s) = \int_0^t \frac{c(u)\exp(-sc(u)G(u, \infty))dG_c(u)}{(1 - e^{-s})} ds \tag{1.51}$$

defines a beta-Stacy process (Walker and Muliere 1997a) with parameters $G(\cdot)$ and $c(\cdot)$, $G_c(\cdot)$ is the continuous part of the right continuous distribution function $G(\cdot)$, and $c(\cdot)$ is a positive function on $[0, \infty)$.

vi. For specific parameters of the beta-Stacy process such that, G is continuous and $c(u) = c$ a constant, then

$$dN_t(s) = \int_0^t \frac{c \exp(-sc(1 - G(u)))dG(u)}{(1 - e^{-s})} ds, \tag{1.52}$$

which upon simplification reduces to

$$dN_t(s) = \frac{e^{-sc}(e^{(1-G(t))} - 1)}{s(1 - e^{-s})} ds \tag{1.53}$$

which is the Levy measure for the Dirichlet process given in Ferguson (1974).

vii. In connection with the Indian Buffet process, Thibaux and Jordan (2007) introduced the Hierarchical Beta process where the Lévy measure of the beta process was modified for continuous A_0 as

$$\upsilon(dz, ds) = \frac{c(z)(1 - s)^{c(z)-1}}{s} A_0(dz)ds$$

on $\Omega \times [0, 1]$, where $c(\cdot)$ is a positive function on Ω, a probability space, known as concentration function, and A_0 is a fixed base measure on Ω.

Completely Random Measures All of the above prior processes were developed by taking different forms of independent increment processes and their Levy measures. Lijoi and Prünster (2010) recast the processes with independent increments in a more general and elegant manner in terms of completely random measures (Kingman 1967) tying together a number of neutral to the right type processes. Let Π^* be a space of finite measures μ^* on $(\mathfrak{X}^*, \mathcal{A}^*)$ such that $\mu^*(A) < \infty$ for all bounded sets $A \in \mathcal{A}^*$, and $\sigma(\Pi^*)$ be the corresponding σ-algebra. A measure μ is said to a *completely random measure* if, $\mu(A)$ is a random variable defined on some probability space $(\Omega, \mathcal{F}, Q)$ into $(\Pi^*, \sigma(\Pi^*))$ for all $A \in \mathcal{A}^*$ (think of a random probability measure P), and for any pairwise disjoint sets $A_1, \ldots, A_k$ in $\mathcal{A}^*$, the random variables $\mu(A_1), \ldots, \mu(A_k)$ are mutually independent. That is, random measures assigned to disjoint subsets are independent. This way the whole distribution over μ is determined once the distributions of random variables $\mu(A)$ are given for all $A \in \mathcal{A}^*$. The completely random measures are discrete with probability one. $\mu((0, t])$ on the real line may be viewed as a stochastic process with independent increments. As in processes with independent increments, the measure μ may be described as having two components: μ_1 representing the part with fixed points of discontinuity c_i's with jumps S_i's, and μ_2 representing random points of discontinuities ξ_i's with jumps J_i's, respectively. Both sets of jumps are random, non-negative and mutually independent. Write $\mu = \mu_1 + \mu_2 = \sum_{i \geq 1} S_i \delta_{c_i} + \sum_{i \geq 1} J_i \delta_{\xi_i}$. The jumps in μ_1 will occur at sample observations, whether the prior process may or may not have them. With the fixed points of discontinuity part removed, μ will admit Levy representation $\log \mathcal{E} e^{-\theta \mu_2(0,t]} = \int_0^\infty (e^{-\theta z} - 1)dN_t(z)$, where $dN_t(z)$ is the Levy measure, at times written as $dN_t(z) = \nu(dz, dt) = \rho_t(dz)\alpha(dt)$, where α

is a measure on $(\mathfrak{X}^*, \mathcal{A}^*)$ and ρ is a transition kernel on $\mathfrak{X}^* \times \mathcal{B}(R^+)$. If $\rho_t = \rho$, then the distribution of jumps in μ_2 are independent of their location and both ν and μ_2 are termed *homogeneous*; otherwise they are *non-homogeneous* (see Lijoi and Prünster 2010). The posterior distribution is also articulated and expressed in terms of μ which clearly reveal how the components are updated after observing the sample.

Now it can be seen that as above different forms of Levy measure ν defines different processes as observed in their paper. In view of the similarity of the completely random measure approach to the one presented above, we prefer to stick with the above approach for two reasons. One, it provides a historical perspective of the development of these processes, and perhaps easy to understand. Two, it also reveals how these measures came about, which is not clear by the completely random measures approach. This is evident for example in the development of the beta and beta-Stacy processes.

1.5.4 Alternate Representation of the Neutral to the Right Process

Now coming back to the neutral to the right process, the random distribution function F, in view of Theorem 1.6, may also be viewed alternatively (Doksum 1974, Ferguson 1974) in terms of a process with independent increments.

Let $Y_t = -\log(1 - F(t))$ where $Y_t = +\infty$ if $F(t) = 1$, so that

$$F(t) = 1 - e^{-Y_t}. \tag{1.54}$$

Then the process Y_t has independent increments, since for any partition $t_1 < t_2 < \ldots < t_m$ of R, the increments $Y_{t_1}, Y_{t_2} - Y_{t_1}, \ldots, Y_{t_m} - Y_{t_{m-1}}$ correspond to the independent normalized increments in F. Furthermore, since $F(t)$ is assumed to be a distribution function a.s., it is non-decreasing a.s., right continuous a.s., $\lim_{t\to-\infty} F(t) = 0$ a.s., and $\lim_{t\to+\infty} F(t) = 1$ a.s. Translating these properties in terms of the Y_t, we may state the following alternative definition of a process neutral to the right.

Definition 1.7 (Doksum) Let Y_t be a process with independent increments, non-decreasing a.s., right continuous a.s., $\lim_{t\to-\infty} Y_t = 0$ a.s., and $\lim_{t\to+\infty} Y_t = \infty$ a.s. Then (1.54) defines a random distribution function neutral to the right. (Y_t is allowed to be $+\infty$ with positive probability for finite t.)

Thus, $F(t)$ may be decomposed using the above representation. The process Y_t has at most countably many fixed points of discontinuity at say, $t_1, t_2, \ldots$ in some order. Let $S_1, S_2, \ldots$ represent the random heights of the jumps at $t_1, t_2, \ldots$ respectively. Then $S_1, S_2, \ldots$ are independent, non-negative, possibly infinite-valued, random variables with densities, say $f_1, f_2, \ldots$ with respect to some convenient measure. The jumps $\{S_j\}$ are also independent of the rest of the process, so with the jumps removed let

$$Z_t = Y_t - \sum_j S_j I_{[t_j,\infty)}(t). \tag{1.55}$$

This process has independent increments, is non-decreasing a.s., and has no fixed points of discontinuity. Therefore, Z_t has infinitely divisible increments and its moment generating function has Lévy formula

$$\log \mathcal{E}e^{-\theta Z_t} = -\theta b(t) + \int_0^\infty \left(e^{-\theta z} - 1\right) dN_t(z), \tag{1.56}$$

where b is non-decreasing and continuous with $b(t) \to 0$ as $t \to -\infty$, and where N_t is a continuous Lévy measure, that is,

i. for every Borel set B, $N_t(B)$ is continuous and non-decreasing,
ii. for every real t, $N_t(\cdot)$ is a measure on the Borel subsets of $(0, \infty)$,
iii. $\int_0^\infty (z/(1+z)) dN_t(z) \to 0$ as $t \to -\infty$.

Thus, going back to the processes neutral to the right, they can be specified by giving four things:

(a) $t_1, t_2, \ldots$, the fixed points of discontinuity
(b) $f_1, f_2, \ldots$, the densities (or distributions) of the jumps there
(c) $b(t)$ the continuous deterministic part, and
(d) $N_t(B)$ the continuous Lévy measure.

The function b corresponds to the continuous deterministic part of the process Y_t. If there are no fixed points of discontinuity and if the Lévy measure vanishes identically, then $F(t)$ is the fixed nonrandom distribution function

$$F(t) = 1 - e^{-b(t)}. \tag{1.57}$$

On the other hand, if $b = 0$, then Y_t and hence $F(t)$ increases only in jumps a.s., so that F is discrete with probability one. In general, except for the fact that an F may have a continuous nonrandom part as a mixture, processes neutral to the right do not avoid the drawback, noted for the Dirichlet process, of choosing discrete probability distributions with probability one.

1.5.5 Posterior Distribution

In the case of the Dirichlet process (also mixtures of Dirichlet processes and Dirichlet invariant process), the conjugacy was parametric and therefore it was easy to describe the posterior, which remained the same as the prior but with an updated parameter α to $\alpha + \sum_{i=1}^n \delta_{X_i}$. However, it is not that simple in the case of the neutral to the right process. Since there are other processes in the class of neutral to the right that are included in the later sections, it seems instructive to report the description in some what more detailed manner. Doksum's description of the posterior distribution is in terms of the posterior distributions of the normalized increments V_j's and is complicated. Ferguson (1974) (and later Ferguson and Phadia 1979) gives an explicit and simpler description and shows that as the uncensored data are handled conveniently by the Dirichlet process, the censored data can also be handled with ease by neutral to the right processes.

Heuristically it can be described as follows. Let $t_1 < t_2 < \ldots < t_k$ represent a large number of partition points. The distribution of $(F(t_1), \ldots, F(t_k))$ is related to $(Y_{t_1}, \ldots, Y_{t_k})$, which may be described even more simply through the variables $(Z_1, \ldots, Z_k)$ where $Z_i = Y_{t_i} - Y_{t_{i-1}}$ represents the i-th increment, and $Y_{t_0} \equiv 0$. In writing the joint distribution of the Z_i and an observation X from F, we approximate by reducing the information about X to that of knowing into which interval $(t_{j-1}, t_j]$ X falls. Then, if $f_i(z_i)$ denotes the density of Z_i, the joint density of $Z_1, \ldots, Z_k$ and X may be written as

$$\begin{aligned}
f(z_1, \ldots, z_k, x) &= \left(\prod_{i=1}^{k} f_i(z_i)\right) P\big(X \in (t_{j-1}, t_j] \mid \mathbf{Z} = \mathbf{z}\big) \\
&= \left(\prod_{i=1}^{k} f_i(z_i)\right) \big(F(t_j) - F(t_{j-1})\big) \qquad (1.58) \\
&= \left(\prod_{i=1}^{k} f_i(z_i)\right) e^{-\sum_{i=1}^{j-1} z_i} \big(1 - e^{-z_j}\big) \\
&= \left(\prod_{i=1}^{j-1} f_i(z_i) e^{-z_i}\right) f_j(z_j)\big(1 - e^{-z_j}\big) \left(\prod_{i=j+1}^{k} f_i(z_i)\right). \qquad (1.59)
\end{aligned}$$

The posterior density of $Z_1, \ldots, Z_k$ given $X \in (t_{j-1}, t_j]$ is then proportional to this quantity. From this, we see that given $X \in (t_{j-1}, t_j]$ the Z_i are still independent (and hence the posterior distribution of F should still be neutral), and that the distributions of the increments to the right of t_j have not changed (a rationale to coin the word 'tailfree'), while the distributions of the increments to the left of t_{j-1} are changed by multiplying the density by e^{-z} and renormalizing. If there is a prior fixed point of discontinuity at x, the posterior density of the jump at x is obtained by multiplying the prior density by $(1 - e^{-zt})$ and normalizing. For a sample of size n, this may be stated in part as follows.

Theorem 1.8 (Doksum 1974) *The posterior distribution of an increment $Z = Y_t - Y_s$ of an interval $(s, t]$ in which no observations fall is obtained by multiplying the prior density of Z by e^{-rz} and renormalizing, where r is the number of observations among the sample of size n that are greater than t.*

If x is a prior fixed point of discontinuity, then the posterior density of the jump Z in Y_t at x is obtained by multiplying the prior density of Z by $e^{-rz}(1 - e^{-z})^m$ and normalizing, where r is the number of observations greater than x, and m is the *number of observations equal to x. That is if $g_x(z)$ is the prior density of the jump Z in Y_t at x, then the posterior density $g_x^*(z)$ is given by*

$$g_x^*(z) = \frac{e^{-rz}(1-e^{-z})^m g_x(z)}{\int e^{-rz}(1-e^{-z})^m g_x(z)dz} \tag{1.60}$$

There is problem in extending this to the case if x is *not* a prior fixed point of discontinuity. If x is a point at which one or more observations fell, then the posterior distribution of Y_t may have a fixed point of discontinuity at x even if the prior did not have fixed point discontinuity at x. On the other hand, there may be no change in the posterior, as is the case when x is in a region where $b(t)$ increases but $N_t(R)$ does not, since then it is known that the observation $X = x$ arose from the nonrandom part of the distribution. The general case is a mixture of these two cases. It is sufficient to state the theorem for a sample of size one, since larger samples may be handled by an application of Theorem 1.8.

Define for each Borel set $B \subset [0,\infty)$ a measure $\mu_B(\cdot)$ on the Borel subsets of R to satisfy

$$\mu_B\big((-\infty,t]\big) = b(t)I_B(0) + \int_B \big(1-e^{-z}\big)dN_t(z). \tag{1.61}$$

Note that $\mu_B \ll \mu_{[0,\infty)}$, so that the Radon-Nikodym derivative

$$\nu_B(t) = \frac{d\mu_B}{d\mu_{[0,\infty)}}(t) \tag{1.62}$$

exists for all B.

Theorem 1.9 (Ferguson) *Let $F(t)$ be a process neutral to the right and let X be a sample of size* 1 *from F. If x is not a prior fixed point of discontinuity, then the posterior distribution of the jump S in Y_t at x, given $X = x$ is given by*

$$H_x(z) = \nu_{[0,z]}(x). \tag{1.63}$$

A complete description of the posterior distribution is given in Sect. 3.3.2 for the right censored data, and formulas are given in terms of the moment generating function (MGF). (Precise derivation for the estimation of survival function is carried out there in two specific cases.) A similar description is given for the beta process prior in Sect. 1.8 and the same holds for the beta-Stacy process of Sect. 1.9.

It may be difficult to evaluate H_x in general, but as noted in Ferguson (1974) and Ferguson and Phadia (1979), it is easy to evaluate in one important special case of an homogeneous process. It is given here in the form of a MGF.

Posterior Distribution for a Simple Homogeneous Process A random distribution function neutral to the right is *homogeneous* if the independent increment process $Y_t = -\log(1-F(t))$ has log moment generating function

$$\log \mathcal{E}e^{-\theta Y_t} = \gamma(t)\int_0^\infty \big(e^{-\theta z}-1\big)dN(z). \tag{1.64}$$

Thus there are no prior fixed points of discontinuities, $b(t) \equiv 0$ and the Lévy measure of Y_t has the simple form, $dN_t(z) = \gamma(t)dN(z)$.

The posterior distribution of the jump S in Y_t at a point x that is not a prior fixed point of discontinuity is given by

$$H_x(s) = \int_0^s \left(1 - e^{-z}\right)dN(z) \Big/ \int_0^\infty \left(1 - e^{-z}\right)dN(z), \tag{1.65}$$

independent of x and of $\gamma(t)$.

For this prior process it is better to describe the posterior in terms of prior and posterior MGF, $M_t(\theta) = \mathcal{E}e^{-\theta Y_t}$. Then it is easy to compute the expectation of F,

$$\mathcal{E}F(t) = \mathcal{E}\left(1 - e^{-Y_t}\right) = 1 - M_t(1). \tag{1.66}$$

Consider first a single observation $X = x$ from F. The posterior MGF of Y_t for a value $t < x$ becomes simply

$$M_t(\theta \mid x) = M_t(\theta + 1)/M_t(1). \tag{1.67}$$

However, for $t > x$, we write Y_t as the sum of three increments, from $-\infty$ to x^-, the jump at x, and from x^+ to t. Independence of these three increments means we can write the posterior MGF as a product (for $t > x$):

$$M_t(\theta \mid x) = \frac{M_x^-(\theta + 1)}{M_x^-(1)} \cdot \varphi_x(\theta) \cdot \frac{M_t(\theta)}{M_x(\theta)}, \tag{1.68}$$

where $M_t^-(\theta)$ represents $\lim_{s \nearrow t} M_s(\theta)$, the MGF of $Y_t^- = \lim_{s \nearrow t} Y_s$, $s < t$; $\varphi_x(\theta)$ represents the posterior MGF of the jump in Y_t at x; and where $M_t(\theta)/M_x(\theta)$ is the prior and hence the posterior MGF of $Y_t - Y_x$ as well (since x is not greater than t).

This argument is easily generalized to a sample of n observations, $X_1, \ldots, X_n$. Let $u_1 < u_2 < \ldots < u_k$ represent the distinct values among $X_1, \ldots, X_n$, and let $\delta_1, \ldots, \delta_k$ denote the numbers of observations at $u_1, \ldots, u_k$ respectively. We refer to $\mathbf{x}$ or to $(\mathbf{u}, \boldsymbol{\delta})$ as the *data*. Let $h_j = \sum_{i=j+1}^k \delta_j$ represent the number of observations greater than u_j, and for fixed t, let $j(t)$ be such that $u_{j(t)} \le t < u_{j(t)+1}$, where for $t < u_1$, $j(t) = 0$ and $u_0 = -\infty$. Then, the posterior MGF is given by

$$\begin{aligned} M_t(\theta \mid data) = \prod_{i=1}^{j(t)} &\left[\frac{M_{u_i}^-(\theta + h_{i-1})}{M_{u_i}^-(h_{i-1})} \cdot \frac{C_{u_i}(\theta + h_i, \delta_i)}{C_{u_i}(h_i, \delta_i)} \cdot \frac{M_{u_i}(h_i)}{M_{u_i}(\theta + h_i)} \right] \\ &\cdot \frac{M_t(\theta + h_{j(t)})}{M_t(h_{j(t)})}, \end{aligned} \tag{1.69}$$

where

$$C_u(\alpha, \beta) = \int_0^\infty e^{-\alpha z}\left(1 - e^{-z}\right)^{\beta - 1} dH_u(z), \tag{1.70}$$

with H_u as in Theorem 1.9 if u is not a prior fixed point of discontinuity of Y_t, and $dH_u(z) = (1 - e^{-z})dG_u(z)$ if u is a prior fixed point of discontinuity with distribution function G_u of the jump there. Further details may be found in Ferguson and Phadia (1979).

Characterization of the Neutral to the Right Process As mentioned in Sect. 1.1.2 that if $X_1, X_2, \ldots$ is an exchangeable sequence of random variables defined on $(0, \infty)$, then from a de Finetti theorem, there exists a random distribution function F conditional on which $X_1, X_2, \ldots \overset{iid}{\sim} F$, and a de Finetti measure μ on $\mathcal{F}^+$, known as prior for F, such that for any n the joint distribution of $X_1, X_2, \ldots, X_n$ is

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \int \prod_{i=1}^{n} F(A_i)\mu(dF). \tag{1.71}$$

Walker and Muliere (1999) proved the following characterization of the process neutral to the right. Let $X_1, X_2, \ldots$ be a sequence of random variables with each X_i defined on $(0, \infty)$, such that

$$P(X_{n+1} > t | X_1, \ldots, X_n) = \prod_{0}^{t} \big[1 - dH\big(s, N\{s\}, N^+(s)\big)\big], \tag{1.72}$$

where $H(\cdot)$ is the cumulative hazard function of F, $N\{s\}$ denote the number of X_i equal to s, $N^+(s)$ the number of X_i greater than s, $i = 1, \ldots, n$,

$$\begin{aligned} &dH\big(s, N\{s\}, N^+(s)\big)\big[1 - dH\big(s, N\{s\} + 1, N^+(s)\big)\big] \\ &\quad = dH\big(s, N\{s\}, N^+(s) + 1\big)\big[1 - dH\big(s, N\{s\}, N^+(s)\big)\big] \end{aligned} \tag{1.73}$$

for all $s > 0$, and $\prod_0^t$ represents a product integral. Then the sequence is exchangeable with de Finetti measure a process neutral to the right. Another characterization is also given by Dey et al. (2003).

1.6 Gamma Process

As noted earlier, a neutral to the right process F may be viewed in terms of a process with nonnegative independent increments Y_t via the representation $F(t) = 1 - e^{-Y_t}$. This approach offered many possibilities that have been exploited. When F is continuous, $H(t) = -\log(1 - F(t))$ is the cumulative hazard function. This shows the possibility of utilizing independent increment processes as priors for $H(t)$. Kalbfleisch (1978) is the first one to explore this possibility by using a specific independent increment process, namely, the gamma process to model $H(t)$ in which the increments are assumed to be distributed as gamma distributions, in the context of Bayesian analysis of survival data using the Cox proportional hazard model. Writing a survival function $S(t) = e^{-H(t)}$, it is clear that $H(t)$ can be modeled as an independent increment process. He does so by choosing the gamma process. However, his interest is in analyzing a regression model—the Cox model. It is expressed as $S(t) = 1 - F(t) = \exp\{-H(t)e^{\boldsymbol{\beta}\mathbf{W}}\}$, where $\mathbf{W}$ is a vector of covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients and $\exp\{-H(t)\} = P(T \geq t|\mathbf{W} = \mathbf{0})$ is the baseline distribution. Kalbfleisch's main objective was the estimation of the regression parameter $\boldsymbol{\beta}$, and H was considered as a nuisance parameter. Estimation of $\boldsymbol{\beta}$ proceeded by determining the marginal posterior distribution of data, having H eliminated. He assumed a gamma process as prior for H. The construction of this prior is based on hazard contributions of the intervals obtained by an arbitrary partition of $[0, \infty)$.

1.6.1 Definition

Let $G(\alpha, \beta)$ denote the gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. Let $\alpha(t), t \geq 0$ be an increasing left continuous function such that $\alpha(0) = 0$.

Definition 1.8 Let Z_t, $t \geq 0$ be a stochastic process such that (i) $Z_0 = 0$, (ii) Z_t has independent increments in non-overlapping intervals; and (iii) for $t > s$, the increment $Z_t - Z_s$ is distributed as $G(c(\alpha(t) - \alpha(s)), c)$, where $c > 0$ is a constant. Then Z_t is said to be a gamma process with parameters, $c\alpha(t)$ and c, and denoted as $\mathcal{G}(c\alpha(t), c)$.

It is clear that $\alpha(t)$ is the mean of the process and c is a precision parameter. Sample paths are a.s. increasing. It is a special case of the independent nonnegative increments process with MGF given by

$$\log \mathcal{E}\big[\exp\{-\theta Z_t\}\big] = -\theta b(t) + \int_0^\infty \big(e^{-\theta s} - 1\big) dN_t(s), \tag{1.74}$$

where the Lévy measure has the form $dN_t(s) = \alpha(t)e^{-cs}ds/s$ and $b(t) \equiv 0$. Other properties of the gamma process are well known.

1.6.2 Posterior Distribution

Assume $H \sim \mathcal{G}(cH_0(t), c)$, where H_0 is the prior guess at H. In deriving the posterior distribution we have to be concerned about the prior fixed points of discontinuities inherent in the processes with independent increments. Kalbfleisch got around it by assuming H_0 to be absolutely continuous, in which case there are no prior fixed points of discontinuities. He shows that the posterior distribution of $H(t)$ is again an independent increments process.

His approach is similar to the one described for the neutral to the right process in the previous section. For an arbitrary partition of the real line, $-\infty = t_0 < t_1 < t_2 < \ldots < t_m = \infty$, let q_j denote the hazard contribution of the interval $[t_{j-1}, t_j)$. Then the cumulative hazard function $H(t_i)$ is the sum of hazard rates r_j's, $H(t_i) = \sum_{j=1}^{i} -\log(1-q_j) = \sum_{j=1}^{i} r_j$, or $r_i = H(t_i) - H(t_{i-1})$. Clearly, H is non-decreasing and by assigning independent distributions to q_j's, a prior neutral to the right emerges for F. Let $r_i \sim G(c(H_0(t_i) - H_0(t_{i-1})), c)$, $i = 1, 2, \ldots, m$. Then by this construction, a gamma process emerges as a prior on the space of cumulative hazard functions $H(t)$. It is clear that like the Dirichlet and neutral to the right processes, the gamma process also yields a random H which is discrete with probability one.

Given a sample from F, the posterior distribution is derived by specifying the posterior distribution of the increments, a strategy used by Doksum (1974). Here it is illustrated for the sample size one. Repeated application of this procedure yields the solution for any sample size. Now for an observation $X = x$ such that $x \in [t_{i-1}, t_i)$, r_i is the sum of three independent components: U, the increment to the left of x, J, the jump at x, and V, the increment to the right of x. U and V are gamma variables. The distributions of V and subsequent increments in H remain unchanged. While the distributions of U and all increments prior to x, have gamma distributions with scale parameter changed from c to $c+1$ (or by $c + e^{\boldsymbol{\beta}\mathbf{W}}$ relative to the observation if the regression model is considered). Thus $r_j \sim G(c(H_0(t_j) - H_0(t_{j-1})), c+1)$, $j = 1, \ldots, i-1$, and $U \sim G(c(H_0(x) - H_0(t_{i-1})), c+1)$. The posterior distribution of the jump J turns out to be a distribution with density

$$f_J(s) = \frac{e^{-sc} - e^{-s(c+1)}}{s(\log((c+1)/c))} \tag{1.75}$$

and MGF $M_J(\theta) = \log((c+1-\theta)/(c-\theta))/\log((c+1)/(c))$. Putting together all the independent variables, the posterior distribution of H given $X = x$ is derived. The extension to the sample size n is obvious (see his paper for details). Clearly the distribution, as one would expect, does not have a closed form.

Dykstra and Laud (1981) presents a more general approach and develops an *Extended Gamma Process*, which is described next.

1.7 Extended Gamma Process

Apart from the Gamma process, the prior processes discussed so far were constructed mainly for the purpose of treating a cumulative distribution function (CDF), F. The Dirichlet process and it's variants were constructed on an arbitrary space of probability measures, and tailfree (to be presented in Sect. 1.10) and processes neutral to the right were constructed on the space of distribution functions defined on the real line. These priors are inadequate if one is interested in density functions, or hazard rates which play an equally important role in the study of life history data. In fact, in the context of reliability theory, hazard rates and cumulative hazard rates play a central role. This led Dykstra and Laud (1981) to investigate the problem of placing a prior on the collection of hazard rates. In their treatment, they use processes with independent increments by treating a random hazard rate as a mixture of gamma processes. A by-product of this approach is that it places a prior on absolutely continuous distribution functions instead of discrete distributions. The Bayes estimators derived with respect to this prior under the usual loss function also turn out to be absolutely continuous. These priors are defined on the real line but are not neutral to the right processes and therefore the results of Doksum (1974) and Ferguson and Phadia (1979) are not directly applicable.

1.7.1 Definition

Let F be a left-continuous CDF with $F(x) = 0$ for $x \leq 0$, $S(x) = 1 - F(x)$, $H(x) = -\ln S(x)$. If $r(t)$ is a right continuous function such that $H(x) = \int_{[0,x)} r(t)dt$, then $r(t)$ is known as the hazard rate. Let $\alpha(t)$, $t \geq 0$ be a nondecreasing, left continuous real valued function such that $\alpha(0) = 0$; $\beta(t)$, $t \geq 0$ be a positive, right continuous real valued function bounded away from zero with left-hand limits existing; and finally, let $Z(t)$, $t \geq 0$ be a gamma process with independent increments corresponding to $\alpha(t)$. It is assumed WLOG that this process has nondecreasing, left-continuous sample paths.

Definition 1.9 (Dykstra and Laud) Let $Z(t) \in \mathcal{G}(\alpha(t), 1)$. Then a stochastic process defined by

$$r(t) = \int_{[0,t)} \beta(s)dZ(s) \tag{1.76}$$

is said to be an extended gamma process and denoted by $r(t) \backsim \Gamma(\alpha(\cdot), \beta(\cdot))$.

$\Gamma(\alpha(\cdot), \beta(\cdot))$ is also known as weighted gamma process or a mixture of gamma processes. Obviously, if $r(t)$ is random, then correspondingly, $F(x) = 1 - \exp\{-\int_{[0,x)} r(t)dt\}$ will also be random.

From Doksum (1974), $F(x)$ will be neutral to the right only if $H(x) = \int_{[0,x)} r(t)dt$ has independent increments. Clearly, even though $r(t)$ has independent increments, $H(x)$ will not, and hence the distributional results of Doksum are not applicable.

1.7.2 Properties

1. The MGF $M_{r(t)}(\theta) = \exp\{-\int_{[0,t)} \log(1 - \beta(s)\theta)d\alpha(s)\}$
2. $\mu(r(t)) = \mathcal{E}(r(t)) = \int_{[0,t)} \beta(s)d\alpha(s)$

3. $\sigma^2(r(t)) = \mathrm{Var}(r(t)) = \int_{[0,t)} \beta^2(s)d\alpha(s)$.
4. The marginal and joint survival functions are stated in this theorem.

Theorem 1.10 (Dykstra and Laud) *If the hazard rate $r(t)$ has the prior distribution $\Gamma(\alpha(\cdot), \beta(\cdot))$, then the marginal survival function of an observation X is given by*

$$S(t) = P(X \geq t) = \exp\left\{-\int_{[0,t)} \log\big(1 + \beta(s)(t-s)\big)d\alpha(s)\right\} \tag{1.77}$$

and the joint survival function given n observations $X_1, \ldots, X_n$ is

$$\begin{aligned} S(t_1, \ldots, t_n) &= P(X_1 \geq t_1, \ldots, X_n \geq t_n) \\ &= \exp\left\{-\int_{[0,t)} \log\left(1 + \beta(s)\sum_{i=1}^{n}(t_i - s)^+\right)d\alpha(s)\right\}, \end{aligned} \tag{1.78}$$

where $a^+ = \max(a, 0)$.

It is easy to derive the posterior distribution given $X \geq x$. However, it has the same difficulty as for the neutral to the right processes when $X = x$.

5. In the case of the Dirichlet process, the parameter F_0 was interpreted as prior guess at the unknown F, and M as the concentration parameter or weight attached to the prior guess. Likewise, by defining $\mu(t)$ and $\sigma^2(t)$ as nondecreasing functions, the authors feel it reasonable to interpret $\mu(t)$ as the best guess of the hazard rate and $\sigma^2(t)$ as a measure of uncertainty or variation in the hazard rate at the point t. Then, if μ, σ^2 and α are assumed to be differentiable, $\alpha(\cdot)$ and $\beta(\cdot)$ may be specified suitably in terms of $\mu(\cdot)$ and $\sigma^2(\cdot)$ as follows:

$$\begin{aligned} \beta(t) &= \left(\frac{d\sigma^2(t)}{dt}\right)\Big/\left(\frac{d\mu(t)}{dt}\right) \quad \text{and} \\ \frac{d\alpha(t)}{dt} &= \left[\frac{d\mu(t)}{dt}\right]^2\Big/\left(\frac{d\sigma^2(t)}{dt}\right). \end{aligned} \tag{1.79}$$

1.7.3 Posterior Distribution

The conjugacy property for this prior holds only in the case of right censored data. For the exact observations, the posterior distribution turns out to be a mixture of extended gamma processes.

Theorem 1.11 (Dykstra and Laud) *Let the prior over the hazard rates be $\Gamma(\alpha(\cdot)$, $\beta(\cdot))$. Then the posterior over the hazard rates,*

(i) *given m censored observations of the form* $X_1 \geq x_1, X_2 \geq x_2, \ldots, X_m \geq x_m$, *is* $\Gamma(\alpha(\cdot), \beta^*(\cdot))$ *where*

$$\beta^*(t) = \frac{\beta(t)}{1 + \beta(t) \cdot \sum_{i=1}^m (x_i - t)^+}; \tag{1.80}$$

(ii) *given m observations of the form* $X_1 = x_1, \ldots, X_m = x_m$, *is a mixture of extended gamma processes.*

$$\begin{aligned} &\mathcal{P}\big(r(t) \in B | X_1 = x_1, \ldots, X_m = x_m\big) \\ &\quad = \frac{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \prod_{i=1}^m \beta^*(z_i) \Psi(B; Q) \prod_{i=1}^m d[\alpha + \sum_{j=i+1}^m I_{(x_j,\infty)}](z_i)}{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \prod_{i=1}^m \beta^*(z_i) \prod_{i=1}^m [d\alpha + \sum_{j=i+1}^m I_{(x_j,\infty)}](z_i)}, \end{aligned} \tag{1.81}$$

where $\Psi(B; Q)$ *denotes the probability of the set* $B \in \mathcal{B}$ *under a stochastic process distributed as* $Q = \Gamma(\alpha + \sum_{i=1}^m I_{(x_i,\infty)}, \beta^*)$.

The effect of censored observations is thus to decrease the slope of the sample paths to the left of the censoring points while leaving it unchanged to the right of censoring points.

The posterior distribution with respect to exact observations is somewhat complicated. However, the methods of Kalbfleisch (1978) and Ferguson and Phadia (1979) may be used to express it in terms of MGFs.

Ammann (1984, 1985) generalizes this approach by recasting the hazard rate as a function of the sample paths of nonnegative processes with independent increments which include an increasing component as well as a decreasing component. This way he is able to define a broad class of priors over a space of absolutely continuous distributions that include IFR, DFR and U-shaped failure rate survival functions.

Poisson Point Process In defining the Dirichlet process, Ferguson (1973) was motivated by the fact that the Dirichlet distribution is conjugate with respect to sampling from a multinomial distribution. Lo (1982) recognized that the gamma distribution is conjugate with respect to sampling from a Poisson distribution. So like the Dirichlet process, it should be possible to define a gamma process to solve inference problems for the Poisson point process from a nonparametric Bayesian point of view. Lo showed that this is possible via the weighted gamma process introduced above, and established the following conjugacy property.

Theorem 1.12 (Lo 1982) *If the prior distribution of the intensity measure* γ *of a Poisson point process is* $\Gamma(\alpha, \beta)$, *then given a sample* $N_1, \ldots, N_n$ *of size n from a Poisson point process with intensity measure* γ, *the posterior distribution of* γ *is* $\Gamma(\alpha + \sum_{j=1}^n N_j, \beta/(1 + n\beta))$.

Weighted Distribution Function In the Bayesian analysis of weighted sampling models (where the probability of including an observation in the sample is proportional to a weighting function), Lo (1993b) shows that the normalized weighted gamma process can be used as a conjugate prior for the sampling from a weighted distribution. The weighted distribution is defined as

$$F(dx|G) = \frac{w(x)G(dx)}{\int w(x)G(dx)}, \tag{1.82}$$

where $w(x)$ is a known weight function, $0 < w(x) < \infty$, and G is the unknown parameter. The normalized weighted gamma process is defined as $\gamma(\cdot) = r(\cdot)/r(+\infty)$ and is denoted by $\Gamma^*(\alpha(\cdot), \beta(\cdot))$, where α and β are shape and weight parameters, respectively. Suppose that we have a random sample $X_1, \ldots, X_n|G \overset{iid}{\sim} F(dx|G)$, i.e. the probability of including an observation in the sample is proportional to the weight function w. Then it is shown that if the prior for G is $\Gamma^*(\alpha, 1/w)$, then the posterior distribution of $G|\mathbf{X}$ is $\Gamma^*(\alpha + \sum_1^n \delta_{x_i}, 1/w)$.

1.8 Beta Processes

As noted in the previous two sections, the hazard rates and cumulative hazard rates play an important role in reliability theory. However, it is not easy to place a prior on them. While the neutral to the right processes were flexible, they were unwielding in practical applications. Nevertheless, as discussed in Sect. 1.6, Kalbfleisch (1978) seems to be the first to use a specific independent increment process—the gamma process, as a prior for the cumulative hazard function (leading to a neutral to the right process on $\mathcal{F}$), which he considered as a nuisance parameter in the treatment of covariate analysis, and wanted to average it out. Dykstra and Laud (1981) followed him by treating the hazard rate as a mixture of gamma processes and constructed a prior on the collection of the hazard rates discussed in the last section.

Hjort (1990) follows a different approach. Note that when F is absolutely continuous, $H = -\ln(1 - F)$ is the cumulative hazard function. However, to accommodate the case when F does not have a density, Hjort chooses to deal with the *cumulative hazard rate*, which is defined as the *sum* of hazard rates in the discrete-time framework (and a limit argument in the continuous case), each having an independent beta distribution, and develops a new class of prior processes called *Beta processes*, thereby placing a prior on the space of cumulative hazard rates.

The construction is based on viewing the cumulative hazard rate as a stochastic process. Clearly, it is a nondecreasing process having independent increments. Thus his approach is parallel to that of Doksum (1974), Ferguson (1974) and Ferguson and Phadia (1979) in which the distribution function was reparametrized in terms of the cumulative hazard function via the representation $F(t) = 1 - e^{-Y_t}$, $t \geq 0$, with Y_t being a non-decreasing process with independent increments. This representation facilitated in developing expressions for the posterior distribution of F as well as for the posterior moment generating function of Y_t. Hjort follows a similar path which allows him to deal not only with censored data but more complex models such as Markov Chains (Sect. 3.5.6) and regression models (Sect. 3.7) as well. His focus however is on the cumulative hazard rates and for reasons described below, the formulas of Ferguson and Phadia do not apply directly to his case.

The idea is as follows. Let the hazard rate be denoted by $h(t) = \frac{F'(t)}{F[t,\infty)} = \frac{dF(t)}{F[t,\infty)}$, $t \geq 0$, and the cumulative hazard function $H(t) = \int_0^t h(s)ds$, where $dF(t)$ is an increment of F at t. To permit the definition of cumulative hazard function when F has no density, he uses a more general form of the definition of H, which is valid when F is absolutely continuous as well, as follows.

$$H(t) = \int_{[0,t]} \frac{dF(s)}{F[s,\infty)}, \quad F(t) = 1 - \prod_{[0,t]} \{1 - dH(t)\}, \tag{1.83}$$

where $\prod_{[0,t]}$ denotes the product integral over the interval $[0,t]$. But this creates a problem: the increments of H cannot exceed one, i.e. $0 \leq dH(t) = H(t) - H(t^-) \leq 1$ for all t. This excludes certain independent increments processes (for example, the gamma process whose increments may exceed 1 resulting in F being greater than 1). This suggests that the increments dH should have a distribution defined on the interval [0, 1] and the Lévy measure of the independent increments process restricted to this interval. A natural choice is the beta distribution. But this distribution does not have the additive property and therefore, the distribution of the increments of H is only approximately beta distribution over any finite interval however small the length of the interval might be. All these considerations, compel him to restricts his space of all cumulative hazard rates to a subspace $\mathcal{H}$, which yields F to be a proper distribution function. He then places a prior on $\mathcal{H}$ and proves the existence (his Theorem 3.1) of a nonnegative, independent increments process $H(\cdot)$, the beta process, whose paths a.s. fall in $\mathcal{H}$. (Hjort denotes the members of $\mathcal{H}$ by $A(\cdot)$. However, we will use $H(\cdot)$ instead since we would be focusing only on $\mathcal{H}$.) Thus H as function of F is a mapping from $\mathcal{F}$ to $\mathcal{H}$. A neutral to the right prior on $\mathcal{F}$ induces a similar prior on $\mathcal{H}$. A formal discussion on this relation is given in Ghosh and Ramamoorthi (2003).

It is worth noting the following distinction. Recall that earlier $F(t)$ was expressed as $F(t) = 1 - e^{-Y_t}$, and Y_t was a nondecreasing process with independent increments and with a countable number of fixed points of discontinuities. If the discontinuities are removed, then the process is nondecreasing with independent increments and hence has a simpler Lévy representation which was exploited in Ferguson and Phadia paper. Hjort deals with the Y_t process itself. In this sense the beta process may be viewed as a process leading to a neutral to the right process on $\mathcal{F}$. However, he designates it as $H(t)$ process to reflect the role of the cumulative hazard rate and highlights the distinction. Recall that Kalbfleisch (1978) also assigned a prior to the cumulative hazard function, but he used a gamma process. In the gamma process, the increments are assumed to be independent gamma random variables and in view of the convolution properties of the gamma distribution, it worked well. Furthermore, by assuming the baseline cumulative hazard function, H_0 to be absolutely continuous, Kalbfleisch avoided the problem of prior fixed points of discontinuities. Here, Hjort takes the distribution of the jumps as well as the increments (infinitely small) of the process as independently but approximately beta distributed.

He shows that the beta process has the usual desirable properties: it has broad support, it is flexible, it has the conjugacy property with respect to the censored data, its parameters have natural interpretations, the formulas can be expressed in closed forms and updating the parameters for posterior distribution is easily accomplished. The posterior distribution can be expressed in terms of what happens to the increments, before, after, and at the observation x. A particular transformation of the Dirichlet process lead to a special case of the beta process. In addition, Hjort points out its applications in nonhomogeneous models such as Markov Chain, competing risks and covariate models. Damien et al. (1995, 1996) provide detail steps for implementation of the beta process in practice. We will present here only the time-continuous case. Further details may be found in his comprehensive paper.

By not taking the cumulative form of the Levy measure, Thibaux and Jordan (2007) proposed a hierarchical beta process (see Sect. 1.8.4) for an application to the Indian buffet process.

1.8.1 Definition

Let H_0 be a cumulative hazard with a finite number of jumps taking place at $t_1, t_2, \ldots$ and let $c(\cdot)$ be a piecewise continuous, nonnegative function on $[0, \infty)$. In a time-discrete model, let X be a random variable taking values in $\chi = \{0, 1, 2, \ldots\}$. (This set can be generalized to the set containing $0, b, 2b, \ldots$ for any arbitrary positive constant b.) Then $h(x) = P\{X = x/X \geq x\}$ for $x = 0, 1, 2, \ldots$ and $H(x) = \sum_{s=0}^{x} h(s)$. Thus, $h(x) = H(x) - H(x^-)$ represents an increment in $H(t)$ (and h_0 in H_0) at $t = x$. Now let

$$h(x) \sim Be\big\{c(x)h_0(x), c(x)\big(1 - h_0(x)\big)\big\}. \tag{1.84}$$

In the time-continuous case, with $dH(x)$ representing an infinitesimal increment in $H(x)$ as well as a jump at t_j, $j \geq 1$, let

$$dH(x) \sim Be\big\{c(x)dH_0(x), c(x)\big(1 - dH_0(x)\big)\big\}. \tag{1.85}$$

These two cases lead to the definition of $H(t)$ viewed as a process with independent increments and having a specific Lévy representation. The advantage now is that $\mathcal{E}(h(x)) = h_0(x) = dH_0(x)$, and $\mathcal{E}(H(x)) = H_0(x)$, the prior guesses of $h(x)$ and $H(x)$, respectively, and $\mathrm{Var}(h(x)) = h_0(x)(1 - h_0(x))/[c(x) + 1]$ as the prior 'uncertainty'. A formal definition is as follows.

Definition 1.10 (Hjort) A process H with independent nonnegative increments is a *beta process* with parameters $c(\cdot)$ and $H_0(\cdot)$, symbolically,

$$H \sim \mathcal{B}e\big\{c(\cdot), H_0(\cdot)\big\}, \tag{1.86}$$

if the following holds: For $t \geq 0$, $\theta \geq 0$, H has a Lévy representation with MGF given by

$$M_t(t) = \log \mathcal{E}\big(e^{-\theta H(t)}\big) = \sum_{t_j \leq t} \log \mathcal{E}\big(e^{-\theta S_j}\big) - \int_0^1 \big(1 - e^{-s\theta}\big) dL_t(s), \tag{1.87}$$

where $S_j = H\{t_j\} = H(t_j) - H(t_j^-) \sim \mathcal{B}e\{c(t_j)H_0\{t_j\}, c(t_j)(1 - H_0\{t_j\})\}$, and $\{L_t; t \geq 0\}$ is a continuous Lévy measure having the form

$$dL_t(s) = \int_0^t c(z)s^{-1}(1-s)^{c(z)-1} dH_{0,c}(z)ds \quad \text{for } t \geq 0, \text{ and } 0 < s < 1, \tag{1.88}$$

where $H_{0,c}(t) = H_0(t) - \sum_{t_j \leq t} H_0\{t_j\}$ is H_0 with its jumps removed.

Here, H_0 can be interpreted as a prior guess at the cumulative hazard and $c(t)$ as a measure of strength in the prior guess (playing the role, respectively, of $F_0 = \overline{\alpha}$ and M in the Dirichlet process). Thus by definition, the beta process has independent increments and at fixed points of discontinuity, each increment has a beta distribution. The Lévy measure is concentrated on the interval [0, 1] instead of the interval $[0, \infty)$.

1.8.2 Properties

1. $\mathcal{E}[H(t)] = \sum_{t_j \leq t} \mathcal{E}S_j + H_{0,c}(t) = H_0(t)$.
2. $\text{var}[H(t)] = \sum_{t_j \leq t} \text{Var}\, S_j + \int_0^t \frac{dH_{0,c}(s)}{c(s)+1} = \int_0^t \frac{dH_0(s)(1-dH_0(s))}{c(s)+1}$.
   In Ferguson and Phadia Y_t was expressed as $Y_t = -\log(1 - F(t))$ and it was shown that the property of nonnegative independent increments is preserved passing from prior to posterior distribution. Here instead $H(t) = -\log(1 - F(t))$. Indeed there is a connection between the two. $H(t)$ is a nonnegative independent increment process if and only if Y_t is. Hence the property of independent increments is preserved in H as well passing from prior to posterior distributions. However, the formulas turn out to be different.
3. A prior for the distribution function is neutral to the right if and only if the corresponding cumulative hazard rate is an independent nonnegative increment process with Lévy measure concentrated on [0, 1].
4. The conjugacy property also holds for the beta process.

Theorem 1.13 (Hjort) *Let $H \sim \mathcal{B}e\{c(\cdot), H_0(\cdot)\}$ as defined above. Then, given a random sample which may include right censored observations, the posterior distribution is given by*

$$H|data \sim \mathcal{B}e\left\{c(\cdot) + R(\cdot), \int_0^{(\cdot)} \frac{c(s)dH_0(s) + dN(s)}{c(s) + R(s)}\right\} \tag{1.89}$$

where $R(t) = \sum_{i=1}^{n} I[X_i \geq t]$, the number of observations available at time t^- and $N(t)$ stands for the number of uncensored observations less than or equal to t.

As was the case with the processes neutral to the right, the posterior process contains fixed points of discontinuities at uncensored points even though the prior may not.

The posterior distribution of a jump at t is

$$H\{t\}|data \sim \mathcal{B}e\big\{c(t)H_0\{t\} + dN(t), c(t)\big(1 - H_0\{t\}\big) + R(t) - dN(t)\big\}. \quad (1.90)$$

Therefore, in describing the posterior distribution care must be taken. Hjort gives the description in his Theorem 4.1 which is reproduced in the next subsection, and is similar to Ferguson (1974) and Ferguson and Phadia's (1979) theorems.

5. Muliere and Walker (1997) have shown that the beta process may also be viewed as a Polya tree process (see Sect. 1.11).

1.8.3 Posterior Distribution

In stating the description of the posterior distribution, Hjort considers a slightly more general case. He assumes as prior a process with independent nonnegative increments with fixed points of discontinuities $\{t_1, \ldots, t_k\}$, and the jumps S_j at t_j having density f_j, $j = 1, \ldots, k$. When these discontinuities are removed, the resulting process will have Lévy measure $dL_t(s) = \int_0^t K(s, z)dG(z)ds$ on $0 < s < 1$, $t \geq 0$ where G is a continuous nondecreasing function with $G(0) = 0$, and $K(s, z)$ is some continuous nonnegative function such that $\int_0^1 s dL_t(s) < \infty$. In the case of beta process priors, f_j's are beta densities, $K(s, z) = c(z)s^{-1}(1 - s)^{c(z)-1}$, and $G = H_0$.

Theorem 1.14 (Hjort) *Given H, let X be a sample of size one from the corresponding distribution function F, and H be a nonnegative independent increment process with parameters consisting of a set $t_1 < \ldots < t_k$ of prior fixed points of discontinuities with jumps S_j at t_j having density f_j, $j = 1, \ldots, k$. Then the posterior is again a nonnegative independent increment process with parameters updated as follows. Here κ is a normalizing constant.*

(i) *Given $X > x$, $f_j(s)$ changes to $f_j^*(s) = \kappa(1 - s)f_j(s)$ if $t_j \leq x$, $= f_j(s)$ if $t_j > x$; and $K(s, z)$ gets multiplied by $(1 - s)$ for only $z \leq x$.*

(ii) *Given $X = x$, and $x = t_i$ for some i, $f_j(s)$ changes to $f_j^*(s) = \kappa(1 - s)f_j(s)$ if $t_j < x$, $= \kappa s f_i(s)$, if $t_j = x$, and $= f_j(s)$ if $t_j > x$; and $K(s, z)$ gets multiplied by $(1 - s)$ for only $z \leq x$.*

(iii) *Given $X = x$, and $x \neq t_i$ for any i, $f_j(s)$ changes to $f_j^*(s) = \kappa(1 - s)f_j(s)$ if $t_j < x$, $= f_j(s)$ if $t_j > x$; and an additional point of discontinuity is added to the set $\{t_1, \ldots, t_k\}$ at x with density of the jump S at x, $f_x^*(s) = \kappa s K(s, z)$, $0 < s < 1$; and $K(s, z)$ gets multiplied by $(1 - s)$ for only $z \leq x$.*

The general case of size n which may include right censored data, can be handled by repeated application of the above theorem. However, as indicated in (iii) a new point is added for every uncensored observation, say u_r not among t_j's, and hence we need to specify the density of the jumps at these new points of discontinuity (assuming no fixed points of discontinuity to start with). This is done and stated in Hjort's Theorem 4.2, which resembles Theorem 4 of Ferguson and Phadia (1979). The density of the jump at u_r is given by $f_r^*(s) = \kappa s^{n_r}(1-s)^{m_r} K(s, u_r)$, where n_r is the number of uncensored observations at u_r and m_r is the number of observations greater than u_r. In Ferguson and Phadia, the posterior distribution was given in terms of the MGF and precise formulas were worked out in two specific cases (see Theorem 3.3 in Sect. 3.3.2).

For practical applications in solving Bayesian inference problems, Damien et al. (1996) provide techniques to simulate the posterior distribution, which can then be used to carry out the analysis.

Sinha (1997) and Ibrahim et al. (2001) present the analysis of Cox model based on interval censored data and assuming a discretized beta process model.

Hjort's treatment is extended in two different directions. Kim (1999) allows for more general censoring scheme leading to a multiplicative intensity model (Aalen 1978). On the other hand, James (2006) associates a new variable $x \in \mathcal{X}$ to the cumulative hazard and proposed a class of neutral to the right processes called spatial neutral to the right processes on $R^+ \times \mathcal{X}$ (see Sect. 1.5.1).

1.8.4 Hierarchical Beta Process

In connection with the Indian Buffet Process (to be discussed in Sect. 1.14) which provides a mechanism of defining a prior on the space of sparse binary matrices encountered in document classification applications, a different form of the beta process called the *Hierarchical Beta process* is introduced by Thibaux and Jordan (2007). Note that in the above development, the focus was on the cumulative integral of the sample realizations of the process, where as in applications to the Indian Buffet Process, the attention is on the realizations themselves. Thus the probability space need not be restricted to the real line only. It is essentially a discrete random probability distribution. Let $(\Omega, \sigma(\Omega))$ be a probability space. A positive random measure Q defined on Ω is said to be an independent increment process if for any disjoint sets $A_1, \ldots, A_k$, $k > 1$ of R, $Q(A_1), \ldots, Q(A_k)$ are independent. Then the beta process is redefined as follows.

Definition 1.11 (Thibaux and Jordan) An independent increment process with positive increments defined on $(\Omega, \sigma(\Omega))$ is said to be a beta process B with parameters $c(\cdot)$ and B_0, denoted by $B \sim BP(c, B_0)$ if its Lévy measure for continuous B_0 is given by

$$\nu(d\omega, dp) = c(\omega) p^{-1}(1-p)^{c(\omega)-1} dp B_0(d\omega)$$

on $\Omega \times [0, 1]$, where $c(\cdot)$ is a positive function on Ω, known as the concentration function, and B_0 is a fixed base measure on Ω.

As a function of p, ν is a degenerate beta distribution. It is $dL_t(\cdot)$ without the integral. When B_0 is discrete a slightly different form of the Lévy measure is used.

To sample a random B is to draw a set of points $(\omega_i, p_i) \in \Omega \times [0, 1]$ from a Poisson process with measure ν, which may be represented as an infinite sum $B = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}$, where δ_{ω_i} is a unit point mass at $\omega_i \in \Omega$ and p_i is it's weight. It is a realization of the beta process and is similar to the infinite sum representation of the Dirichlet process. It shows that like the Dirichlet process, B is also discrete. In a similar way a *Bernoulli process* is defined.

Definition 1.12 (Thibaux and Jordan) Let B be a measure on Ω. An independent increment process on $(\Omega, \sigma(\Omega))$ is said to be a Bernoulli process with parameter (hazard measure) B, denoted by $Z \sim BeP(B)$ if its Lévy measure is given by

$$\mu(dp, d\omega) = \delta_1(dp)B(d\omega).$$

If B is continuous, then $Z = \sum_{i=1}^{N} \delta_{\omega_i}$ where $N \sim Poisson(B(\Omega))$ and ω_i's are independently distributed as $B(\cdot)/B(\Omega)$. That is a realization of Bernoulli process is a collection of atoms distributed according to $B(\cdot)/B(\Omega)$ each of unit mass. If B is discrete of the form $B = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}$, then $Z = \sum_{i=1}^{\infty} q_i \delta_{\omega_i}$ where q_i are independent Bernoulli random variables taking value 1 with probability p_i.

The connection of these processes to the Indian Buffet Process will be discussed in Sect. 1.14.

Beta process is shown to have the conjugacy property. Let $Z_1, \ldots, Z_n$ be a set of independent Bernoulli processes, $Z_i|B \sim BeP(B)$, $i = 1, \ldots, n$ and $B|c, B_0 \sim BP(c, B_0)$, then the posterior is $B|\mathbf{Z}, c, B_0 \sim BP(c + n, \frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^{n} Z_i)$ which resembles expression (1.89). It's connection to the Indian Buffet Process and an algorithm to generate beta processes is given in their paper.

The prediction rule based on independent Bernoulli processes drawn from B, and after eliminating B, may be stated as

$$\begin{aligned} Z_{n+1}|\mathbf{Z}_n, c, B_0 &\sim BeP\left(\frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^{n} Z_i\right) \\ &= BeP\left(\frac{c}{c+n} B_0 + \sum_{j=1}^{K} \frac{m_j}{c+n} \delta_{\omega_j}\right), \end{aligned} \tag{1.91}$$

where K is the number of distinct observations among $Z_1, \ldots, Z_n$ with atom ω_j having m_j multiplicities.

The authors also indicate some applications of the beta process in hierarchical modeling problems.

1.9 Beta-Stacy Process

Alternative to the Dirichlet process and processes neutral to the right, Walker and Muliere (1997a) introduced a new stochastic process, called the *beta-Stacy process*, which places a prior on the space $\mathcal{F}$ of all distributions functions defined on $[0, \infty)$ (extension to the whole real line is trivial). It is derived by using an independent nonnegative increment process Z which is taken to be a log-beta process defined below as opposed to the gamma process in Kalbfleisch (1978) and the beta process in Hjort (1990). As in the case of beta process, they also consider in their construction, discrete and continuous cases separately. But unlike the beta process they take the increments to be distributed as *beta-Stacy distribution* with parameters, α and β and with density given by

$$f(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha} \frac{(x-y)^{\beta-1}}{x^{\alpha+\beta-1}} I_{(0,x)}(y) \quad \text{for } 0 < x \leq 1, \tag{1.92}$$

where $B(\alpha, \beta)$ is the usual beta function. If $x = 1$, the density reduces to that of a beta distribution. They give interpretation of the parameters in terms of the mean and variance of F. A key feature is that it gives positive probability to continuous distribution functions.

The beta-Stacy process generalizes the Dirichlet process in two respects: it has a broader support so that more flexible prior information may be represented; and unlike the Dirichlet process, it is conjugate to the right censored data. The parameter of the Dirichlet process M, a positive constant, is replaced by a positive function $c(\cdot)$. Thus a particular case of beta-Stacy process yields the Dirichlet process. Also, when the prior process is assumed to be Dirichlet, the posterior distribution given the right censored data turns out to be a beta-Stacy process. It is worth noting that while the Dirichlet process is defined by the joint distribution of probabilities of sets of any finite measurable partition of $\mathfrak{X}$, the beta-Stacy process is defined on the interval $[0, \infty)$ via the independent nonnegative increment process Z and the representation of the neutral to the right process (as also the Dirichlet process definable on R). Thus the beta-Stacy process belong to the class of neutral to the right processes. It's connection to the beta process is given in property 4. Their result is as follows.

1.9.1 Definition

Let $\alpha(\cdot)$ $(\alpha(0) = 0)$ be a right continuous measure and $\beta(\cdot)$ be a positive function, both defined on $[0, \infty)$. Let $\{t_1, t_2, \ldots\}$ be a countable set of discontinuities of $\alpha(\cdot)$ and define a continuous measure $\alpha_c(t) = \alpha(t) - \sum_{t_j \leq t} \alpha\{t_j\}$. Log-beta process is defined as follows.

Definition 1.13 (Walker and Muliere) A stochastic process Z is a log-beta process on $([0, \infty), \mathcal{B}_+)$, with parameters $\alpha(\cdot)$ and $\beta(\cdot)$, if Z is an independent nonnegative increment process with log-Laplace transform

$$\log \mathcal{E}\big(e^{-\theta Z_t}\big) = \sum_{t_j \le t} \log \mathcal{E}\big(e^{-\theta S_j}\big) - \int_0^\infty \big(1 - e^{-v\theta}\big) dN_t(v) \tag{1.93}$$

where S_j is the size of the jump at t_j, $1 - \exp(-S_j) \backsim Be(\alpha\{t_j\}, \beta(t_j))$, and the Lévy measure is given by,

$$dN_t(v) = \frac{1}{(1 - e^{-v})} \int_0^t \exp\big(-v\big(\beta(s) + \alpha\{s\}\big)\big) d\alpha_c(s) dv, \quad v > 0. \tag{1.94}$$

To define the beta-Stacy process in terms of the parameters $c(\cdot)$ and G, G a distribution function to be interpreted as a prior guess of F (similar to F_0 in the Dirichlet process), we require the following.

Let $c(\cdot)$ be a positive function, $G \in \mathcal{F}$ be a right continuous function with a countable set of discontinuities at $\{t_1, t_2, \ldots\}$, and $G_c(t) = G(t) - \sum_{t_j \le t} G\{t_j\}$ so that $G_c(t)$ is continuous. Let Z_t be an independent nonnegative increments process with log Laplace transform

$$M_t(\theta) = \log \mathcal{E}\big(e^{-\theta Z_t}\big) = \sum_{t_j \le t} \log \mathcal{E}\big(e^{-\theta S_j}\big) - \int_0^\infty \big(1 - e^{-v\theta}\big) dN_t(v), \tag{1.95}$$

where $1 - \exp(-S_j) \backsim Be(c(t_j)G\{t_j\}, c(t_j)G[t_j, \infty))$ and the Lévy measure given by, for $v > 0$,

$$dN_t(v) = \frac{1}{(1 - e^{-v})} \int_0^t \exp\big(-vc(s)G(s, \infty)\big) c(s) dG_c(s) dv. \tag{1.96}$$

Then the definition of beta-Stacy process is

Definition 1.14 (Walker and Muliere) Let Z_t be an independent nonnegative increment process as defined above. Then F is a beta-Stacy process on $([0, \infty), \mathcal{B}_+)$ with parameters $c(\cdot)$ and G, denoted by $F \in \mathcal{S}(c(\cdot), G)$, if for all $t \ge 0$, $F(t) = 1 - e^{-Z_t}$.

Walker and Muliere prove the existence of such a process on the line of Hjort's proof for the existence of beta process. Note that in defining the neutral to the right process, F was reparametrized in terms of Y_t, a non-negative independent increment process. Here, essentially, Y_t is taken to be the log-beta process.

The parameters α and β of the log-beta process are related, under certain condition, to the parameters of the beta-Stacy process $c(\cdot)$ and G as follows.

$$G(t) = 1 - \prod_{t_k \le t} \left(1 - \frac{\alpha\{t_k\}}{\beta(t_k) + \alpha\{t_k\}}\right) \exp\left(-\int_0^t \frac{d\alpha_c(s)}{\beta(s) + \alpha\{s\}}\right)$$

and $c(t) = \beta(t)/G[t, \infty)$. α and β can be recovered from $c(\cdot)$ and G via

$$\alpha(t) = \int_0^t c(s) dG_c(s) + \sum_{t_j \le t} c(t_j) G\{t_j\}, \qquad \beta(t) = c(t) G[t, \infty).$$

In the beta process of previous section, the jumps S_j were taken to be beta distributed. Here $1 - \exp(-S_j)$, $j \geq 1$ are taken to be beta distributed. The following properties are highlighted in their paper.

1.9.2 Properties

1. If we take $c(\cdot) = c$ a constant $(= \alpha(R^+))$ and $G(\cdot) = \alpha(\cdot)/\alpha(R^+)$ continuous, then $dN_t(v)$ reduces to

$$\begin{aligned} dN_t(v) &= \frac{dv}{(1-e^{-v})} \int_0^t \exp\bigl(-vcG(s,\infty)\bigr) c\, dG_c(s) \\ &= \frac{e^{-v\alpha(0,\infty)}}{v(1-e^{-v})} \bigl(e^{v\alpha(t,\infty)} - 1\bigr), \end{aligned} \tag{1.97}$$

   which is the Lévy measure for the Dirichlet process (Ferguson 1974) with parameter α, and thus $F(t) = 1 - e^{-Z_t}$ is a Dirichlet process viewed as neutral to the right process. Here the generality is gained by taking the parameter c as a positive function instead of a constant.
2. If we replace $1 - e^{-v}$ by v in the Lévy measure for log-beta process and assume α to be continuous, upon integrating out v, it can be shown that

$$\log \mathcal{E}\bigl(e^{-\theta Z_t}\bigr) = -\int_0^t \log\bigl(1 + \theta/\beta(s)\bigr) d\alpha(s), \tag{1.98}$$

   which characterizes the extended gamma process (Dykstra and Laud 1981).
3. The Lévy measure for the beta process $dL_t(s)$, with support (0, 1), can be obtained via a simple transformation of the Lévy measure of log-beta process, with α assumed to be continuous,

$$dL_t(s) = \frac{1}{1-s} dN_t\bigl(-\log(1-s)\bigr). \tag{1.99}$$

4. If H is a beta process and $dZ = -\log(1 - dH)$, then $F(t) = 1 - e^{-Z_t}$ is a beta-Stacy process.
5. By taking $\theta = 1$ in the MGF, the prior mean can be seen to be

$$\begin{aligned} \mathcal{E}\bigl(F(t)\bigr) &= 1 - \prod_{t_k \leq t} \left(1 - \frac{G\{t_k\}}{G(t_k,\infty)}\right) \exp\left(-\int_0^t dG_c(s)/G(s,\infty)\right) \\ &= 1 - \prod_{[0,t]} \left(1 - \frac{dG(s)}{G(s,\infty)}\right) = G(t). \end{aligned} \tag{1.100}$$

   The second equality is in the product integral notation.

6. The conjugacy property of the beta-Stacy process with respect to the data which may possibly include right censored observations, is stated in the following theorem.

Theorem 1.15 (Walker and Muliere) *Let $X_1, \dots, X_n$ be a random sample, possibly with right censoring, from an unknown distribution function F on $[0, \infty)$ and let $F \backsim \mathcal{S}(c(\cdot), G)$. Then the posterior distribution of F is again a beta-Stacy process with parameter $c^*(\cdot)$ and G^*, where*

$$G^*(t) = 1 - \prod_{[0,t]} \left\{ 1 - \frac{c(s)dG(s) + dN(s)}{c(s)G[s, \infty) + R(s)} \right\}, \tag{1.101}$$

$$c^*(t) = \frac{c(t)G[t, \infty) + R(t) - N\{t\}}{G^*[t, \infty)}, \tag{1.102}$$

and where as before, $N(\cdot)$ is the counting process for uncensored observations and $R(t) = \sum_i^n I[X_i \geq t]$.

This generalizes Susarla and Van Ryzin (1976) result where F was assumed to have a Dirichlet process prior.

7. A similar conjugacy result, parallel to that for the gamma and beta processes holds for the log-beta process as well.

Theorem 1.16 (Walker and Muliere) *Given a sample of size n from F with a log-beta process prior with parameters $\alpha(t)$ and $\beta(t)$, then the posterior distribution is a log-beta process with parameters updated as $\alpha(t) + N(t)$ and $\beta(t) + R(t) - N\{t\}$.*

8. The posterior mean, which is the Bayes estimate of $F(t)$ under the weighted quadratic loss function, is given in Chap. 3 and is the same estimator as obtained by Hjort (1990) for the beta process.

Characterization Recall that for the Dirichlet process we had $\mu(t) = \mathcal{E}[F(t)] = F_0(t)$ and $\mathrm{Var}[F(t)] = F_0(t)(1 - F_0(t))/(M + 1)$ and therefore we may set $\alpha(\cdot) = MF_0(\cdot)$. The neutral to the right process was described in terms of the stochastic process with Lévy measure $N_t(\cdot)$ as indicated earlier. As $\mu(t)$ and $\mathrm{Var}[F(t)]$ characterize the Dirichlet process, Walker and Damien (1998) defines two functions $\mu(t)$ and $\mathrm{Var}[F(t)]$ in terms of the Lévy measure that characterize the neutral to the right process, and more generally, the beta-Stacy process. Note that $S(t) = 1 - F(t) = e^{-Z_t}$, $\mathcal{E}[S(t)] = \mathcal{E}[e^{-Z_t}]$. So with no fixed points of discontinuity, Walker and Damien consider functions

$$\mu(t) = -\log \mathcal{E}\big[S(t)\big] = -\log \mathcal{E}\big[e^{-Z_t}\big] = \int_0^\infty \big(1 - e^{-z}\big) dN_t(z) \tag{1.103}$$

and

$$\lambda(t) = -\log \mathcal{E}\big[S^2(t)\big] = \int_0^\infty \big(1 - e^{-2z}\big) dN_t(z), \tag{1.104}$$

where $N_t(\cdot)$ is a Lévy measure as before. Also, since $(\mathcal{E}[S(t)])^2 < \mathcal{E}[S^2(t)] < \mathcal{E}[S(t)]$, μ and λ satisfy $0 < \mu(t) < \lambda(t) < 2\mu(t)$. Thus to characterize the process, it is required to find N_t satisfying these two equations. They consider Lévy measures of the type

$$dN_t(z) = \left(1 - e^{-z}\right)^{-1} \int_0^t e^{-z\beta(s)} d\alpha(s) dz, \tag{1.105}$$

where $\beta(\cdot)$ is a nonnegative function and $\alpha(\cdot)$ is a finite measure and show that this type of N_t characterizes the beta-Stacy process and covers many neutral to the right type processes. In particular, the Dirichlet process arises when $\beta(t) = \alpha(t, \infty)$, and the simple homogeneous process (Ferguson and Phadia 1979) emerges when β is constant.

They prove the existence of such $\alpha(\cdot)$ and $\beta(\cdot)$ which satisfy

$$\begin{aligned} \mu(t) &= \int_0^\infty \int_0^t e^{-z\beta(s)} d\alpha(s) dz \quad \text{and} \\ \lambda(t) &= \int_0^\infty \int_0^t \frac{(1 - e^{-2z})}{(1 - e^{-z})} e^{-z\beta(s)} d\alpha(s) dz. \end{aligned} \tag{1.106}$$

If β is constant, then μ and λ are related to $\alpha(\cdot)$ and β through $\mu(t) = \alpha(t)/\beta$ and $\lambda(t) = c\alpha(t)/\beta$ where $c = (1 + 2\beta)/(1 + \beta)$. In general, they discuss a meaningful way to choose $\alpha(\cdot)$ and $\beta(\cdot)$. It is important to note that this method allows one to specify the mean and variance for $F(t)$ which in general is not possible for the neutral to the right processes.

1.9.3 Posterior Distribution

On the basis of the above theorems, the authors restate Ferguson's (1974) and Ferguson and Phadia's (1979) theorems for obtaining the posterior distribution of F given the data, when F is assumed a priori to be neutral to the right. It is similar to the one given for the beta process. In restating the posterior distribution they take the Lévy measure of the type $dN_t(z) = (\int_0^t K(z, s) ds) dz$ and note that the beta-Stacy process with parameters $\alpha(\cdot)$ and $\beta(\cdot)$ arises when $K(z, s) ds = e^{-z(\beta(s) + \alpha\{s\})} d\alpha(s)/(1 - e^{-z})$. Let $\{t_1, t_2, \ldots\}$ be the set of fixed points of discontinuities and $f_j(z)$ be the density of jump at t_j. Then

(i) *Given $X > x$, $f_j(z)$ changes to $f_j^*(z) = \kappa e^{-z} f_j(z)$ if $t_j \leq x$, $= f_j(z)$ if $t_j > x$; and $K(z, s)$ gets multiplied by e^{-z} for only $s \leq x$.*

(ii) *Given $X = x$, and $x = t_i$ for some i, $f_j(z)$ changes to $f_j^*(z) = \kappa e^{-z} f_j(z)$ if $t_j < x$, $= k(1 - e^{-z}) f_i(z)$, if $t_j = x$, and $= f_j(z)$ if $t_j > x$; and $K(z, s)$ gets multiplied by e^{-z} for only $s \leq x$.*

(iii) *Given* $X = x$, *and* $x \neq t_i$ *for any* i, $f_j(z)$ *changes to* $f_j^*(z) = \kappa e^{-z} f_j(z)$ *if* $t_j < x$, $= f_j(z)$ *if* $t_j > x$; *and an additional point of discontinuity is added to the set* $\{t_1, t_2, \ldots\}$ *at* x *with density of the jump* at x, $f_x^*(z) = k(1 - e^{-z})K(z, x)$, *and* $K(z, s)$ *gets multiplied by* e^{-z} *for only* $s \leq x$.

This approach would allow one to carry out full Bayesian analysis via simulation. Walker and Damien (1998) give details to carry out such analyzes when the posterior is a beta-Stacy process. They illustrate the method by a numerical example in which they rework Kaplan and Meier (1958) data and compare the results with those of Ferguson and Phadia (1979).

It should be noted in this connection, as was indicated in Sect. 1.5, that modification of the Lévy measure give rise to different stochastic processes which may serve as priors for the unknown distribution function or cumulative hazard function. For example, Z_t is a log-beta process on $(R, \mathcal{B})$ with parameters $\alpha(\cdot)$ and $\beta(\cdot)$ if Z is an independent nonnegative increment process with Lévy measure given by (1.94)

If $\alpha(\cdot)$ is continuous, then G is continuous and is given by $G(t) = 1 - \exp(-\int_0^t d\alpha(s)/\beta(s))$. If $(1 - e^{-v})$ is replaced by v in (1.94) then (1.98) characterizes the extended gamma process (Dykstra and Laud 1981) as mentioned earlier.

1.10 Tailfree Processes

In view of the limitations of the Dirichlet process that it selects a discrete probability distribution with probability one, efforts were focused to discover some alternatives. One of them, the tailfree processes offer some hope. Like the neutral to the right processes, they are also defined on the real line. Earlier attempts for constructing tailfree processes can be traced to Freedman (1963) and Fabius (1973) 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. As mentioned earlier, *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. They are defined on the real line based on a sequence of nested partitions of the real line and the property of independence of variables between partitions. (Like the neutral to the right process, a random probability is defined via the joint distribution of families of independent random variables.) Their support includes absolutely continuous distributions. They are flexible and are particularly useful when it is desired to give greater weights to the regions where it is deemed appropriate, by selecting suitable partitions. They possess the conjugacy property. However, unlike the case of the Dirichlet and other processes, the Bayesian results based on these priors are strongly influenced by the partitions chosen. Furthermore, it is difficult to derive expressions in close forms and lack adequate interpretation of the parameters involved. The Dirichlet process is essentially the only process which is tailfree with respect to every sequence of partitions. Computations of tailfree processes are generally more difficult than those of Dirichlet priors.

1.10.1 Definition

In describing the tailfree processes, we follow Ferguson (1974). Let $\{\pi_m; m = 1, 2, \ldots\}$ be a tree of nested measurable partitions of $(R, \mathcal{B})$; that is $\pi_1, \pi_2, \ldots$ be a sequence of measurable partitions such that π_{m+1} is a refinement of π_m for each m, and $\bigcup_0^\infty \pi_m$ generates $\mathcal{B}$. Simplest form of partitions are when π_{m+1} is obtained by splitting each set of the partition π_m into two pieces.

Definition 1.15 (Ferguson) The distribution of a random probability P on $(R, \mathcal{B})$ is said to be tailfree with respect to $\{\pi_m\}$ if $\exists$ a family of nonnegative random variables $\{V_{m,B}; m = 1, 2, \ldots, B \in \pi_m\}$ such that

(1) the families $\{V_{1,B}; B \in \pi_1\}, \{V_{2,B}; B \in \pi_2\}, \ldots$ are independent, and
(2) for every $m = 1, 2, \ldots$, if $B_j \in \pi_j$, $j = 1, 2, \ldots, m$ is such that $B_1 \supset B_2 \supset \ldots \supset B_m$, then $P(B_m) = \prod_{j=1}^m V_{j,B_j}$.

The simplest tailfree process is when the sequence of partitions is defined by splitting randomly the right most set at each level and then defining the random variables $V_{m,B}$'s by the conditional probability $P(B_{mj}|B_{m-1(mj)})$ where $B_{m-1(mj)}$ refers to the set in π_{m-1} containing B_{mj}. The sequence may be constructed by a stick-breaking method. Let $\theta_1, \theta_2, \ldots$ be iid uniform random variables defined on $(0, 1)$. Take a stick of unit length and cut a piece of length $p_1 = \theta_1$. Of the remaining length $1 - p_1$, cut a piece of length $p_2 = \theta_2(1 - p_1)$. Continue this process. Here $\pi_0 = (0, 1]$, $\pi_1 = \{(0, p_1], (p_1, 1]\}$, $\pi_2 = \{(0, p_1], (p_1, p_2], (p_2, 1]\} \ldots$. Another important special case of tailfree processes on $(0, 1]$, is when the partitions points are taken as the dyadic rationals.

1.10.2 The Dyadic Tailfree Process

Ferguson (1974) constructed a dyadic tailfree process on the interval $(0, 1]$ relative to the sequence $\{\pi_m\}$ where π_m is the set of all dyadic intervals of length $1/2^m$, $\pi_m = \{((i-1)/2^m, i/2^m]; i = 1, \ldots, 2^m\}, m = 1, 2, \ldots$. The intervals are expressed in binary notations as follows. Let $.\epsilon_1\epsilon_2 \ldots \epsilon_m$ denote the binary expansion of the dyadic rational $\sum_{j=1}^m \epsilon_j 2^{-j}$, where each ϵ_j is zero or one. Thus $B_0 = (0, 1/2]$, $B_1 = (1/2, 1]$, $B_{00} = (0, 1/4]$, $B_{01} = (1/4, 1/2]$, etc. A set $B \in \pi_m$ is of the form $B_{\epsilon_1\epsilon_2\ldots\epsilon_m} = (.\epsilon_1\epsilon_2 \ldots \epsilon_m, .\epsilon_1\epsilon_2 \ldots \epsilon_m + 2^{-m}]$. The random probability P is defined via the joint distribution of all the random variables, $P(B_{\epsilon_1\epsilon_2\ldots\epsilon_m})$. Let $Y, 0 \leq Y \leq 1$ denote $P(B_0)$ and $1 - Y$ denote $P(B_1)$, that is, Y and $1 - Y$ are the probabilities of events $X \in B_0$ and $X \in B_1$, respectively. Next let $Y_0 \in [0, 1]$ denote the conditional probability $P(B_{00}|B_0)$, $Y_1 = P(B_{10}|B_1)$, so that $P(B_{00}) = YY_0$ and $P(B_{01}) = Y(1 - Y_0)$, $P(B_{10}) = (1 - Y)Y_1$, etc. Following this pattern, use $Y_{\epsilon_1\epsilon_2\ldots\epsilon_m}$ with $\epsilon_m = 0$ for $V_{m,B}$, and $1 - Y_{\epsilon_1\epsilon_2\ldots\epsilon_m}$ with $\epsilon_m = 1$ for $V_{m,B}$ for a set $B \in \pi_m$.

Then $P(B)$ is the product of all the variables associated with the path in the tree leading to B from $(0, 1]$. Thus

$$P(B) = \left(\prod_{j=1,\epsilon_j=0}^{m} Y_{\epsilon_1\epsilon_2\ldots\epsilon_{j-1}} \right) \left(\prod_{j=1,\epsilon_j=1}^{m} (1 - Y_{\epsilon_1\epsilon_2\ldots\epsilon_{j-1}}) \right), \tag{1.107}$$

where for $j = 1$, $Y_{\epsilon_1\epsilon_2\ldots\epsilon_{j-1}}$ stands for Y. For example $P((\frac{1}{2}, \frac{5}{8}]) = (1 - Y)Y_1Y_{10}$. The Y random variables are taken so that they are independent between different layers of partitions. Thus the choice of distributions of Y's should be such that $P(B_{\epsilon_1\epsilon_2\ldots\epsilon_m \underbrace{0000}_{j \text{ terms}}\ldots}) = P(B_{\epsilon_1\epsilon_2\ldots\epsilon_m}) \prod_j Y_{\epsilon_1\epsilon_2\ldots\epsilon_m \underbrace{0000}_{j \text{ terms}}\ldots} \xrightarrow{\text{a.s.}} 0$.

When P is extended to be defined over the algebra of sets generated by the dyadic intervals, P will be σ-additive. Finally, it is extended in the usual manner to a unique probability defined on the class of Borel sets on $(0, 1]$. The distribution of P will be tailfree with respect to the sequence of partitions, $\{\pi_m\}$. If *all* the Y's are chosen mutually independent with $Y_{\epsilon_1\epsilon_2\ldots\epsilon_m} \sim Be(\alpha_{\epsilon_1\epsilon_2\ldots\epsilon_m 0}, \alpha_{\epsilon_1\epsilon_2\ldots\epsilon_m 1})$, for some suitable nonnegative real numbers α's, the process yields a Polya tree process, discussed in the next section. The bivariate extension is presented in Sect. 1.16.

1.10.3 Properties

1. The Dirichlet process is tailfree with respect to every sequence of nested measurable partitions (Doksum 1974). This can be seen as follows.

   For each $m = 1, 2, \ldots$ let $\{A_{m1}, \ldots, A_{mk_m}\}$ denote the partition π_m of R such that π_m is a refinement of π_{m-1}. We need to show that for each m, there exists independent family of random variables $\{Z_{1i}; i = 1, 2, \ldots, k_1\}, \ldots, \{Z_{mi}; i = 1, 2, \ldots, k_m\}$, such that the distribution of the vector $(P(A_{m1}), \ldots, P(A_{mk_m}))$ is the same as that of $(\prod_{j=1}^m Z_{j1}, \ldots, \prod_{j=1}^m Z_{jk_m})$, namely, the Dirichlet distribution. But for any i, $i = 1, 2, \ldots, k_m$, $P(A_{mi})$ has a $Be(\alpha(A_{mi}), \alpha(R) - \alpha(A_{mi}))$. Therefore we have to show that there exist random variables Z's such that $\prod_{j=1}^m Z_{ji}$ is also distributed as $Be(\alpha(A_{mi}), \alpha(R) - \alpha(A_{mi}))$. For this purpose, we define $Z_{ji} = P(A_{ji}|A_{j-1(ji)})$ as independent beta distributed random variables with parameter $(\alpha(A_{ji}), \alpha(A_{j-1(ji)}) - \alpha(A_{ji}))$, for $j = 1, 2, \ldots, m$, where $A_{j-1(ji)}$ is some set of the partition π_{j-1} which contains A_{ji} of π_j and $A_0 = R$. Now taking the product of these variables, and using the properties of beta random variables, it can be seen that $\prod_{j=1}^m Z_{ji} \sim Be(\alpha(A_{mi}), \alpha(R) - \alpha(A_{mi}))$ as was to be shown.
2. The Dirichlet process is the only tailfree process, except for the three trivial processes belonging to class $\mathcal{C}$ of Sect. 1.2, that is independent of the defining partitions (Doksum 1974).
3. Tailfree processes are particularly useful when it is desired to give greater weights to the regions where it is deemed appropriate, by selecting suitable partitions in the construction of the prior.

4. Besides being difficult in applications, the main drawback is that the points of subdivision play a strong role in the posterior distributions. Thus the behavior of estimates will depend upon the type of partition used in describing the process.
5. It is easy to check that a distribution function F is neutral to the right if and only if F is tailfree with respect to every sequence of partitions $\pi_m : \{-\infty < t_1 < \ldots < t_m < \infty\}$ such that π_{m+1} is obtained from π_m by splitting the right most interval (t_m, ∞) into two pieces $(t_m, t_{m+1}]$ and (t_{m+1}, ∞).
6. Tailfree processes are conjugate.

Theorem 1.17 *If the distribution of P is tailfree with respect to the sequence of partition $\{\pi_m\}$, and if $X_1, \ldots, X_n$ is a sample from P, then the posterior distribution of P given $X_1, \ldots, X_n$ is also tailfree with respect to $\{\pi_m\}$.*

The posterior distribution of the V's given $X_1, \ldots, X_n$ can easily be calculated.

1.11 Polya Tree Processes

The Polya tree is a tailfree process in which all (not just between partitions) variables are assumed to be independent. The original idea was contained in Ferguson (1974), but it was Lavine (1992, 1994) who defined it formally and studied in detail to serve as a prior for an unknown distribution function. Since the Dirichlet process is also a tailfree process, the Polya tree process may be considered as an intermediate between the Dirichlet process and a tailfree process. It has advantage over the Dirichlet process since, with proper choice of parameters, it can select continuous and absolutely continuous distributions with probability one. Thus unlike the Dirichlet process, it could serve as a potential candidate to put priors over density functions. On the other hand, it has advantage over tailfree processes since it provides a greater tractability. It also has the conjugacy property with respect to the right censored data which is not true for the Dirichlet process. With enlarged base, it is a generalization of the Dirichlet process (albeit on the real line) as prior and thus has potential to replace the Dirichlet process as prior in various applications. Mauldin et al. (1992) considered its multivariate analog and instead of beta they deal with the Dirichlet distribution. We will however limit ourselves to Lavine's formulation. Similar to the Dirichlet process, Lavine also defines *mixtures* of Polya trees for certain applications, and *finite* or *partially specified* Polya trees for computational feasibility. Some recent activity of their use in modeling data is mentioned in the end.

1.11.1 Definition

Let $E = \{0, 1\}$, $E^0 = \emptyset$, E^m be the m-fold product $E \times \ldots \times E$, $E^* = \bigcup_0^\infty E^m$ and E^N be the set of infinite sequences of elements of E. Let $\mathfrak{X}$ be a separable measurable space, $\pi_0 = \mathfrak{X}$ and $\Pi = \{\pi_m; m = 0, 1, \ldots\}$ be a separating binary tree of partitions of $\mathfrak{X}$; that is, let $\pi_0, \pi_1, \ldots$ be a sequence of partitions such that $\bigcup_0^\infty \pi_m$ generates the measurable sets and that every $B \in \pi_{m+1}$ is obtained by splitting some $B' \in \pi_m$ into two pieces. Let $B_\emptyset = \mathfrak{X}$ and, for all $\epsilon = \epsilon_1 \ldots \epsilon_m \in E^*$, let $B_{\epsilon 0}$ and $B_{\epsilon 1}$ be the two pieces into which B_ϵ splits. Degenerate splits are allowed, for example, $B_\epsilon = B_{\epsilon 0} \cup \emptyset$.

Definition (Lavine) A random probability measure P on $\mathfrak{X}$ is said to have a Polya tree distribution, or a Polya tree prior, with parameter (Π, A), written $P \backsim PT(\Pi, A)$, if there exist nonnegative numbers $A = \{\alpha_\epsilon : \epsilon \in E^*\}$ and random variables $Y = \{Y_\epsilon : \epsilon \in E^*\}$ such that the following holds:

(i) all the random variables in Y are independent;
(ii) for every $\epsilon \in E^*$, Y_ϵ has a Beta distribution with parameters $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$;
(iii) for every $m = 1, 2, \ldots$ and every $\epsilon \in E^m$,

$$P(B_{\epsilon_1 \ldots \epsilon_m}) = \left(\prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_1 \ldots \epsilon_{j-1}}\right)\left(\prod_{j=1;\epsilon_j=1}^{m} (1 - Y_{\epsilon_1 \ldots \epsilon_{j-1}})\right), \tag{1.108}$$

where the first term in the products i.e. $j = 1$ is interpreted as $Y_\emptyset$ or as $1 - Y_\emptyset$.

Random variables $\theta_1, \theta_2, \ldots$ are said to be a sample from P if, given P, they are iid with distribution P.

The Y_ϵ's have the following interpretation: For any $i = 1, 2, \ldots$, $Y_\emptyset$ and $1 - Y_\emptyset$ are, respectively, the probabilities that $\theta_i \in B_0$ and $\theta_i \in B_1$, and for $\epsilon \neq 0$, Y_ϵ and $1 - Y_\epsilon$ are the conditional probabilities that $\theta_i \in B_{\epsilon 0}$ and $\theta_i \in B_{\epsilon 1}$, respectively, given that $\theta_i \in B_\epsilon$. Polya trees are shown to be conjugate and therefore can easily be updated. The new Polya tree has the same structure but the parameters are altered.

The new updated Polya tree gives the distribution of $P \mid \theta_i$. When θ_i is observed exactly there are infinitely many α_ϵ's to update. If θ_i is observed to be in one of the sets, say, B_δ, there are only finitely many to update.

1.11.2 Properties

Most of these properties, unless specified otherwise, are established by Lavine (1992, 1994).

1. The Dirichlet process is a special case of Polya trees. A Polya tree is a Dirichlet process if, for every $\epsilon \in E^*$, $\alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$. Many of the properties proved for the Dirichlet process may be extended to the Polya tree process as well.
2. Muliere and Walker (1997) show that the Polya tree priors generalize the beta process when viewed at an increment level. For any increment $\Delta s > 0$, let $B_0 = [0, \Delta s)$ and $B_1 = [\Delta s, \infty)$. Now partition B_1 into $B_{10} = [\Delta s, 2\Delta s)$ and $B_{11} = [2\Delta s, \infty)$. Continue partitioning the right partition. For $m > 1$,

let $B_{\epsilon_1\dots\epsilon_m 0} = [m\Delta s, (m+1)\Delta s)$ and $B_{\epsilon_1\dots\epsilon_m 1} = [(m+1)\Delta s, \infty)$ where $\epsilon_i = 1$ for all $i = 1, 2, \dots, m$. Let G be a measure on $\Omega = [0, \infty)$ and set $\alpha_{m-10} = \gamma_{m-1} G(B_{\epsilon_1\dots\epsilon_m 0})$ and $\alpha_{m-11} = \gamma_{m-1} G(B_{\epsilon_1\dots\epsilon_m 1})$, where $\gamma_{m-1} = \gamma((m-1/2)\Delta s)$ for some positive function $\gamma(\cdot)$. Now define a sequence of independent beta random variables $Y_m \sim Be(\alpha_{m-10}, \alpha_{m-11})$, and let $X_m = Y_m \prod_{j=1}^{m-1}(1 - Y_j)$. Finally set $H(0) = 0$ and $H(t) = \sum_{m\Delta s \leq t} Y_m$. Then H can be considered as a beta process (Hjort 1990) with parameter $c(\cdot) = \gamma(\cdot) G[\cdot, \infty)$ and $H_0(\cdot) = \int_0^{(\cdot)} dG(s)/G[s, \infty)$ viewed at an incremental level of Δs (Walker and Muliere 1997b; Muliere and Walker 1997). The corresponding distribution function results from $F(t) = 1 - \prod_{m\Delta s \leq t}(1 - Y_m) = \sum_{j\Delta s \leq t} X_j$.

3. With proper choice of parameters, Polya trees assign probability 1 to the set of all continuous distributions. Kraft (1964), Ferguson (1974) and Mauldin et al. (1992) give sufficient conditions (see the next property) for a random probability measure P to be continuous or absolutely continuous with probability one. This would make Polya trees to be appropriate candidates to place priors on density functions, and since the posterior distributions for this priors are obtained by simply updating the parameters, they would make Bayesian estimation of densities feasible, which was not possible with the Dirichlet prior because of discreteness. Lo (1984) used kernel mixture of the Dirichlet processes to place priors on densities (see Sect. 2.5.4).
4. Polya tree CDF can be made uniformly close to a given CDF with high probability. Also, using a Polya tree its probability density function can be made close to a given density function which is impossible for the Dirichlet process. That is, he shows for a given probability measure Q with density q, for any $\epsilon > 0$ and $\eta \in (0, 1)$, there exists a Polya tree P with density p such that $\mathcal{P}(\text{ess}\,\text{Sup}_\theta \,|\log(p(\theta)/q(\theta))| < \epsilon) > \eta$.
5. It allows one to specify a prior which places greater weight on the subsets of real line where it is deemed appropriate, by selecting suitable partitions accordingly.
6. Polya trees have an advantage that for some sampling situations where the posterior turn out to be a mixtures of Dirichlet processes, Polya trees lead just to a single Polya tree.
7. Recall that in the case of Dirichlet process, the parameter α was a finite measure representing prior guess F_0 at the unknown distribution function F via the relation $\alpha(\cdot) = MF_0(\cdot)$ and $M = \alpha(R)$. Therefore, $\mathcal{E}[F(\cdot)] = F_0(\cdot) = \alpha(\cdot)/\alpha(R)$. Similarly in the case of Polya Tree, we can define a probability measure $\beta = \mathcal{E}[P]$ by defining $\beta(B) = \mathcal{E}[P(B)]$ for any measurable set B. Lavine does this by defining first for any $\epsilon \in E^*$, $\beta(B_\epsilon)$ and then extending it to any measurable set B.

$$\beta(B_\epsilon) = \mathcal{E}\left[\left(\prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_1\dots\epsilon_{j-1}}\right)\left(\prod_{j=1;\epsilon_j=1}^{m} (1 - Y_{\epsilon_1\dots\epsilon_{j-1}})\right)\right]$$

$$= \prod_{j=1;\epsilon_j=0}^{m} \frac{\alpha_{\underline{\epsilon}_{j-1}0}}{\alpha_{\underline{\epsilon}_{j-1}0} + \alpha_{\underline{\epsilon}_{j-1}1}} \prod_{j=1;\epsilon_j=1}^{m} \frac{\alpha_{\underline{\epsilon}_{j-1}1}}{\alpha_{\underline{\epsilon}_{j-1}0} + \alpha_{\underline{\epsilon}_{j-1}1}} \qquad (1.109)$$

defines β on the elements of $\bigcup \pi_m$ and since the latter generates measurable sets, β is thus extended to the measurable sets.

Now $\mathcal{P}[\theta_i \in B] = \mathcal{E}[P[\theta_i \in B]|P] = \mathcal{E}[P[B]] = \beta(B)$. Thus the unconditional distribution of an observation θ_i is given for $\epsilon = \underline{\epsilon}_m \in E^*$, as

$$\mathcal{P}[\theta_i \in B_\epsilon] = \mathcal{E}\big[P[B_\epsilon]\big] = \mathcal{E}\big[P(B_{\epsilon_1})P(B_{\epsilon_1\epsilon_2}|B_{\epsilon_1})\dots P(B_\epsilon|B_{\epsilon_1\dots\epsilon_{m-1}})\big]$$
$$= \frac{\alpha_{\epsilon_1}}{\alpha_0+\alpha_1}\cdots\frac{\alpha_{\underline{\epsilon}_m}}{\alpha_{\underline{\epsilon}_{m-1}0}+\alpha_{\underline{\epsilon}_{m-1}1}}. \tag{1.110}$$

8. A major drawback is that the partition Π plays an essential role in developing inferential procedures (Dirichlet processes are the only tailfree processes in which Π does not) and thus the conclusions are contaminated by the type of partition used. On the other hand, it enables one to design partitions so that the random distribution centers around a desired known distribution.
9. Polya trees are conjugate and so for the posterior distribution we need only to update the parameters which can be done easily. For example, $Y_\emptyset$ is the probability that $\theta \in B_0$ and is a beta random variable. Therefore the conditional distribution of $Y_\emptyset$ given θ, is a beta distribution with one of the parameters of the beta distribution will have increased by one. If $\theta \in B_\epsilon$ for some $\epsilon \in E^*$, the scheme of updating follows the same rule. However, the difference is that if θ is observed exactly, then infinitely many α's get updated. On the other hand if we know only that $\theta \in B_\epsilon$, only finitely many α's need to be updated. To overcome the problem of updating infinitely many parameters, Lavine suggests two possible recourses. One is to take α's large enough ($\geq m^2$) at level m and stop at some level M (say of order $\log_2 n$); other is to consider only finitely many levels of a Polya tree (see below).
10. As noted in the last section, Ferguson (1974) constructed a simple dyadic tailfree process P on the interval (0, 1] with respect to $\{\pi_m\}$ where $\pi_m = \{((i-1)/2^m, i/2^m]; i = 1, \dots, 2^m\}$, $m = 1, 2, \dots$. Let $.\epsilon_1\epsilon_2\dots\epsilon_m$ denote the binary expansion of dyadic rational $\sum_{j=1}^m \epsilon_j.2^{-j}$, where ϵ_j is zero or one. If all the Y random variables defined there are taken to be mutually independent distributed as $Y_{\epsilon_1\dots\epsilon_{m-1}} \sim Be(\alpha_{\epsilon_1\dots\epsilon_{m-1}0}, \alpha_{\epsilon_1\dots\epsilon_{m-1}1})$, for nonnegative real numbers α's, then the posterior distributions of Y variables given a sample of size n have the same structure, with $Y_{\epsilon_1\dots\epsilon_{m-1}} \sim Be(\alpha_{\epsilon_1\dots\epsilon_{m-1}0}+r, \alpha_{\epsilon_1\dots\epsilon_{m-1}1}+n-r)$, where r is the number of observations falling in the interval $(.\epsilon_1\epsilon_2\dots\epsilon_{m-1}0, .\epsilon_1\epsilon_2\dots\epsilon_{m-1}1]$ and $n-r$ falling in the interval $(.\epsilon_1\epsilon_2\dots\epsilon_{m-1}1, .\epsilon_1\epsilon_2\dots\epsilon_{m-1}1+2^{-m}]$. He points out that the choice of α's have the following consequences.

    (a) $\alpha_{\epsilon_1\dots\epsilon_m} = \frac{1}{2^m}$ yields a P which is discrete with probability 1 (Blackwell 1973)
    (b) $\alpha_{\epsilon_1\dots\epsilon_m} = 1$ yields a P which is continuous singular with probability 1 (Dubins and Freedman 1966)
    (c) $\alpha_{\epsilon_1\dots\epsilon_m} = m^2$ yields a P which is absolutely continuous with probability 1 (Kraft 1964).

Thus the selections of α's have consequences. They affect the rate at which the updated predictive distribution moves from the prior distribution to the sample distribution, and how closely the distribution of P is concentrated about its mean (see below).

11. The predictive probability is easy to compute (Muliere and Walker 1997). Suppose we are given the data $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, then

$$\mathcal{P}[\theta_{n+1} \in B_{\underline{\epsilon}_m} | data] = \frac{\alpha_{\epsilon_1} + n_{\epsilon_1}}{\alpha_0 + \alpha_1 + n} \cdots \frac{\alpha_{\underline{\epsilon}_m} + n_{\underline{\epsilon}_m}}{\alpha_{\underline{\epsilon}_{m-1}0} + \alpha_{\underline{\epsilon}_{m-1}1} + n_{\underline{\epsilon}_{m-1}}}, \tag{1.111}$$

where n_ϵ is the number of observations among θ's in B_ϵ.

12. Drăghici and Ramamoorthi (2000) give conditions for the prior and posterior Polya tree processes to be mutually continuous, and mutually singular.

13. *Construction of a Polya Tree*: If a prior guess of β, say continuous β_0 is available, Lavine gives a construction of Polya tree such that $\theta_1 \sim \beta_0$. Choose B_0 and B_1 such that $\beta_0(B_0) = \beta_0(B_1) = \frac{1}{2}$. Then for every $\epsilon \in E^*$, choose $B_{\epsilon 0}$ and $B_{\epsilon 1}$ to satisfy $\beta_0(B_{\epsilon 0} \mid B_\epsilon) = \beta_0(B_{\epsilon 1} \mid B_\epsilon) = \frac{1}{2}$, and so on.

   If $\mathfrak{X} = R$ and β_0 has a CDF G, the elements of π_m can be taken to be the intervals $\{G^{-1}(k/2^m), G^{-1}((k+1)/2^m)\}$ for $k = 0, \ldots, 2^m - 1$.

   In the case of censored observations, suppose we know only $\Theta_1 > \theta_1, \ldots, \Theta_k > \theta_k$. WLOG assume $\theta_1 < \theta_2 < \ldots < \theta_k$. Then Π may chosen so that $B_1 = (\theta_1, \infty)$, $B_{11} = (\theta_2, \infty)$, $\ldots$, $B_{11\ldots1} = (\theta_k, \infty)$. Then $P \mid data \sim PT(\Pi, A^*)$ where now $\alpha_1^* = \alpha_1 + k$, $\alpha_{11}^* = \alpha_{11} + k - 1, \ldots, \alpha_{11\ldots1}^* = \alpha_{11\ldots1} + 1$. Although, the posterior distribution is a single Polya tree, it is a mixture of Dirichlet processes. He provides some guidance on the choice of α's.

   α's are chosen according to how quickly the updated predictive distribution moves away from the prior predictive distribution. If α's are large it will be closed to the prior and if α's are small it will be close to sample distribution, a behavior found in the Dirichlet process. If $\alpha_{\epsilon 0} = \alpha_{\epsilon 1}$, then the beta distribution is symmetric. Large values of $\alpha_{\epsilon 0} = \alpha_{\epsilon 1}$ will make P smooth as noted in Ferguson (1974). The choice of α's also governs how closely the distribution of P is concentrated around its mean. See Lavine (1992, 1994) for further discussion.

   Since for exact observations we have to update infinitely many α's, Polya trees may not be suitable unless finite Polya trees are used. For the right censored data we have more choices—Polya tree priors, along with beta, beta-Stacy and neutral to the right processes are preferable over the Dirichlet process. However, the construction of Polya tree priors where the partitions are based on observed data should be a cause for concern.

   Lavine provides examples of calculations involved in using Polya tree priors in place of the Dirichlet priors. He describes the posterior distribution when Polya trees are used to model the errors in regression models $Y_i = \phi(\mathbf{X}_i, \beta) + \epsilon_i$, where $\mathbf{X}_i$ is a known vector of covariates, β is an unknown vector of parameters and ϕ is a known function of $\mathbf{X}_i$ and β, and ϵ_i are independent with unknown distribution P, $P \sim PT(\Pi_\beta, A_\beta)$. He also reworks Antoniak's (1974) empirical Bayes problem in which a mixture of Dirichlet processes prior is replaced by a mixture of Polya trees prior, and shows how posteriors can be computed via the Gibbs sampler thus demonstrating advantages of this substitution.

Characterization Walker and Muliere (1997b) give the following characterization. Let r_{kt} for $k = 1, 2, \dots, m$ represent the number of observations in $B_{\epsilon_1 \dots \epsilon_k}$ (where $B_{\epsilon_1 \dots \epsilon_m} = B_{mt}$) given that there are n_j observations in B_{mj} (for $j = 1, \dots, t$).

$F \backsim \mathcal{PT}(\Pi, A)$ if and only if, there exists non-negative numbers $A = (\alpha_0, \alpha_1, \dots)$ such that, for all $m = 1, 2, \dots,$ and $t \in \{1, \dots, 2^m\}$, and non-negative integers $n_1, \dots, n_t$,

$$\begin{aligned}&\mathcal{P}[\theta_{n+1} \in B_{mt} | n_1 \in B_{m1}, \dots, n_t \in B_{mt}] \\ &\quad = \frac{\alpha_{\epsilon_1} + r_{1t}}{\alpha_0 + \alpha_1 + n} \frac{\alpha_{\epsilon_1 \epsilon_2} + r_{2t}}{\alpha_{\epsilon_1 0} + \alpha_{\epsilon_1 1} + r_{1t}} \cdots \frac{\alpha_{\underline{\epsilon}_m} + r_{mt}}{\alpha_{\underline{\epsilon}_{m-1} 0} + \alpha_{\underline{\epsilon}_{m-1} 1} + r_{m-1t}},\end{aligned} \tag{1.112}$$

where $\underline{\epsilon}_m = \epsilon_1 \dots \epsilon_m$ and thus $\alpha_{\epsilon_1 \dots \epsilon_m}$ is written as $\alpha_{\underline{\epsilon}_m}$.

Finite and Mixtures of Polya Trees Just as Antoniak (1974) (see Sect. 1.4) defined mixtures of Dirichlet processes by indexing the parameter of the Dirichlet process α with a variable θ having a parametric distribution, so also the *mixtures of Polya trees* are defined by Lavine.

Definition 1.16 (Lavine) The distribution of a random probability P is said to be a mixture of Polya trees if there exist a random variable U with distribution H, known as the mixing distribution, and for each u, Polya trees with parameters $\{\Pi_u, A_u\}$ such that $P|U = u \sim PT\{\Pi_u, A_u\}$.

Thus for any measurable set $B \in \mathcal{A}$, $\mathcal{P}(P \in B) = \int \mathcal{P}(P \in B|u)H(du)$. Obviously if H is degenerate at a point, then the mixture reduces to a single Polya tree. For the posterior distribution, we not only need to update A_u but also H must be updated. The mixtures of Polya trees produce a smoothening effect and thus the role of partition may not be that critical.

In practical applications such as generating a sample from the Dirichlet process using the infinite sum Sethuraman representation, the truncation was necessary to proceed with the Bayesian analysis (Sect. 1.2, and Ishwaran and James 2001). Similarly, in using the Polya Tree prior, truncation is essential. This is possible by choosing α's appropriately so that they increase rapidly toward the end of the tree. Simplification is achieved by terminating and updating the partition Π up to some level M and the resulting Polya Trees $\mathcal{PT}(\Pi_M, A_M)$ are termed by Lavine as *finite or partially specified* Polya trees (see Lavine for formal definition). Mauldin et al. (1992) also define finite Polya trees. Lavine offers guidance on how this can be done to a desired degree of accuracy. For example, up to level M, define random G according to the Polya tree and there after it may be taken as uniform or use G_0 the base distribution restricted to this set. Hanson and Johnson (2002) recommend M to be of order $\sim \log_2 n$ as a rule of thumb for sample size n.

Since the base of Polya tree priors include absolutely continuous distributions, it is found to be favorable over the Dirichlet process. Consider the general linear model $Z = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\mathbf{X}$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of regression coefficients, and ϵ is the error term. Traditionally the practice is to assume the error term to be distributed as a parametric distribution, typically normal distribution with mean zero. The nonparametric Bayesian approach is to assume the error term having an unknown distribution, and a prior is placed on the unknown distribution (see Antoniak 1974 for example) centered around a base distribution which may be taken as normal with mean zero. There are several papers along this line using different priors.

Walker and Mallick (1997b) use a finite Polya tree prior for the random errors in a hierarchical generalized linear model centered around a known base probability measure (by taking partitions to coincide with the percentiles of the corresponding distribution function) and find this approach to be more appropriate than a parametric approach. They extend this approach to an accelerated failure time model (Walker and Mallick 1999) where the error term is assumed to have a Polya tree prior and show how to implement MCMC procedure with applications to survival data. Procedure to simulate a random probability measure P from $\mathcal{PT}(\Pi, A)$ is also indicated in their paper. This is done by first generating a finite set of beta random variables and defining the random measure P_M by $P(B_{\epsilon_1 \ldots \epsilon_M})$ for each $\epsilon_1 \ldots \epsilon_M$ according to (1.108). Then one of the 2^M sets is picked according to the random weights $P(B_{\epsilon_1 \ldots \epsilon_M})$ and then a uniform random variate is taken from this set. If one of the set chosen happens to be an extreme set, then the random variate is chosen according to the base measure G_0 restricted to this set. α's are chosen such that they increase rapidly down toward level M. See their paper for details.

Hanson and Johnson (2002) argue that in practice it may be difficult to specify a single centering/base distribution G_0. Therefore, they recommend modeling the error term in a linear model as a mixture of Polya trees. A mixture of Polya tree distribution G is specified by allowing parameters of the centering distribution G_0 and/or the family of real numbers α's to be random. That is, $G|U, C \sim PT(\Pi_u, A_c)$, $U \sim f_u(u)$, $C \sim f_C(c)$. They consider mixtures of Polya trees in which the partitions are constructed by a parametric family of probability distributions with variance U. The effect of taking mixtures is to smooth out the partitions of a simple Polya tree. Hanson (2006) further justify the efficiency of using mixtures of Polya trees alternative to using parametric models and provide computational strategies to carry out the analysis and illustrate them by discussing several examples.

There are numerous papers published since 2006 demonstrating the utility of using mixtures of Polya trees just like mixtures of the Dirichlet processes, in modeling regression models and certain types of large and complex data. Computational procedures are demonstrated with real data and efficiency of such methods is discussed. For example, they are used in reliability and survival data analysis (Hanson 2007), and multivariate mixtures of Polya trees are used for modeling ROC data (Hanson et al. 2008). For an introduction and some applications, the reader is referred to a paper by Christensen et al. (2008).

1.12 Ferguson-Sethuraman Processes

In this section, we describe briefly several new processes which have their origin in Ferguson (1973) and Sethuraman (1994) infinite sum mixture representation of the Dirichlet process. Some of them have garnered significant interest in many fields outside the mainstream statistics including, machine learning, ecology, population genetics, document classification, etc. Thus it is safe to say that it has revolutionized nonparametric Bayesian approach to modeling and inference.

Taking a cue from the Ferguson-Sethuraman representation, we consider here a random probability measure P defined on $(\mathcal{X}, \mathcal{A})$ which has the following general form of a countable mixture of unit masses placed at (random) points $\xi_1, \xi_2, \ldots$ with random weights, $p_1, p_2, \ldots,$

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot), \quad p_j \geq 0, j = 1, 2, \ldots, \sum_{j=1}^{\infty} p_j = 1. \tag{1.113}$$

The manner in which the weights are assigned or constructed and locations are defined, determine different processes. A number of processes having this representation or originating from it, in which the weights are constructed using the stick-breaking construction, have received considerable attention in the last two decades. They are termed as *stick-breaking* processes (Ishwaran and James 2001). The phrase, stick-breaking is used in connection with the construction of weights and goes back to several authors cited by Ishwaran and James, and is used to signify the method of construction of the weights. However, to include random discrete probability distributions (ξ_j are not random) and vectors $\mathbf{p} = (p_1, p_2, \ldots)$ of proportions (of species in a population) where the weights need not be constructed using the stick-breaking construction and yet have the above representation, we enlarge the family and call it as *Ferguson-Sethuraman* processes, and save the stick-breaking phrase to indicate the method of construction. This terminology seems to be more appropriate than the *Stick-breaking processes*. Ferguson-Sethuraman processes encompass Ishwaran and James' $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ family and include additional processes such as the Chinese Restaurant and Indian Buffet processes. In the context of completely random measures, P may be viewed here as measure μ which is either μ_1 or μ_2 but not both, where μ_1 represents the part with fixed points of discontinuity and μ_2 represents the part with random points of discontinuities (see Sect. 1.5.3).

The above representation has four essential ingredients: infinite sum; random weights p_j's; (random) locations ξ_i's; and unit masses at ξ_i's. By varying these ingredients, various generalizations have emerged to serve as priors. They include, a class of discrete priors, such as the Dirichlet-multivariate process; multi-parameter priors such as, the beta two-parameter process and two-parameter Poisson-Dirichlet process; covariate processes such as the dependent and spatial Dirichlet processes; and hierarchical and mixture processes such as the Kernel based stick-breaking processes. These and other processes are discussed in this and subsequent sections.

Two immediate generalizations are apparent. First, the sum in the above representation could be truncated at a positive integer N; and second, the weights p_i's may be constructed using distributions that are more general than the one-parameter beta distribution used in the Sethuraman representation. By truncating the sum to a finite number of terms, we get a class of discrete distribution priors (Ongaro and Cattaneo 2004) and it provides a way to approximate the Dirichlet process for implementing MCMC algorithm in generating a sample from it. This in turn has facilitated analysis of complex hierarchical and mixture models in practice.

The weights in Ferguson's infinite sum representation of a random probability (1.3) were constructed using the normalized gamma variables. But they were impossible to use in practice as they involved an infinite sum in the denominator. On the other hand, the Sethuraman representation of the Dirichlet process consists of $p_1 = V_1$, $p_i = V_i \prod_{j=2}^{i-1}(1 - V_j)$, $i \geq 2$, V_i's are independent with each distributed as $Be(1, \alpha(\chi))$. This representation is devoid of the infinite sum in its weights and thus is more amenable to use in practice. This suggests a natural generalization by taking a more flexible distribution of V_i, the beta distribution with parameters a_i and b_i i.e. $V_i \sim Be(a_i, b_i)$, $a_i > 0$, $b_i > 0$, $i = 1, 2, \ldots$. By varying the values of a_i and b_i, Ishwaran and James (2001) have shown that a large class of priors can thus be constructed which includes some well known processes. During the last few years, this remarkable construction has further turned up in some unexpected prior processes in the fields outside statistics. They include the Chinese restaurant process (CRP) and the Indian buffet process (IBP) (Griffiths and Ghahramani 2006).

The third and fourth generalizations involve the iid locations ξ_i's and the unit mass attached to each of them. By making the locations ξ_j depend upon covariates, MacEachern (1999) has shown that it is possible to carry out the Bayesian analysis of regression type models. He calls the resulting process a *Dependent Dirichlet process*. Several authors have further generalized this approach by making ξ_j and/or p_j dependent upon auxiliary information and proposed *spatial Dirichlet process* (Gelfand et al. 2005), *generalized spatial Dirichlet process* (Duan et al. 2007), *order-based dependent Dirichlet process* (Griffin and Steel 2006), *multivariate spatial Dirichlet process* (Reich and Fuentes 2007) and *latent stick-breaking process* (Rodriguez et al. 2010), to name a few. Finally, by replacing the unit mass at ξ_j with a finite nondegenerate measure, Dunson and Park (2008) introduced a *kernel based Dirichlet process* and have shown that it is feasible to carry out the Bayesian analysis of more complex models.

Many of the above generalizations may be considered as special cases of a discrete random probability measure studied by Pitman (1996a) to model species sampling problems in ecology and population genetics, where the data arise from a discrete distribution. His *species sampling model* is defined as

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot) + \left(1 - \sum_{j=1}^{\infty} p_j\right) Q(\cdot), \tag{1.114}$$

where Q is a probability measure corresponding to a non-atomic distribution G, $\xi_j \overset{iid}{\sim} G$, and weights p_j's are constrained by the conditions, $p_j \geq 0$ and $\sum_{j=1}^{\infty} p_j \leq 1$. In the context of population genetics, it is considered that a population consists of an infinite number of species identified by $\eta_1, \eta_2, \ldots$, and p_j represent the proportion of the j-th species in the population. Clearly, if $\sum_{j=1}^{\infty} p_j = 1$, it reduces to the model (1.113).

The weights p_j's have some interesting interpretations. Ferguson's weights were constructed using a gamma process and they were in decreasing order, $p_1 > p_2 > \ldots$. Sethuraman weights were constructed using beta random variables and they need not be in any particular order. On the other hand, if $\mathbf{p} = (p_1, p_2, \ldots)$ is viewed as a random vector of probabilities, the joint distribution of p_i's or of $\mathbf{p}$ determined by V_i's (when $V_i \overset{iid}{\sim} Be(1,\theta)$ and $p_1 = V_1$, $p_i = V_i \prod_{j=2}(1-V_j)$, $i \geq 2$) is known as the GEM distribution named after McCloskey (1965), Engen (1978), and Griffiths (1980) who introduced it in the context of ecology and population genetics (see Johnson et al. 1997). Also, when interpreted as a probability model in ecology, it is known as Engen's (1975) model. The distribution of ranked permutation $\overline{\mathbf{p}} = (p_{(1)}, p_{(2)}, \ldots)$ of $\mathbf{p}$ is the Poisson-Dirichlet distribution (Kingman 1975). Further connection of these weights is discussed below.

Multiplicities in a Sample The underlying distribution being discrete, we would naturally expect ties among an observed sample $\mathbf{X} = (X_1, \ldots, X_n)$ drawn from such a distribution. Let $X_1^*, \ldots, X_{K(n)}^*$ denote the $K(n) = K$ distinct observations among n and let $n_1, \ldots, n_K$ denote the number of ties at those distinct observations, respectively, so that $n_1 + \ldots + n_K = n$. Earlier it was noted that Antoniak (1974) derived a formula which can be expressed as

$$\mathcal{P}(N_1 = n_1, \ldots, N_K = n_K | K, n) = p_{n,K}(n_1, \ldots, n_K), \tag{1.115}$$

where $p_{n,K}$ is a function of $n_1, \ldots, n_K$ and varies with the underlying distribution.

This may also be viewed as partitioning a set of n integers $\{1, 2, \ldots, n\}$ into K cells $C_1, \ldots, C_K$ and $n_i = \text{Card}(C_i) \geq 1$, $i = 1, \ldots, K$. Then $p_{n,K}$ is labeled as an *exchangeable partition probability function* (EPPF) and comes across in many different applications. A significant study of random partitions was published by Pitman (1995). X_i^*'s have different interpretation in different applications. In a sample of n animals, $X_1^*, \ldots, X_{K(n)}^*$ may represent different species and $n_1, \ldots, n_K$ being the number of animals of each type. In a Polya urn scheme, $X_1^*, \ldots, X_{K(n)}^*$ may represent distinct colors of balls and $n_1, \ldots, n_K$ being the number of balls drawn of each color in n draws. In Chinese restaurant and Indian buffet processes, $X_1^*, X_2^*, \ldots$ may represent distinct tables (dishes) in a restaurant and $n_1, n_2, \ldots$ are then the number of patrons sitting at those tables (relishing dishes).

Formulas for $p_{n,K}$ are derived for different distributions. For example, for the Dirichlet process with parameter α, it is (Antoniak 1974)

$$p_{n,K}(n_1, \ldots, n_K | K) = \frac{\alpha^K}{\alpha(\mathfrak{X})^{(n)}} \prod_{i=1}^{K} \Gamma(n_i), \tag{1.116}$$

where $s^{(n)} = \Gamma(s+n)/\Gamma(s) = s(s+1)\ldots(s+n-1)$. Others are stated in the latter sections.

The infinite sum mixture representation provides a general framework to describe numerous processes. It may also be characterized via specification of predictive distribution based on the Polya urn scheme. The manner in which the weights are specified determine the type of Ferguson-Sethuraman process generated. We describe here some of them organized as follows.

a. Discrete and Finite dimensional priors
b. Beta two-parameter process
c. Poisson-Dirichlet processes (to be discussed in the next section)
d. Dependent and Spatial Dirichlet processes
e. Kernel based stick-breaking processes

1.12.1 Discrete and Finite Dimensional Priors

A family of random probability measures having the representation $P(\cdot) = \sum_{j=1}^{N} p_j \delta_{\xi_j}(\cdot)$ is discussed in Ishwaran and James (2001), where the weights are obtained by the stick-breaking construction. As before, p_i's and ξ_i's are all considered to be mutually independent. The integer N can be finite or infinite, and it could be a fixed integer or a random variable having a particular distribution. By assuming $V_i \overset{iid}{\sim} Be(a_i, b_i)$, the authors have shown that it could generate several different processes.

Generalized Dirichlet Distribution A class of finite dimensional priors may be defined by restricting the sum in (1.113) to a finite number of terms $N < \infty$. In this case if the weights are defined by the stick-breaking construction with $p_1 = V_1$, $p_i = V_i \prod_{j=2}^{i-1}(1 - V_j)$, $i = 2, \ldots, N-1$, then V_N will necessarily have to be defined as $V_N = 1$ to guarantee that $\sum_{j=1}^{N} p_j = 1$ a.s. With p_i thus defined, the vector $\mathbf{p}_N = (p_1, \ldots, p_N)$ has a generalized Dirichlet distribution $GD(\mathbf{a}, \mathbf{b})$, with density given by

$$\left(\prod_{i=1}^{N-1} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)}\right) \prod_{j=1}^{N-1} p_j^{a_{j-1}} p_N^{b_{N-1}-1} \times \prod_{j=1}^{N-2} \left(1 - \sum_{k=1}^{j} p_k\right)^{b_j - (a_{j+1} + b_{j+1})}, \tag{1.117}$$

and it constitutes a conjugate prior to multinomial sampling. If $\mathbf{p}_N \sim D(a_1, \ldots, a_N)$, it can also be represented as having a $GD(\mathbf{a}, \mathbf{b})$ with $\mathbf{a} = (a_1, a_2, \ldots, a_{N-1})$ and $\mathbf{b} = (\sum_{k=2}^{N} a_k, \sum_{k=3}^{N} a_k, \ldots, a_N)$. In particular, if the parameters of the Dirichlet distribution are $a_1 = \ldots = a_N = \alpha/N$ for some $\alpha > 0$ (symmetric Dirichlet distribution), then P is called a *finite dimensional Dirichlet* prior with parameter α (and N) (see Ishwaran and James 2001). In the symmetric case, a random probability measure P can also be expressed as $P_N = \sum_{k=1}^{N} (Y_k / \sum_{k=1}^{N} Y_k) \delta_{\xi_j}$, where $Y_k \overset{iid}{\sim} G(\alpha/N, 1)$, a gamma distribution. This symmetric prior has been used by other authors (for example, Kingman 1975 and Patil and Taillie 1977) in constructing prior distributions.

Discrete Prior Distributions The integer N itself may be considered as a discrete random variable having a specific distribution. In this case, Ongaro and Cattaneo (2004) present a unified general class of *discrete prior distributions* Π_d as follows. Let P_0 be a non-atomic probability measure on $(\mathfrak{X}, \mathcal{A})$. A random probability measure P belongs to this class Π_d if it can be represented as $P = \sum_{i=1}^{N} p_i \delta_{\xi_i}$ where p_i, and ξ_i are independent, p_i having a specified distribution, $\xi_i \overset{iid}{\sim} P_0$ and N being an extended value positive integer or a random variable. As a particular case, given N they take the vector $\mathbf{p}_N = (p_1, \ldots, p_N)$ distributed as an arbitrary distribution on the $(N-1)$-dimensional simplex

$$S_N = \left\{ \mathbf{p}_N : \sum_{i=1}^{N} p_i = 1, p_i \geq 0, i = 1, \ldots, N \right\}. \tag{1.118}$$

They prove some of the properties which are similar to the ones that hold for the Dirichlet process. However, it does not satisfy the conjugacy property and therefore to include the posterior distributions, they show how to create an enlarged family embedding the class of priors.

Residual Allocation Models The weights in the finite case, when viewed as a vector of proportions $\mathbf{p} = (p_1, p_2, \ldots, p_N)$, have been found quite useful in ecology (see for example, Patil and Taillie 1977). One particular model involving these weights seems to appear very frequently. It is called the *residual allocation model* (RAM) which is also known as the stick-breaking model. Let $\mathbf{p}_N$ be as above and define the *residual fractions* as follows: $v_1 = p_1$, $v_2 = p_2/(1-p_1)$, $v_3 = p_3/(1-p_1-p_2), \ldots, v_N = p_N/(1-p_1-\ldots-p_{N-1}) = 1$. A random probability $\mathbf{p}_N$ is said to be a RAM if $v_1, \ldots, v_{N-1}$ have independent distributions with $\mathcal{P}(0 < v_i < 1) = 1$, $i = 1, 2, \ldots, N-1$ and $\mathcal{P}(v_N = 1) = 1$. In the case $N = \infty$, it is necessary to have $\mathcal{P}(\lim_{n\to\infty}(1-p_1-p_2-\ldots-p_n) = 0) = 1$. Examples of RAM are (Patil and Taillie 1977):

(1) The symmetric Dirichlet distribution with parameters N and $\alpha > 0$ is an obvious RAM with $v_i \sim Be(\alpha, (N-i)\alpha)$.

(2) Engen's (1975) model is also a RAM with $v_i \overset{iid}{\sim} Be(1, \alpha)$, $\alpha > 0$.

(3) *Size-biased permutation* (to be defined later) of $\mathbf{p}_N$ having a symmetric Dirichlet distribution with parameter α, is another RAM with $v_i \sim Be(\alpha + 1, (N - i)\alpha)$.

Finite sum representation is also used as a tool to approximate the Dirichlet process in carrying out computational algorithms.

Generalization of the Beta Distribution When $N = \infty$, the construction of several processes may be achieved by varying the parameters of the beta distribution used in the stick-breaking construction. Assumption of independent $V_i \sim Be(a_i, b_i)$, $a_i > 0, b_i > 0, i = 1, 2, \ldots$, leads to the following processes which are discussed below. If $a_i = a$ and $b_i = b$, it gives rise to a distribution called the *beta two-parameter distribution* (Ishwaran and Zarepour 2000). On the other hand if we set $a_i = 1 - \alpha$ and $b_i = \theta + i\alpha$, with $0 \leq \alpha < 1, \theta > -\alpha$, it gives rise to a *two-parameter Poisson-Dirichlet* distribution denoted as $PD(\alpha, \theta)$ (Pitman and Yor 1997). By setting $\alpha = 0$ and $\theta = \mu$ in $PD(\alpha, \theta)$, the Dirichlet process $\mathcal{D}(\mu)$ is recovered; $PD(\alpha, 0)$ yields a stable law process; and $PD(0, \theta)$ yields the (*one parameter*) *Poisson-Dirichlet* distribution (Kingman 1975).

1.12.2 Beta Two-Parameter Process

In the Sethuraman representation of the Dirichlet process, the random variables V_i, $i = 1, 2, \ldots$ are taken to be iid $Be(1, \alpha)$. A two-parameter generalization is introduced by Ishwaran and Zarepour (2000) by replacing the distribution $Be(1, \alpha)$ of V_i with $Be(a, b)$, $a > 0$, $b > 0$ for all $i \geq 1$. They call it as *beta two-parameter process*. It is discussed in connection with the approximation of the Dirichlet process by truncating the infinite sum (1.113) to a finite N terms in order to implement MCMC algorithm in fitting certain nonparametric hierarchical and mixture models. They also suggest that this process may be suitable for finite mixture modeling since the number of distinct sample values tend to match with the number of mixture components.

1.12.3 Dependent and Spatial Dirichlet Processes

There are a number of different approaches to incorporate covariates in data analysis. Most popular approach in survival data analysis is to incorporate covariates through the Cox model and its Bayesian analogue was introduced in Kalbfleisch (1978) who assumed a gamma process prior for the baseline distribution and averaged it out in estimating the regression parameters (see Sect. 1.6).

There has been an intense interest in recent years to provide alternate approaches. MacEachern (1999) proposed using the Sethuraman representation to accommodate covariates. Consider the above representation in terms of a distribution, $F(t) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(-\infty, t]$. Since $\sum_{j=1}^{\infty} p_j = 1$ a.s., F is a distribution function a.s. Let the covariate be denoted by $\mathbf{x} \in \chi$, where χ is a subset of k-dimensional Euclidean space R^k. In the single covariate model, i.e. when $k = 1$, replace each ξ_j by $\xi_{j\mathbf{x}}$, and define $F_{\mathbf{x}}(t) = \sum_{j=1}^{\infty} p_j \delta_{\xi_{j\mathbf{x}}}(-\infty, t]$. That is, each ξ_j is replaced by a stochastic process $\xi_{j\mathbf{x}}$, indexed by $\mathbf{x} \in \chi$. Then the collection of $F_{\mathbf{x}}$ form a class of random distribution functions indexed by $\mathbf{x} \in \chi$. In this formulation, the locations of point masses have been indexed by the covariates, but the weights p_j are undisturbed. The Dirichlet process priors on the collection of $F_{\mathbf{x}}$ are now termed as

Dependent Dirichlet processes. The parameters F_0—the base measure, and M—the total mass of the Dirichlet process may or may not depend on the covariate $\mathbf{x}$. If they do, the marginal distribution of $F_{\mathbf{x}}$ will be a Dirichlet process with parameters $M_{\mathbf{x}}$ and $F_{0\mathbf{x}}$. If they do not, they evolve stationary processes. In another paper, MacEachern (1998) also discusses computational procedures and some properties of the Dependent Dirichlet process. Dunson and Park (2008) indicate that when the weights are independent of the covariates, this method has been successfully applied in problems related to the analysis of variance, spatial modeling and time series. Having only the locations depend on $\mathbf{x}$ constitutes the first kind of extension.

The second kind of extension MacEachern suggests is to have the random weights p_j also vary with $\mathbf{x}$. This can be accomplished by replacing the individual variates V_j in Sethuraman representation by stochastic processes $V_{j\mathbf{x}}$, $\mathbf{x} \in \chi$. In doing so, note that the total mass M will have to be allowed to depend on $\mathbf{x}$ as well, since the distribution of V_j involves the parameter M. This will yield a modified parameter M_χ. Further, it can be extended to include multi-dimensional covariates as well.

Gelfand et al. (2005) saw the need to extend the above definition of the Dependent Dirichlet process by allowing the locations $\mathbf{x}$ to be drawn from random surfaces to create a random spatial process. They named it as *Spatial Dirichlet process*. It is essentially a Dirichlet process defined on a space of surfaces. Using their terminology, let $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, $D \subseteq R^d$ be a stochastic process denoted by Y_D and let $\mathbf{s}^{(k)} = (\mathbf{s}_1, \ldots, \mathbf{s}_k)$ be the specific locations in D at which observations are collected in n replications. Thus the data is of form $\{Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_k)\}^T$, $t = 1, \ldots, n$. The Sethuraman representation is given by $G = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}$ where p_j's are as defined earlier and ξ_j are iid F_0, the base distribution and independent of p_j's. Now to model Y's, they replace ξ_j with $\xi_{j,D} = \{\xi_j(\mathbf{s}) : \mathbf{s} \in D\}$. The resulting random process G for Y_D is given by $\sum_{j=1}^{\infty} p_j \delta_{\xi_{j,D}}$. This means that for $\mathbf{s}^{(k)}$ as above, G induces a random probability measure $G^{(k)}$ on the space of distribution functions for the vector $\{Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_k)\}^T$ with $G^{(k)} \sim \mathcal{D}(MG_0^{(k)})$, where $G_0^{(k)}$ is the k-variate distribution for $\{Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_k)\}^T$ induced by G_0. Covariance structure is thus incorporated by pooling information from nearby locations. The distinction highlighted is that in MacEachern (2000), the Dirichlet processes are univariate at each index value ($F_{\mathbf{x}}$). Here, G induces a random distribution $G(Y(\mathbf{s}))$ for each $\mathbf{s}$, hence the set $G_D \equiv \{G(Y(\mathbf{s})) : \mathbf{s} \in D\}$. The advantage here is that the random marginal distribution $G(Y(\mathbf{s}_i))$ of $Y(\mathbf{s}_i)$ given G and $G(Y(\mathbf{s}_j))$ of $Y(\mathbf{s}_j)$ given G are such that the difference between them tend to zero as $\|\mathbf{s}_i - \mathbf{s}_j\| \to 0$. See also Wolpart and Ickstad (1998).

Duan et al. (2007) argue that the above model uses the same set of weights p_j's thus inducing common surface selection for all locations in the collection, which may not be appropriate in certain situation. Therefore they introduce a random distribution for the spatial effects such that the surface selection can vary with the choice of locations. This is done by introducing a generalization of the above model in which p_j's are also made to vary with locations. Thus a random probability measure on the space of surfaces over D is defined such that for any sets $A_1, \ldots, A_k$

of $\mathcal{B}$

$$\mathcal{P}\big(Y(s_1) \in A_1, \ldots, Y(s_k) \in A_k\big)$$
$$= \sum_{i_1=1}^{\infty} \cdots \sum_{i_k=1}^{\infty} p_{i_1,\ldots,i_k} \delta_{\xi^*_{i_1}(s_1)}(A_1) \ldots \delta_{\xi^*_{i_k}(s_k)}(A_k), \tag{1.119}$$

where i_j stands for $i(s_j)$, $j = 1, \ldots, k$ and the weights $\{p_{i_1,\ldots,i_k}\}$, independent of locations, represent the site-specific joint selection probabilities having a distribution defined in infinite-dimensional simplex $\mathcal{S} = \{\mathbf{p} : p_{i_1,\ldots,i_k} \geq 0;$ $\sum_{i_1=1}^{\infty} \cdots \sum_{i_k=1}^{\infty} p_{i_1,\ldots,i_k} = 1\}$.

In contrast, Griffin and Steel (2006) propose models which allow the weights to change over the domain of the covariates by permuting the V_i's, and does not require replication. They define an *order-based stick-breaking prior* on a space D by a sequence $\{a_j, b_j\}_{j=1}^{N}$ (where N is potentially infinite) and centering distribution F_0, by defining a random distribution $F_{\mathbf{x}}$ for $\mathbf{x} \in D$ as $F_{\mathbf{x}} = \sum_{j=1}^{N} p_j(\mathbf{x})\delta_{\xi_j}$, where ξ_j's and V_j's are independent $\xi_j \overset{iid}{\sim} F_0$, and $V_j \sim Be(a_j, b_j)$, $p_j(\mathbf{x}) = V_{\pi_j(\mathbf{x})} \prod_{i<j}(1 - V_{\pi_i(\mathbf{x})})$ and $\pi(\mathbf{x}) = \{\pi_1(\mathbf{x}), \ldots, \pi_{n(\mathbf{x})}(\mathbf{x})\}$ with $n(\mathbf{x}) \leq N$, the ordering at $\mathbf{x}$. This ordering associate with each atom (ξ_j, V_j) a position in the covariate space and the ordering is defined through ranking atoms by their distance from $\mathbf{x}$.

Reich and Fuentes (2007) extend the stick-breaking prior to the multivariate spatial setting and use it to analyze wind field data. They use bivariate normal priors for the location ξ_j, and similar to Griffin and Steel have the weights p_j vary spatially, but instead of permuting V_j's, they introduce a series of kernel functions to allow the masses to change with space. That is they replace $V_j(\mathbf{x})$ in the stick-breaking construction by a kernel $w_j(\mathbf{x})V_j$. This is similar to the approach of Dunson and Park (2008) discussed below.

Rodriguez et al. (2010) propose models for prior distributions on stochastic processes on an index space D with constant marginal distributions. That is, they no longer build different distributions for each possible value of the index space, but instead construct a stochastic process where observations at different locations are dependent but have a common unknown marginal distribution. In other words, instead of having marginal distribution of the process change at each location, they assume a constant marginal across the index space, but allow samples from it to be dependent. They call the resulting process a *latent stick-breaking process*.

All of the above authors provide rationale for the proposed models and describe procedures for carrying out analyzes with practical examples. These extensions offer a great deal of flexibility in modeling data and may be used as components in hierarchical Bayesian models. They have direct use in modeling the residual structure in the traditional linear model.

1.12.4 Kernel Based Stick-Breaking Processes

A different way of incorporating the covariates is through a kernel mixture used by Lo (1984) in density estimation (see Sect. 2.5.4). Dunson and Park (2008) follow this line of extension. In the problem of density estimation, Lo (1984) defines a random density function $f(y) = \int K(y,u)dG(u)$, where $K(y,u)$ is a known kernel and G is taken to be a random distribution function. By assuming $G \sim \mathcal{D}(\alpha)$, he places a prior on the space of density functions. These type of kernel mixture models are dense (in L_1 norm) in the space of absolutely continuous distributions (Lo 1984). Now a covariate x can be accommodated via letting $f(y|x) = \int K(y,u)dG_x(u)$, where G_x is chosen according to the Dependent Dirichlet process discussed above.

Taking a cue from kernel mixture models, a further more complex generalization is proposed by Dunson and Park (2008) in which, in addition to the locations and mixing weights depending on covariate $x \in \chi$, they replace the point mass at ξ_j by a nondegenerate probability measure G_j^*. Consider a sequence of mutually independent random components $\{\xi_j, V_j, G_j^*; j = 1, 2, \ldots\}$, where for each j, $\xi_j \sim H$ is a location parameter with base distribution H, $V_j \sim Be(a_j, b_j)$ defines a probability weight, and $G_j^* \sim Q$ a probability measure, all defined on appropriate spaces. For a bounded kernel $K : R^p \times R^p \to [0,1]$, let $U_{jx}(\xi_j, V_j) = K(x, \xi_j)V_j$ for all $x \in \chi$.

Then they define a *kernel stick-breaking process* as

$$P_x = \sum_{j=1}^{\infty} U_{jx}(\xi_j, V_j) \prod_{i<j} \big(1 - U_{ix}(\xi_i, V_i)\big) G_j^*. \tag{1.120}$$

They note some obvious special cases. For example if the kernel $K(x, \xi_j) = 1$ and $G_j^* = \delta_{\xi_j}$, the process reduces to the one discussed above. This representation may be recognized as a covariate dependent mixture of an infinite sequence of base probability measures with G_j^* located at ξ_j.

In a recent paper, Chung and Dunson (2011) develop what they call the *local Dirichlet process*. The reader may refer to their paper for further information.

1.13 Poisson-Dirichlet Processes

The Dirichlet process was defined as a random probability measure P on quite a general and arbitrary measurable space $\mathfrak{X}$, and Π was the collection of all such measures. It's infinite mixture representation is $P = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}$. In contrast, this section deals with a random discrete probability distribution $\mathbf{p} = (p_1, p_2, \ldots)$ defined on a countably infinite set and let Π denote here the collection of such discrete probability distributions. Π may also be considered as a collection of vectors $\mathbf{p}$ or a set of sequences $(p_1, p_2, \ldots)$ of real numbers subject to restrictions $p_i \geq 0$, $i = 1, 2, \ldots,$ and $\sum_{i=1}^{\infty} p_i = 1$. The interest is in finding the joint distribution of the components of vector $\mathbf{p}$ or their permutations, which arise naturally in many fields of applications. For example, p_i's may be considered as the proportion of species in a population encountered in ecology or in study of abundances of genes in population genetics, and the interest is in their representation in the order of dominance among a sample drawn from the population.

Two types of permutations are encountered. The rank-ordered permutation of $\mathbf{p}$ in which p_i's are arranged in descending order, and is denoted by $\overline{\mathbf{p}} = (p_{(1)}, p_{(2)}, \ldots)$. Recall that the weights in Ferguson's alternative definition (see Sect. 1.2) of a random probability measure P forms such an ordered sequence. The second is a *size-biased* permutation of $\mathbf{p}$ obtained sequentially as follows. Pick an element p_i of $\mathbf{p}$ at random and set $w_1 = p_i = p_{i(1)}$ (i.e. $w_1 = p_i$ with probability p_i). Remove $p_{i(1)}$. Now pick p_i at random from the remaining components and set $w_2 = p_i = p_{i(2)}$ (i.e. $w_2 = p_i I[p_i \neq w_1]$ with probability $p_i I[p_i \neq w_1]/(1 - w_1)$). Continue this way. Then the resulting sequence $(w_1, w_2, \ldots)$ is said to be a *size-biased random permutation* of $\mathbf{p}$.

Kingman (1975) derived the distribution of $\overline{\mathbf{p}}$ involving one single parameter under certain conditions and called it a *Poisson-Dirichlet distribution*. Its extension to the two-parameter case was treated by Pitman and Yor (1997), which is also a two-parameter generalization of the Dirichlet process. The objective of this section is to introduce briefly one and two-parameter Poisson-Dirichlet distributions which may be considered as priors on Π^*, a subset of Π such that $p_1 \geq p_2 \geq \ldots$.

These distributions may also be constructed by the stick-breaking construction and thus are considered as Ferguson-Sethuraman type distributions.

1.13.1 One-Parameter Poisson-Dirichlet Process

Kingman (1975) points out that it is impossible to choose at random a $\mathbf{p}$ which is invariant under permutation of that set. Therefore his approach is to consider first the class of finite dimensional probability distributions $\mathbf{p}_N = (p_1, p_2, \ldots, p_N)$ and then letting N increase indefinitely. He shows that under appropriate conditions, the vector $\mathbf{p}_N$, rearranged in decreasing order, has a non-degenerate limiting distribution involving one parameter and named it as a Poisson-Dirichlet distribution. He gives an interesting motivator—the problem of 'heaps', which may be described as follows.

Suppose we have a heap of N items, $I_1, I_2, \ldots, I_N$ stacked up on the desk. Periodically, we seek an item which is I_k with probability p_k and is searched through the heap, starting at the top. After its use, it is placed back on the top of the heap and a subsequent search starts. All searches are assumed to be independent. The process is repeated, every time the item after use is being placed at the top. Eventually, the system will stabilize and items will have been stacked up in the order of their popularity, the most sought after item will tend to be placed on the top, second most popular item will be placed next, and so on. This is essentially the rearrangement of p_i's in decreasing order.

It also has ecological applications where p_i in the random vectors $\mathbf{p} = (p_1, p_2, \ldots)$ may represent the proportion of i-th species η_i in an infinite population, and presumably, there are unlimited species. It is desired to find the distribution of the random vector $\overline{\mathbf{p}} = (p_{(1)}, p_{(2)}, \ldots)$, where $p_{(1)} \geq p_{(2)} \geq \ldots$ and where $p_{(k)}$ for $k = 1, 2, \ldots$, represents the proportion of k-th dominant species η_k encountered in a sample draw.

Let $\mathbf{p}_N = (p_1, p_2, \ldots, p_N)$ be a vector such that $p_i \geq 0$, $i = 1, 2, \ldots, N$ and $\sum_{i=1}^{N} p_i = 1$. The usual prior distribution taken for $\mathbf{p}_N$ is $D(\alpha_1, \alpha_2, \ldots, \alpha_N)$, the Dirichlet distribution with parameter vector $(\alpha_1, \alpha_2, \ldots, \alpha_N)$, on the $(N-1)$-dimensional simplex. However, Kingman (1975) notes that if in particular all α's are the same and equal to α, then p_i's have an exchangeable symmetric distribution $D(\alpha, \alpha, \ldots, \alpha)$, and $\mathcal{E}(p_i) = N^{-1}$. When α is large, the distribution tends to degenerate at $(N^{-1}, \ldots, N^{-1})$, while small values of α indicate p_j's to be small but there is a high probability that a few may not be small. In that case, what is the distribution of $\mathbf{p}_N$? These observations lead him to consider the asymptotic case and establish the following result.

Theorem 1.18 (Kingman) *For $\lambda > 0$, there is a probability measure $\mathbf{P}_\lambda$ on Π^* with the following property. For each N, the finite dimensional random vector $\mathbf{p}_N = (p_1, p_2, \ldots, p_N)$ subject to $p_i \geq 0$, $i = 1, 2, \ldots, N$ and $\sum_{i=1}^{N} p_i = 1$, having symmetric distribution $D(\alpha, \alpha, \ldots, \alpha)$ and $N\alpha \to \lambda$, as $N \to \infty$ and $\alpha \to 0$, then for any n, the distribution of the random vector $(p_{(1)}, p_{(2)}, \ldots, p_{(n)})$ converges to that of $(p_1, p_2, \ldots, p_n)$ under $\mathbf{P}_\lambda$.*

Thus $\mathbf{P}_\lambda$ defines a prior on Π^*, the space of vectors $(p_1, p_2, \ldots)$ just as the Dirichlet process is a prior on Π.

This theorem exhibits the limiting joint distribution of order statistics $p_{(1)} \geq p_{(2)} \geq \ldots$. That is, the distribution of the infinite sequence $(p_1, p_2, \ldots)$ with $p_1 \geq p_2 \geq \ldots$, and $\sum_{i=1}^{\infty} p_i = 1$, depends only on λ and is named as the Poisson-Dirichlet distribution with parameter λ, and denoted by $PD(\lambda)$. Another way to look at $p_{(1)} \geq p_{(2)} \geq \ldots$ is to recognize them as ordered normalized independent increments of a gamma process with shape parameter λ. That is if $Y_{(i)}$ denotes the normalized and ordered size of jump of a gamma process, then $(p_{(1)}, p_{(2)}, \ldots)$ may be considered same as $(Y_{(1)}, Y_{(2)}, \ldots)$. It is clear from the above that the decreasing order statistics of the $D(\alpha_1, \alpha_2, \ldots, \alpha_n)$ provide an approximation to those of $PD(\lambda)$ as long as n is large and α_i are small with their sum $\alpha_1 + \alpha_2 + \ldots + \alpha_n$ being close to λ (Kingman 1993).

It is difficult to derive $PD(\lambda)$ directly. However, following Patil and Taillie (1977), Kingman (1993) gives a stick-breaking construction. Let $V_i \overset{iid}{\sim} Be(1, \lambda)$ and define $q_1 = V_1$ and for $i \geq 2$, $q_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$. q_i's are the weights in the Sethuraman representation of the Dirichlet process. The sequence $\mathbf{q} = (q_1, q_2, \ldots)$ is the size-biased permutation of $\mathbf{p}$ (in the limiting sense). Then it is shown that the vector of its order statistics $\overline{\mathbf{q}} = (q_{(1)}, q_{(2)}, \ldots)$, has the $PD(\lambda)$ distribution. This is shown by first considering the N-dimensional vector $\mathbf{p}_N$ having a symmetric Dirichlet distribution with parameter α and then passing to the limit $N \to \infty$, $\alpha \to 0$ such that $N\alpha \to \lambda$. The distribution of the sequence $(q_1, q_2, \ldots)$ is known as the GEM distribution, as noted earlier.

The vector $\mathbf{p}$ is said to be *invariant under size-biased permutation* (ISBP) if the sequence $\widetilde{\mathbf{p}}$ obtained by size-biased permutation of $\mathbf{p}$ has the same finite dimensional distribution as $\mathbf{p}$. McCloskey (1965) had established an important property regarding $\mathbf{p}$. For any sequence $\mathbf{p}$ with $p_1 = v_1,\ p_n = v_n \prod_{i=1}^{n-1}(1-v_i),\ n \geq 2$ and v_i are independent and identically distributed on $[0, 1]$, then $\mathbf{p}$ is ISBP if $v_i \overset{iid}{\sim} Be(1, \lambda),\ \lambda > 0,\ i > 1$. The 'only if' part was established by Pitman (1996a). Thus, if $\widetilde{\mathbf{p}} = (\widetilde{p}_1, \widetilde{p}_2, \ldots)$ is the vector obtained by the *size-biased permutation*, then it is equal to $\mathbf{p}$ in distribution if and only if $\mathbf{p}$ has the GEM distribution.

It's relation with the residual allocation scheme was highlighted by Patil and Taillie (1977). As observed there, the rank-ordered permutation of Engen's model (in which residual fractions are iid $Be(1, \alpha),\ \alpha > 0$) is equal in distribution to the Kingman's limit. And the size-biased permutation of Kingman's limit equals Engen's model. Thus the two are permutations of each other, and that Engen's model itself is invariant under the size-biased permutation.

A sample of size n drawn from $PD(\lambda)$ will have ties. It will be composed of m_1 singletons (i.e. belonging to distinct species), m_2 pairs, m_3 triplets, and so on, so that $\sum_{j=1}^{n} jm_n = n$. Let $\mathbf{m} = (m_1, m_2, \ldots, m_n)$. Then the probability of observing such a sample is given by

$$\mathcal{P}(\mathbf{m} = \mathbf{m}) = \frac{n!\Gamma(\lambda)}{\Gamma(n+\lambda)} \prod_{j=1}^{n} \left(\frac{\lambda^{m_j}}{j^{m_j} m_j!} \right), \tag{1.121}$$

and is known as Ewen's (1972) sampling formula derived in the context of genetics. This formula was independently discovered by Antoniak (1974) (see Eq. (1.16)). This formula is generalized when sampling is from a two-parameter Poisson-Dirichlet distribution and is given in Eq. (1.125).

The predictive distribution of a future observation from $PD(\lambda)$ is a special case of the one given for the two-parameter Poisson-Dirichlet distribution below.

1.13.2 Two-Parameter Poisson-Dirichlet Process

As noted in Sect. 1.2 (Property 19) the predictive distribution produced by the Dirichlet process selects a new observation with probability $M/(M+n)$ and coincides with one of the previous observations with probability $n/(M+n)$. These probabilities do not depend on K or n_j missing out on some valuable information. Inclusion of this information is achievable via the *two-parameter Poisson-Dirichlet distribution* (also know as *Pitman-Yor process*) developed by Pitman and Yor (1997) (see also, Perman et al. 1992, and Pitman 1995, 1996a) as an extension of one parameter Poisson-Dirichlet distribution. It is a probability distribution over the set of decreasing sequences of positive numbers adding to 1. It is also a two parameter generalization of the Dirichlet process, along with other generalizations mentioned earlier. The parameter λ is replaced by two parameters: a discount parameter α, and a concentration parameter θ, such that $0 \leq \alpha < 1$ and $\theta > -\alpha$ and the distribution is denoted by $PD(\alpha, \theta)$. The discount parameter α governs the power-law behavior which makes this process more suitable than the Dirichlet process for many applications. If $\alpha = 0$, then $PD(0, \theta)$ is the Dirichlet process; and if $\theta = 0$, it yields a random probability whose weights are based on a stable law with index $0 < \alpha < 1$. It may be constructed using the stick-breaking construction in the same way as the one-parameter Poisson-Dirichlet Distribution mentioned above, where the iid variables v_i having the distribution $Be(1, \lambda)$ is replaced by independent variables V_i such that $V_i \sim Be(1 - \alpha, \theta + i\alpha)$, for $i = 1, 2, \ldots$.

Consider again an idealized population of species where $p_1 > p_2 > \ldots$ denote the frequencies of species $\eta_1, \eta_2, \ldots$, respectively, in the population with $\sum_{n=1}^{\infty} p_n = 1$. That is p_1 represents the proportion of the largest size of species in the population, p_2 the second largest, and so on. Consider now the size-biased sampling of the previous section. Let $\widetilde{p}_1$ be a random variable (*size-biased* pick from the sequence $\mathbf{p} = (p_1, p_2, \ldots)$) such that $\mathcal{P}(\widetilde{p}_1 = p_j | \mathbf{p}) = p_j$, $j = 1, 2, \ldots$, and for each $n = 1, 2, \ldots$ and $j = 1, 2, \ldots$,

$$\mathcal{P}(\widetilde{p}_{n+1} = p_j | \widetilde{p}_1, \widetilde{p}_2, \ldots, \widetilde{p}_n, \mathbf{p}) = \frac{p_j I[p_j \neq \widetilde{p}_i, \text{ for all } 1 \leq i \leq n]}{(1 - \widetilde{p}_1 - \widetilde{p}_2 - \ldots - \widetilde{p}_n)}. \quad (1.122)$$

This means, select randomly an individual from the population and set $\widetilde{p}_1 = p_j$ if the selected individual belongs to species labeled η_j. Remove the species corresponding to $\widetilde{p}_1$ from the population and select randomly a second individual from the remaining species. Put $\widetilde{p}_2 = p_j$ if the second individual belongs to species labeled η_j which is obviously different from the label of species of the first individual. The probability that the second individual belongs to j-th species is $p_j/(1 - \widetilde{p}_1)$. Continue the process. This process produces a sequence $\widetilde{p} = (\widetilde{p}_1, \widetilde{p}_2, \ldots)$ which is the size-biased permutation *of* $(p_1, p_2, \ldots)$. By this it is understood that $\widetilde{p}$ is a sequence of proportions of species in order of their appearance in random sampling from the population. Pitman and Yor (1997) gave the following definition of $PD(\alpha, \theta)$ in terms of independent beta random variables.

Definition 1.17 (Pitman and Yor) For $0 \leq \alpha < 1$ and $\theta > -\alpha$, let a probability $P_{\theta,\alpha}$ govern independent random variables V_k with $V_k \sim beta(1-\alpha, \theta + k\alpha)$ and set $\widetilde{p}_1 = V_1$, $\widetilde{p}_n = V_n \prod_{k=1}^{n-1}(1 - V_k)$, $n \geq 2$. Further let $p_1 \geq p_2 \geq \ldots$ be the ranked values of $\widetilde{p}_1, \widetilde{p}_2, \ldots$. Then the Poisson-Dirichlet distribution with parameters (α, θ), denoted as $PD(\alpha, \theta)$, is defined to be the $P_{\theta,\alpha}$ distribution of the sequence $(p_1, p_2, \ldots)$.

Thus $PD(\alpha, \theta)$ is a two-parameter prior distribution on Π^*. This definition reveals that the sequence $(\widetilde{p}_1, \widetilde{p}_2, \ldots)$ obtained by the size-biased random permutation of $(p_1, p_2, \ldots)$ is a simple Engen's (1975) residual allocation model.

They show (their Proposition 2) that under $P_{\theta,\alpha}$ governing $(V_1, V_2, \ldots)$, the sequence $(p_1, p_2, \ldots)$ is such that $p_1 > p_2 > \ldots$ and $\sum_{n=1}^{\infty} p_n = 1$ a.s. and $(\widetilde{p}_1, \widetilde{p}_2, \ldots)$ is a size-biased permutation of $(p_1, p_2, \ldots)$. It is further concluded that if $(p_1, p_2, \ldots)$ is any sequence having $PD(\alpha, \theta)$ distribution with $0 \leq \alpha < 1$ and $\theta > -\alpha$ and $\widetilde{\mathbf{p}}$ is a size-biased permutation of $\mathbf{p}$ and V_n defined as $V_n =$

$\widetilde{p}_n/(\widetilde{p}_n + \widetilde{p}_{n+1} + \ldots)$, then the sequences $\mathbf{p}$, $\widetilde{\mathbf{p}}$ and $(V_1, V_2, \ldots)$ have the same distributions as those in the definition. $V_n = \widetilde{p}_n/(1 - \widetilde{p}_1 - \ldots - \widetilde{p}_{n-1})$ can also be viewed as residual fractions.

It's connection to a simple residual allocation model (stick-breaking scheme) is identified by taking the distribution of independent V_i in the Sethuraman representation as two-parameter beta distribution with parameters $1-\alpha$ and $\theta + i\alpha$. In-depth investigation of $PD(\alpha, \theta)$ has been carried out by Pitman and Yor (1997). Also related papers by Perman et al. (1992) and Pitman (1995, 1996a) shed more light on this distribution.

In the case of one-parameter Poisson-Dirichlet distribution, it was shown that the joint distribution of order statistics of a random discrete probability distribution is the Poisson-Dirichlet distribution. Similarly, it is shown here that the joint distribution of ranked values of size-biased picks is the Poisson-Dirichlet distribution. Pitman and Yor have also shown that the size-biased permutation of rank ordered elements of $\mathbf{p}$ having this distribution produces a sequence of p_i's derived by the residual allocation scheme. It's connection to the random weights in infinite sum representation of a probability measure may be described as follows. $\widetilde{p}_1, \widetilde{p}_2, \ldots$ generated by the stick-breaking construction are the size-biased permutation of weights arranged in decreasing order, or that the size-biased permutation of weights that are in decreasing order may be obtained by the stick-breaking construction.

As $PD(\alpha, \theta)$ is a two-parameter generalization of the Dirichlet process, it offers more flexibility and may offer tremendous advantage over the Dirichlet process in some data modeling. Various applications discussed in Chap. 2 and Chap. 3 may be reworked with replacing the Dirichlet process there with $PD(\alpha, \theta)$. It's application in hierarchical modeling is discussed in Teh and Jordan (2010).

Polya Urn Characterization Now going back to the infinite sum representation of a random probability measure (1.113) with $\xi_i \overset{iid}{\sim} H$, H nonatomic, the prediction rule for $PD(\alpha, \theta)$ can be derived. Suppose we have a sample from the above process, that is, $X_i|P \overset{iid}{\sim} P$, $i = 1, 2, \ldots, n$, $P \sim PD(\alpha, \theta)$. The sample will have some ties. Let $X_1^*, \ldots, X_K^*$ be the K distinct observations among the sample and $n_1, \ldots, n_K$ be their multiplicities, respectively. Integrating out P, the marginal distribution of X_i's will satisfy the following prediction rule (Pitman 1995).

$$X_{n+1}|X_1, X_2, \ldots, X_n \sim \sum_{i=1}^{K} \frac{n_k - \alpha}{\theta + n} \delta_{X_k^*} + \frac{\theta + K\alpha}{\theta + n} H. \tag{1.123}$$

From this it is clear (Pitman 1995, 1996b) that the Poisson-Dirichlet process can be characterized in terms of a generalized Polya urn scheme. Given $X_1, X_2, \ldots, X_n$, choose X_{n+1} at the $(n+1)$-th step to be a new observation with probability $(\theta + K\alpha)/(\theta + n)$ and equal to the previous observation X_k^* with probability $(n_k - \alpha)/(\theta + n)$, $k = 1, 2, \ldots, K$. Thus the probability that X_{n+1} will be a new distinct observation depends upon the number of clusters K and is monotonically increasing in K. α serves as the moderating parameter for this dependence. Its special case, $\alpha = 0$ and $\theta = \alpha(\chi)$ corresponds to the Dirichlet process with parameter α and yields the Blackwell-MacQueen (1973) predictive rule.

The formula for $p_{n,K}$ in the case of $PD(\alpha, \theta)$ is given by Pitman (1995)

$$p_{n,K}(n_1, \dots, n_K|K) = \frac{\prod_{i=1}^{K-1}(\theta + i\alpha)}{(\theta+1)^{(n-1)}} \prod_{j=1}^{K} (1-\alpha)^{(n_j-1)}, \tag{1.124}$$

where as before, $s^{(n)} = \Gamma(s+n)/\Gamma(s) = s(s+1)\dots(s+n-1)$.

This probability may also be expressed alternatively. Suppose in applications to population genetics, a sample of size n is classified in terms of the number of different species, $m_i \geq 0$ consisting of i animals, $i = 1, \dots, n$, represented in the sample, then $\sum_{i=1}^{n} m_i = K$ and $\sum_{i=1}^{n} i m_i = n$. This is same as in the context of partitioning n integers, where m_i represents the number of cells in the partition which contains i integers, $i = 1, \dots, n$. That is m_1 represents singleton cells, m_2 represents the number of cells having a pair, etc. Then we get (Pitman 1995)

$$\mathcal{P}(\mathbf{m} = \mathbf{m}) = n! \frac{\prod_{i=1}^{K-1}(\theta + i\alpha)}{(\theta+1)^{(n-1)} \prod_{i=1}^{n} m_i!} \prod_{j=1}^{n} \left(\frac{(1-\alpha)^{(j-1)}}{j!} \right)^{m_j}, \tag{1.125}$$

known as *Pitman's sampling formula*. This is a two-parameter extension of Ewen's (1972) formula (1.121). When $\alpha \to 0$, we get Ewen's formula, which was also discovered by Antoniak (1974) as given in Sect. 1.2, property 17.

Applications of the Poisson-Dirichlet distribution in hierarchical modeling, approximation to the posterior distribution given a sample from it, and computational aspects using Gibbs sampling algorithm, are discussed in Ishwaran and James (2001). For example, given a sample $X_1, X_2, \dots, X_n$ from $PD(\alpha, \theta)$, the posterior random measure may be approximated by a finite sum given below (Pitman 1996b; Ishwaran and James 2001).

$$P(\cdot|\mathbf{X}) = \sum_{j=1}^{K} p_j^* \delta_{X_j^*}(\cdot) + p_{K+1}^* Q(\cdot), \tag{1.126}$$

where $X_1^*, \dots, X_K^*$ are as before K distinct observations with multiplicities $n_1, \dots, n_K$, respectively,

$$\left(p_1^*, \dots, p_K^*, p_{K+1}^*\right) \sim D(n_1 - \alpha, \dots, n_K - \alpha, \theta + K\alpha) \tag{1.127}$$

which is independent of random measure Q, and $Q \sim PD(\alpha, \theta + K\alpha)$.

1.14 Chinese Restaurant and Indian Buffet Processes

Two new processes have recently garnered a lot of interest from outside the statistical community and found useful in a wide range of interesting applications. In a culinary metaphor they are popularly known as the *Chinese restaurant process* (CRP) and the *Indian buffet process* (IBP). The first one is a simple process which defines a distribution over partitions of N objects into K classes. The second process results in defining a prior over the space of binary $N \times K$ matrices. In both cases K is assumed to be potentially unbounded. The main strategy in their derivation is to find the distributions assuming K to be finite first and then taking the limit as $K \to \infty$. Other simpler methods of derivation are also available in the literature. A very good account of their derivation, applications, as well as their respective connections to the Dirichlet process and beta process is given in Griffiths and Ghahramani (2011).

1.14.1 Chinese Restaurant Process

Imagine a stream of customers arriving at a Chinese restaurant which has unlimited number of tables to sit at. The Chinese restaurant process is a simple scheme to allocate N customers to K tables, where K is presumed to be unlimited. It defines a distribution over partitions of N objects into K classes or partitioning integers $\{1, 2, \ldots, N\}$ into K distinct sets. It is same as the extended Polya urn scheme of Blackwell and MacQueen (1973) discussed in Sect. 1.2. It is also a scheme to generate a sample from the Dirichlet process.

A formal derivation of the CRP may be found in Griffiths and Ghahramani (2011). Their strategy is to define an N-dimensional vector $\mathbf{C} = \{c_1, c_2, \ldots, c_N\}$ with entries representing the classes to which each object is assigned to, i.e. $c_i \in \{1, 2, \ldots, K\}$ for $i = 1, 2, \ldots, N$. Since the objects are exchangeable, they derive the distribution over the equivalence class of $\mathbf{C}$ vectors. The partitioning of N objects in K classes is a multinomial experiment with parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. Therefore, its conjugate prior the Dirichlet distribution is taken to be symmetric with parameter vector $(\alpha/K, \ldots, \alpha/K)$, where α is a positive measure. By integrating out $\boldsymbol{\theta}$, and taking the limit as $K \to \infty$, the marginal probability of an assignment vector belonging to an equivalence class is obtained as

$$\mathcal{P}(\mathbf{C} = \mathbf{c}) = \alpha^{K_+} \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{k=1}^{K_+} (m_k - 1)!, \tag{1.128}$$

where m_k is the number of objects assigned to class k, $k = 1, 2, \ldots, K_+$ and K_+ is the number of classes among K with $m_k > 0$. These probabilities define a distribution over the partitions or sitting N patron at K tables. It also can be considered as providing a prior over class assignments for an infinite mixture model.

This distribution can be alternatively derived by the predictive distribution of a future random variable, conditioned on the past observed variables via the Blackwell-MacQueen Polya urn process. Suppose we have an urn containing infinite number of balls of different colors, denoted by $c_1, c_2, \ldots$. The colors are distributed according to $\overline{\alpha} = \alpha(\cdot)/\alpha(\mathfrak{X})$. At the first step, a ball X_1 is drawn from this set according to the distribution $\overline{\alpha}$, and its color is noted. The ball is replaced along with an additional ball of the same color. At the $(n+1)$-th step, either a ball which is of one of the observed colors is picked with probability $n/(\alpha(\mathfrak{X})+n)$ or a ball of new color is picked with probability $\alpha(\mathfrak{X})/(\alpha(\mathfrak{X})+n)$. In both cases, the ball is replaced along with another ball of the same color, and the step is repeated. Thus a sequence $X_1, X_2, \ldots,$ of random variables is generated where X_i is a random color from the set of colors $\{c_1, c_2, \ldots\}$. Thus we can say that $X_1 \sim \alpha/\alpha(\mathfrak{X})$ and $X_{n+1}|X_1, X_2, \ldots, X_n \sim (\alpha + \sum_{i=1}^n \delta_{x_i})/(\alpha(\mathfrak{X})+n)$, which can be written equivalently as

$$X_{n+1}|X_1, X_2, \ldots, X_n \sim \sum_{i=1}^{n} \frac{1}{\alpha(\mathfrak{X})+n}\delta_{x_i} + \frac{1}{\alpha(\mathfrak{X})+n}\alpha. \tag{1.129}$$

Some colors, say K will be repeated in n draws. Denote the distinct colors among them by $X_1^*, \ldots, X_K^*$ and let m_k be the number of times the color X_k^* is repeated, $k = 1, 2, \ldots, K$, $m_1 + \ldots + m_K = n$. Then the above expression can be written in terms of K, as

$$X_{n+1}|X_1, X_2, \ldots, X_n, K \sim \sum_{k=1}^{K} \frac{m_k}{\alpha(\mathfrak{X})+n}\delta_{X_k^*} + \frac{1}{\alpha(\mathfrak{X})+n}\alpha. \tag{1.130}$$

This process is sequential and continues indefinitely. The limit of the joint distribution will yield the above probability. Note that $\alpha(\mathfrak{X})$ represents the α measure of the above derivation.

This process has been interpreted in practical terms and popularized in culinary metaphor by the catchy name, Chinese Restaurant process (attributed to Jim Pitman and Lester Dubins by Griffiths and Ghahramani 2006). The correspondence is as follows.

The sequence $X_1, X_2, \ldots,$ of draws of balls represent incoming patrons at a Chinese restaurant, different colors of balls represent tables with different dishes (one dish per table), each of unlimited sitting capacity (that is there are infinite many balls of each color). Each customer sits at a table. The first customer at a table orders randomly a dish for the table according to the distribution $\alpha(\cdot)/\alpha(\mathfrak{X})$. The $(n+1)$-th customer chooses to join previous customers with probability $n/(\alpha(\mathfrak{X})+n)$ or chooses a new table with probability $\alpha(\mathfrak{X})/(\alpha(\mathfrak{X})+n)$ and orders a dish. If he joins previous customers and there are already K tables occupied, then he joins the table k with dish X_k^* with probability $m_k/(\alpha(\mathfrak{X})+n)$, where m_k is the number of customers already enjoying the dish X_k^*, $k = 1, 2, \ldots, K$ i.e. $X_{n+1} = X_k^*$ with probability $m_k/(\alpha(\mathfrak{X})+n)$. If he chooses a new table, he orders a random dish distributed according to $\alpha(\cdot)/\alpha(\mathfrak{X})$. This yields the above two expressions.

Patrons are exchangeable as are the random variables X_i's in the Polya urn sequence. The probability of a particular sitting arrangement depends only on m_k, which is a function of n, and not on the order in which they arrive and sit. As n increases, there is greater probability that the new patron will choose an occupied table rather than a new table. After n steps, the output of the CRP is a partition of n customers across K tables, or grouping of n balls in K distinct colors, or simply, partitioning of integers $\{1, 2, \ldots, n\}$ into K distinct sets and CRP is the induced distribution over partitions. The expected number of tables occupied by first n customers is $\sum_{i=1}^{n}[\alpha(\mathfrak{X})/(\alpha(\mathfrak{X}) + i - 1)]$ (Antoniak 1974) and tend to infinity along with n. That is, as $n \to \infty$, the number of partitions tend to infinity as well.

The CRP can be used to generate a realization of the Dirichlet process. In the case of Sethuraman representation, locations are generated as $\xi_i \overset{iid}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$ and the weights are generated by

$$p_k = \lim_{n\to\infty} \frac{m_k(n)}{\alpha(\mathfrak{X}) + n}. \tag{1.131}$$

The CRP is obtained by integrating out the Dirichlet process and thus it describes the marginal distributions in terms of random partitions determined by K tables in a restaurant. Samples from the Dirichlet process are probability measures and samples from the CRP are partitions. Teh et al. (2006) proposed a further generalization as *CRP franchised* in which different restaurants are linked together in terms of a common menu. This corresponds to a hierarchical Dirichlet process in which the base measure of the Dirichlet process F_0 is itself considered as having a Dirichlet process prior with hyper parameters, say, M^* and G_0.

1.14.2 Indian Buffet Process

In the Chinese restaurant process each object (patron) can posses only one feature (choose one table or order one dish), and thus can be represented as a matrix with rows representing objects (patrons) and columns representing features (the tables they occupy or dishes they order). Thus the matrix will have a single entry of one in each row and all other entries will be zeros. However, in certain applications each object may possess an unlimited number of features, therefore a generalization of the above model is desirable. This leads to the development of the *Indian Buffet process* proposed by Griffiths and Ghahramani (2006).

The IBP is essentially a process to define a prior distribution on the equivalence classes of sparse binary matrices (entries of the matrix have binary responses) consisting of a finite number of rows and an unlimited number of columns. Rows are interpreted as objects and columns as potentially unlimited features. In contrast to the CRP, here the matrix may have entries of 1's in more than one column in each row. Thus it can serve as a prior for probability models involving objects and features encountered in certain applications in machine learning, such as image processing. It also provides a tool to handle nonparametric Bayesian models with large number of latent variables. The catchy phrase is coined by the authors as a culinary metaphor in view of the similarity with the CRP. It is claimed that Indian restaurants in London offer buffet with almost unlimited number of dishes. In their analogy, patrons visiting the restaurants are objects and dishes they choose are features. In the Chinese restaurant process each patron chooses a single table (or dish) and stream of patron continues. In the IBP, a patron may choose any number of dishes and the number N of patrons is fixed. An expanded review of the process, its applications and the method of generating samples can be found in their paper (Griffiths and Ghahramani 2011).

Let $\mathbf{Z}$ be a binary response matrix of N rows and an unlimited number of columns. It's elements z_{ik} denotes the fact that object i posses k-th feature, $i = 1, \ldots, N$ and $k = 1, 2, \ldots,$ and takes on values 1 or 0 according to whether the feature is present or not. The task is to derive the probability distribution of a random matrix $\mathbf{Z}$.

The authors' approach in deriving the probability distribution of $\mathbf{Z}$ is to start with a finite number of columns K and then consider the limit as K tends to infinity. Let μ_k denote the probability that an object possesses k-th feature and that the features are generated independently. Thus, $z_{ik} \sim Ber(\mu_k)$, $k = 1, 2, \ldots, K$ for each $i = 1, 2, \ldots, N$. Under this model and given $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$, the probability of $\mathbf{Z} = \mathbf{z}$ is given by

$$P(\mathbf{Z} = \mathbf{z}|\boldsymbol{\mu}) = \prod_{k=1}^{K}\prod_{i=1}^{N} P(z_{ik}|\mu_k) = \prod_{k=1}^{K} \mu_k^{m_k}(1-\mu_k)^{N-m_k}, \tag{1.132}$$

where $m_k = \sum_{i=1}^{N} z_{ik}$ is the number of objects possessing feature k. Assigning a beta prior, $Be(\frac{\alpha}{K}, 1)$ to each μ_k, where α is a strength parameter of the IBP yields

$$P(\mathbf{Z} = \mathbf{z}) = \prod_{k=1}^{K} \frac{(\alpha/K)\Gamma(m_k + \alpha/K)\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \alpha/K)}, \tag{1.133}$$

which tends to 0 as K tends to ∞. However, this distribution depends only on m_k and not on the order of the columns. That is the probability remains unchanged if the columns are permuted. Permutation of columns is an equivalence relation on the set of $N \times K$ matrices.

An analogy with customer-dishes can be made as follows. Patrons (objects) stream in one by one and taste dishes (features). If the i-th patron tastes k-th dish, then $z_{ik} = 1$; otherwise $z_{ik} = 0$. Thus m_k is the number of patrons tasting k-th dish. If the interest is only on what dishes are tasted and not in the order in which they are tasted, then the interest might be on the probability of observing any matrix $\tilde{\mathbf{z}}$ in the equivalence class of $\mathbf{z}$. This probability is shown to be

$$P(\mathbf{Z} = \tilde{\mathbf{z}}) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!} P(\mathbf{Z} = \mathbf{z}), \tag{1.134}$$

where K_h denotes the number of columns having the full history h, $h = 0, 1, \ldots,$ $2^N - 1$ (total number of histories is 2^N), a typical history being $(z_{1k}, \ldots, z_{Nk})$ with $z_{ik} = 0$ or 1, and K_0 being the number of features for which $m_k = 0$. This is a distribution over the equivalence classes obtained by partitioning a set of binary matrices according to column-permutation.

In order to define a distribution over infinite dimensional binary matrices, the authors take into account how customers choose dishes and define *left-ordered* binary matrices. A typical *left-ordered* binary matrix is generated from $\mathbf{Z}$ by first accumulating to the left all columns of $\mathbf{Z}$ for which $z_{1k} = 1$, i.e. all dishes tried by the first patron. Next put together all columns on the left for which $z_{2k} = 1$, and so on. Equivalence classes are defined with respect to these matrices and the probability of producing a specific matrix belonging to the equivalence class by this process, as $K \to \infty$ and K_h held fixed, is shown to be

$$P(\mathbf{Z} = \tilde{\mathbf{z}}) = \frac{\alpha^{K_+}}{\prod_{h=0}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}, \qquad (1.135)$$

where $K_+ = \sum_{h=0}^{2^N-1} K_h$, the number of features for which $m_k > 0$ (thus $K = K_0 + K_+$) and $H_N = \sum_{j=1}^{N} \frac{1}{j}$, the N-th harmonic number. So this distribution represents a prior with parameter α over the binary matrices with N rows and an unlimited number of columns in the same way as the Dirichlet process is a prior over the class of all probability measures, and $PD(\lambda)$ is the prior over all discrete probability distributions $\mathbf{p}$ with $p_1 > p_2 > \ldots$.

A random draw from the Dirichlet process can be obtained by any one of the methods mentioned in Sect. 1.2. Likewise, the above probability distribution can be derived from the IBP with parameter α, as follows. The derivation by the stick-breaking construction is mentioned thereafter.

In the IBP, N customers enter a restaurant sequentially which has the choice of infinitely many dishes arranged in a line. The first customer starts at the left and tastes the number of dishes according to the Poisson distribution with parameter α. The n-th customer $n = 1, \ldots, N$ moves along the buffet sampling dishes according to the popularity, and tasting dish k with probability m_k/n, where m_k is the number of previous $n-1$ customers who have tasted the same dish k. In addition, he samples a number of new dishes, not tasted before by any customer, according to the Poisson law with parameter α/n. The selection of different dishes by different customers can be indicated by a binary matrix $\mathbf{Z}$.

The probability of any particular matrix generated by the IBP is then shown to be

$$P(\mathbf{Z} = \mathbf{z}) = \frac{\alpha^{K_+}}{\prod_{n=1}^{N} K^{(n)}!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}, \qquad (1.136)$$

where $K^{(n)}$ is the number of new dishes sampled by the n-th customer and K_+ is the number of dishes for which $m_k > 0$. These matrices are not in the left-ordered form. Therefore an adjustment is made by multiplying this probability by the factor $\prod_{n=1}^{N} K^{(n)}! / \prod_{h=0}^{2^N-1} K_h!$ yielding the desired probability of Eq. (1.135) (see Griffiths and Ghahramani 2006 for details).

A two-parameter generalization of IBP is derived by Ghahramani et al. (2007) by assigning a $Be(\frac{\alpha\beta}{K}, \beta)$ (instead of $Be(\frac{\alpha}{K}, 1)$) prior to each μ_k. In the metaphor of Indian buffet, the first customer starts at left of buffet lay out and samples $Poisson(\alpha)$ dishes. The n-th customer tastes any dish k from previously sampled dishes by $m_k > 0$ customers with probability $m_k/(\beta + n - 1)$, and in addition tastes $Poisson(\alpha\beta/(\beta + n - 1))$ new dishes. The authors provide useful interpretation of these parameters.

Teh and Gorur (2010) introduce a 3-parameter generalization of the IBP with power-law behavior. The parameters are α, β, and σ such that $\alpha > 0$ and $\beta > -\sigma$ and $\sigma \in [0, 1)$. In the context of Indian buffet, the first customer tries $Poisson(\alpha)$ dishes; the n-th customer tries previously tried dish k ($m_k > 0$) with probability $(m_k - \sigma)/(\beta + n - 1)$, $k = 1, 2, \ldots, K^+$ and tries a number of new dishes according to $Poisson(\alpha r)$, where $r = (\Gamma(1 + \beta)\Gamma(n - 1 + \beta + \sigma))/(\Gamma(n + \beta)\Gamma(\beta + \sigma))$. When $\sigma = 0$, it reduces to the two parameter IBP. This is akin to Pitman-Yor process discussed in the last section which generalized the Dirichlet process by the introduction of an additional discount parameter σ. The authors also give the stick-breaking construction for this process which is the same as for the IBP and show that their power-law IBP is a good model for word occurrences in document corpora.

Thibaux and Jordan (2007) connects the Indian Buffet process to a particular form of the beta process B (discussed in Sect. 1.8), in the same way the CRP is connected to the Dirichlet process. It is shown that the IBP is an iid mixture of the Bernoulli processes with mixing (de Finetti) measure the beta process, as the Dirichlet process is the mixing measure for the CRP. Here again customers and dishes are identified with objects and features, respectively. Let Z_i be a binary row vector of $\mathbf{Z}$, $i = 1, \ldots, n$. Let B be the beta process with parameter $c(\cdot)$ a positive function and B_0, a fixed measure, and given B, let Z_i be distributed as the Bernoulli process with parameter B. That is, $B \sim \mathcal{B}e\{c, B_0\}$ and $Z_i|B \sim BeP(B)$, for $i = 1, \ldots, n$ are independent Bernoulli draws from B. Integrating out B, we have the marginal predictive distribution as

$$Z_{n+1}|Z_1, \ldots, Z_n, c, B_0 \sim BeP\left(\frac{c}{c+n}B_0 + \sum_{k=1}^{K}\frac{m_k}{c+n}\delta_{\omega_k}\right), \tag{1.137}$$

where m_k is the number of customers among n having tried dish k and ω_k stands for ‘tastes dish k’. It’s interpretation in the terminology of Indian Buffet process is as follows. Suppose B_0 is continuous and c is constant such that $\lambda = B_0(\Omega)$ is finite. Since $Z_1 \sim BeP(B_0)$ and B_0 is continuous, Z_1 is a Poisson process (B_0), and the total number of features of Z_1 is $Z_1(\Omega) \sim P(\lambda)$. That is the first customer will taste $P(\lambda)$ number of dishes. For the $(n + 1)$-th customer, Z_{n+1} is sum of two components: U the number of dishes already tasted by n customers, and V the number of new dishes he will taste. $U \sim BeP(\sum_k \frac{m_k}{c+n}\delta_{\omega_k})$ and $V \sim BeP(\frac{c}{c+n}B_0)$. U will have mass $\frac{m_k}{c+n}$ at ω_k i.e. he will taste dish k already tried by previous customers with probability $\frac{m_k}{c+n}$, $k = 1, \ldots, K$, and will taste additionally $P(\frac{c\lambda}{c+n})$ number of new dishes.

Here the underlying Dirichlet/multinomial structure for the CRP is replaced by beta/Bernoulli structure. For an application to document classification problem, their paper should be consulted.

Stick-Breaking Construction of IBP To sample a binary matrix from the distribution of **Z** we need μ_k's (similar to p_i's in the Sethuraman representation of the Dirichlet process). But since we do not care for the ordering of columns, it is sufficient to generate ordered μ_k's. Then these ordered μ_k's are given in terms of θ_k's with each $\theta_k \sim Be(\alpha, 1)$. Let $\mu_{(1)} > \mu_{(2)} > \ldots > \mu_{(K)}$ be a decreasing reordering of $\mu_1, \mu_2, \ldots, \mu_K$, where a $Be(\frac{\alpha}{K}, 1)$ prior is placed on each μ_k. As $K \to \infty$, Teh et al. (2007) construct a stick-breaking representation of the Indian buffet process as follows. Let

$$\theta_k \overset{iid}{\sim} Be(\alpha, 1); \qquad \mu_k = \theta_k \mu_{k-1} = \prod_{l=1}^{k} \theta_l, \tag{1.138}$$

and θ_k is independent of $\mu_1, \mu_2, \ldots, \mu_{k-1}$, $k = 1, 2, \ldots, K$. This construction may be viewed in terms of breaking a stick of unit length. At the first stage, cut the stick at point θ_1 chosen randomly according to $Be(\alpha, 1)$, and discard the cut piece and label the length of the remaining part of stick as μ_1. At the second stage, cut the remaining part of the stick at point $\theta_2 \sim Be(\alpha, 1)$ relative to the current length of the stick, and discard the cut piece. Label the length of the remaining part of stick as μ_2. Continue this process.

They connect this process to the one for the Dirichlet process. If at stage k, we denote the length of the discarded piece as p_k, then we have

$$p_k = (1 - \theta_k)\mu_{k-1} = (1 - \theta_k) \prod_{l=1}^{k-1} \theta_l. \tag{1.139}$$

Now making a change of variable $V_k = 1 - \theta_k$, we have $V_k \overset{iid}{\sim} Be(1, \alpha)$ and setting $p_k = V_k \prod_{l=1}^{k-1}(1 - V_l)$, p_k's turn out to be the weights in stick-breaking construction of the Dirichlet process.

As pointed out by them, in both constructions, the weights are obtained as the lengths of sticks. In Dirichlet process, the weights p_k are the lengths of discarded pieces, whereas in Indian buffet processes, the weights μ_k are the lengths of sticks remaining. Thus in the Dirichlet process construction, the p_k's necessarily add to 1 but need not have any order among them. In contrast, the μ_k need not add to 1, but are in decreasing order. This duality between the Dirichlet process and the IBP may be exploited for further extensions of the IBP mirroring the extensions of the Dirichlet process. For example, Pitman-Yor (1997) extension of the Dirichlet process may be adapted to the IBP by taking $\theta_k \sim Be(\alpha + k\sigma, 1 - \sigma)$ and $\mu_k = \prod_{l=1}^{k} \theta_l$.

1.15 Some Other Processes

In this section we briefly mention a few other prior processes that have been proposed and investigated in the literature. With the exception of the hierarchical and mixture processes, these processes have not attracted as much attention as the above mentioned processes have. A common practice in hierarchical Dirichlet process modeling is to assign a prior (often taken as a Dirichlet process) to the parameter α of the Dirichlet process so that Dirichlet processes across the board could be linked. These models may be collectively viewed as *mixture models* which can be traced back to Antoniak (1974) who introduced the mixtures of Dirichlet processes. Such models have been particularly proved useful in modeling the distribution of words in text documents. These are introduced in this section briefly.

1.15.1 Dirichlet-Multinomial Process

In the context of Bayesian inference for sampling from a finite population, Lo (1986) defines a finite dimensional process. Assume that $F \in \mathcal{D}(\alpha)$ and given F, $X_1, \ldots, X_N$ is a iid sample from F. The marginal distribution of $X_1, \ldots, X_N$ is then symmetric and is a function of N and α. A point process $N(\cdot) = \sum_{j=1}^{N} \delta_{X_j}(\cdot)$ defined on $(R, \mathcal{B})$ is called a *Dirichlet-multinomial process* with parameters (N, α) if for any k and any partition $(B_1, \ldots, B_k)$ of R, the random vector $(N(B_1), \ldots, N(B_k))$ is a Dirichlet-multinomial with parameters $(N; \alpha(B_1), \ldots, \alpha(B_k))$. Alternatively it may also be stated as follows. Let F_0 be a distribution function, $M > 0$ a real number, and N a positive integer. A point process $N(\cdot)$ on $(R, \mathcal{B})$ is said to be a *Dirichlet-multinomial process* with parameters (M, F_0, N) if for any k and any partition $(B_1, \ldots, B_k)$ of R, the random vector $(N(B_1), \ldots, N(B_k))$, given F has a conditional multinomial distribution with parameters $(N; F(B_1), \ldots, F(B_k))$ and $F \in \mathcal{D}(\alpha)$ on R, with $\alpha(R) = M$ and $F_0(\cdot) = \alpha(\cdot)/M$. He shows that this process is conjugate and the Dirichlet process is the limit of the Dirichlet-multinomial processes. It also appears in connection with Bayesian bootstrap (Lo 1987).

1.15.2 Dirichlet Multivariate Process

In the process of dealing with nonparametric Bayesian estimation in a competing-risks model, Salinas-Torres et al. (2002) introduced a multivariate version of the Dirichlet process called *Dirichlet Multivariate process* as follows.

Let $(\mathfrak{X}, \mathcal{A})$ be a measurable space and $\alpha_1, \ldots, \alpha_k$ be finite, non-null and nonnegative measures defined on $(\mathfrak{X}, \mathcal{A})$. Let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_k)$, $P_1, \ldots, P_k$ be mutually independent random elements defined on a suitable probability space. Suppose further that $\boldsymbol{\rho}$ has a singular Dirichlet distribution $D(\alpha_1(\mathfrak{X}), \ldots, \alpha_k(\mathfrak{X}))$, and $P_j \in \mathcal{D}(\alpha_j)$, $j = 1, \ldots, k$. Set $\mathbf{P}^* = (P_1^*, \ldots, P_k^*) = (\rho_1 P_1, \ldots, \rho_k P_k)$. Then $\mathbf{P}^*$ is said to be a *Dirichlet multivariate* (*k-variate*) process with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$. Also, $P_1^*, \ldots, P_k^*$ are subprobability measures and $\sum_{j=1}^{k} P_j^*$ is a probability measure on $(\mathfrak{X}, \mathcal{A})$. They show that the posterior distribution is obtained simply by updating the parameters of the priors. They also derive some weak convergence results of $\mathbf{P}^*$ and use it in deriving nonparametric Bayesian estimators in competing-risks models (to be presented in Sect. 3.5.3).

1.15.3 Generalized Dirichlet Process

Hjort (1990) defines a prior for the distribution function F via it's cumulative hazard function H. Let H be a beta process prior with parameters $c(\cdot)$ and $H_0(\cdot)$, that is symbolically, $H \sim \mathcal{B}e\{c(\cdot), H_0(\cdot)\}$, and consider the random CDF $F(t) = 1 - \prod_{[0,t]}\{1 - dH(s)\}$. Then $\mathcal{E}[F(t)] = F_0(t) = 1 - \prod_{[0,t]}\{1 - dH_0(s)\}$ as shown in Sect. 1.5. It is then noted in Hjort (1990) that F is a Dirichlet process with parameter $\kappa F_0(\cdot)$ and κ is a positive constant chosen so that $c(s) = \kappa F_0[s, \infty)$. Thus in this case F is identified as a *generalized Dirichlet process* with two parameters, $c(\cdot)$ and $F_0(\cdot)$. This fact is the key idea in developing the beta-Stacy (Walker and Muliere 1997a) process. Hjort also notes that if $H \sim \mathcal{B}e\{c(\cdot), H_0(\cdot)\}$, then $B = -\log(1 - F)$ is a Lévy process (independent nonnegative increments process) with Lévy representation

$$\mathcal{E}\big(e^{-\theta B(t)}\big) = \left(\prod_{j:t_j \le t} \mathcal{E}(1 - S_j)^{\theta}\right) \times \exp\left\{-\int_0^{\infty} \big\{\psi\big(c(s) + \theta\big) - \psi\big(c(s)\big)\big\} c(s) dH_{0c}(s)\right\}, \quad (1.140)$$

where $\psi(x)$ is the digamma function $\Gamma'(x)/\Gamma(x)$. Various properties and applications of the Dirichlet process developed in Sect. 1.2 may be investigated for the generalized Dirichlet process as well.

1.15.4 Beta-Neutral Process

In Sect. 1.5, neutral to the right processes were defined in which the hazard contributions of arbitrary intervals of a partition of the real line were assumed to be independent with arbitrary distributions. If instead they are assumed to have independent beta distributions, Lo (1993a) called the resulting processes as *beta-neutral* processes. He uses the beta-neutral process to define a random survival function via two independent gamma processes, one representing the complete data and the other representing the censored data. He shows that this family is conjugate and the posterior distribution, given the data, can be evaluated by simply updating the parameters.

1.15.5 Bernstein-Dirichlet Prior

Petrone (1999) introduces a class of prior distributions on the space $\mathcal{F}[0, 1]$ of distribution functions F defined on the closed interval [0, 1]. This is done via constructing a Bernstein polynomial. Given a function F on the closed interval [0, 1], a *Bernstein polynomial of order k* of F is defined as

$$B(x; k, F) = \sum_{j=0}^{k} F\left(\frac{j}{k}\right)\binom{k}{j} x^j (1-x)^{k-j}. \tag{1.141}$$

If F is a random distribution function on [0, 1], and k is also random, then clearly, so is the polynomial $B(x; k, F)$. As $k \to \infty$, $B(x; k, F) \to F(x)$ at each point of continuity x of F. It's derivative can be written as a mixture of beta densities $\sum_{j=1}^{k} w_{j,k} Be(x : j, k-j-1)$ where

$$w_{j,k} = F\left(\frac{j}{k}\right) - F\left(\frac{j-1}{k}\right), \quad j = 1, 2, \ldots, k, F(0) = 0, \tag{1.142}$$

$w_{j,k} \geq 0$ and $\sum_{j=1}^{k} w_{j,k} = 1$. The mixture is a probability density. By randomizing k and the weights $w_{j,k}$ of the mixture, a prior on the space of densities on [0, 1] can be constructed. A probability measure induced by B is called a *Bernstein prior* and it's construction is described. For example, if k and F are dependent, given k, F can be chosen as a Dirichlet process prior with parameter α_k. If they are independent, then a joint distribution for the pair (k, F) may be assigned on the corresponding product space. If in this latter case, $F \in \mathcal{D}(\alpha)$, she calls such a prior as *Bernstein-Dirichlet prior*. It is shown to have full support and it can also select an absolutely continuous distribution function with a continuous and smooth derivative.

1.15.6 Hierarchical and Mixture Processes

In nonparametric Bayesian analysis, we usually come across three types of mixtures. The first type is a parametric mixture discussed in Antoniak (1974) (see Sect. 1.4), where the parameter of the Dirichlet process was considered to be random and the mixing of Dirichlet processes was done with respect to a parametric distribution. The random probability measure so obtained was called a mixture of Dirichlet processes. That is, $P(\cdot) = \int \mathcal{D}(\alpha_u) H(du)$, H a distribution function.

To place a prior over density functions, Lo (1984) took a second kind of the mixture, a mixture with respect to a known kernel. He (Lo 1984, 1986) approaches this problem by defining a kernel representation of the density function as, $f(x|G) = \int_R K(x, u)G(du)$, where G is distribution function on R. By treating G to be random with a $\mathcal{D}(\alpha)$ prior, a random mixing distribution is defined. Here $K(x, u)$ is a kernel defined on $(\mathcal{X} \times R)$ into R^+ such that for each $u \in R$, $\int_{\mathcal{X}} K(x, u)dx = 1$

and for each $x \in \mathcal{X}$, $\int_R K(x,u)\alpha(du) < \infty$. The mixture is of parametric families with respect to a nonparametric distribution. By integrating out G, we get a kernel mixture of the Dirichlet processes. The difference between the two is that the mixing components and mixing weights are interchanged. (To place a prior over hazard rates, Dykstra and Laud (1981) took the mixing distribution to be a gamma process (Sect. 1.7.1).)

A third type of mixture is the one which leads to hierarchical models where the parameters of the prior distributions themselves are assigned priors with hyper parameters. It has a long history of applications in parametric and semiparametric set ups. It's adaptation to the case of nonparametric (infinite-dimensional parameters) was undertaken in Teh et al. (2004, 2006) where they discuss hierarchical Dirichlet processes and indicate their extensions to other priors. Their interest stems from a need to model group data where each group is associated with a mixture model and the desire is to link them together. In such cases, the discreteness of the Dirichlet process turns out to be an asset.

In the usual nonparametric Bayesian models we assume $X_1, \ldots, X_n \overset{iid}{\sim} F$ and F a Dirichlet process with parameters M and F_0, denoted by $F \sim \mathcal{D}(M, F_0)$ (a slight departure from the usual notation of $\mathcal{D}(\alpha)$ to emphasize the roles of M and F_0). As before M is a concentration parameter and F_0 the baseline distribution, which is generally taken to be a parametric distribution. However priors may as well be assigned to either M or F_0, or both. Teh et al. (2006) justify doing that from practical aspects and identify different types of hierarchical and mixture models.

If the baseline distribution F_0 in the $\mathcal{D}(M, F_0)$ is itself assumed to be a Dirichlet process with hyper parameters, M^* and G_0, it is referred to as a *Hierarchical Dirichlet process*. Thus we have

$$\begin{aligned} &X_1, \ldots, X_n | P \overset{iid}{\sim} P, \qquad P|M, F_0 \sim \mathcal{D}(M, F_0) \quad \text{and} \\ &F_0 | M^*, G_0 \sim \mathcal{D}\big(M^*, G_0\big), \end{aligned} \tag{1.143}$$

each conditionally independent.

A second possibility useful in clustering problems is to model the data where it is subdivided into groups with each group's random measure P_i such that

$$P_i | M, F_0 \overset{iid}{\sim} \mathcal{D}(M, F_0) \quad \text{and} \quad F_0 | M^*, G_0 \sim \mathcal{D}\big(M^*, G_0\big), \tag{1.144}$$

and each conditionally independent. In a Chinese restaurant analogy, where the groups may represent different restaurants sharing a menu of a set of common dishes, it is referred to as the *Chinese Restaurant franchise*. The common dishes are modeled by treating F_0 to be a discrete distribution with its atoms representing the dishes. This model is found to be useful in the information retrieval field.

A third possibility is to have

$$\begin{aligned} &X_j | \theta_j \sim F(\theta_j), \quad \text{for each } j, \qquad \theta_j | P \overset{iid}{\sim} P, \quad \text{and} \\ &P | M, F_0 \sim \mathcal{D}(M, F_0). \end{aligned} \tag{1.145}$$

This model is referred to as a *Dirichlet Process mixture model.* An additional level may also be introduced as a model with

$$\begin{aligned}
&X_{ji}|\theta_{ji} \sim F(\theta_{ji}) \quad \text{for each } j \text{ and } i,\\
&\theta_{ji}|P_j \overset{iid}{\sim} P_j, \qquad P_j|M, F_0 \overset{iid}{\sim} \mathcal{D}(M, F_0), \quad \text{and}\\
&F_0|M^*, G_0 \sim \mathcal{D}\big(M^*, G_0\big),
\end{aligned} \tag{1.146}$$

where, as usual all are assumed to be conditionally independent. This model is referred to as a *Hierarchical Dirichlet process mixture model.*

Teh et al. (2006) provide stick-breaking constructions of these models and their analogs in terms of the Chinese restaurant process and Chinese restaurant franchise models. The predictive distributions and inferential procedures of these models, and applications in document modeling are also discussed in the paper.

The Dirichlet process in the above analysis may be replaced with its two-parameter generalization, $PD(\alpha, \theta)$ to gain more flexibility. This is found to be useful in areas such as natural language modeling and image processing. This aspect is pursued by Teh and Jordan (2010), where further material may be found.

Ishwaran and James (2001) extended the hierarchical modeling approach by replacing the Dirichlet process with a broader class of priors called *species sampling models,* developed by Pitman (1995, 1996a) corresponding to a random probability measure of the form (1.114). These priors include several other processes discussed in earlier sections. The same can be done in the above hierarchical models as well. In fact, there is a growing interest in recent years in exploring mixture models in modeling large scale data sets as they offer more flexibility and at the same time are computationally feasible.

Parallel to the hierarchical Dirichlet process, Thibaux and Jordan (2007) proposed a *hierarchical beta process*. Recall that in the beta process (Sect. 1.8), we have vectors Z_i's of binary responses $z_{ik} = 1$ or 0, $k = 1, 2, \ldots$ according as whether the feature is present or not. Thus, the vectors Z_i's are distributed as Bernoulli processes with parameter B, and B is assumed to have a beta process prior with parameters c and B_0. In symbols, $Z_i \overset{iid}{\sim} BeP(B)$ and $B \sim BP(c, B_0)$. In modeling documents by the sets of words or phrases they contain, assume that the documents are classified into K categories, $A_1, \ldots, A_K$ and Z_{ij} represents the j-th document in i-th category, $i = 1, \ldots, K$, each Z_{ij} being a vector of binary responses with p^i_ω being the probability, specific to the i-th category, of feature ω (say a word) being present. Then the hierarchical beta process model may be expressed as $Z_{ij} \sim BeP(B_i)$, $j = 1, \ldots, n_i$, $B_i \sim BP(c_i, B)$, $i = 1, \ldots, K$ and $B \sim BP(c, B_0)$, subject to certain conditions on the variables and parameters involved (see Thibaux and Jordan 2007 for details).

Alternative to hierarchical modeling, empirical Bayes approach also offers some advantage. Instead of putting a prior on the baseline distribution F_0, it may be estimated from the (past) data itself. This is done in the empirical Bayes approach in various applications discussed in Chap. 2 and Chap. 3. This approach has some merit over the hierarchical modeling methods in the sense that the data itself guides the value(s) of unknown parameter(s) as opposed to assuming certain arbitrary priors. In the empirical Bayes applications of the Dirichlet process, M and F_0 were estimated consistently using the past data and the analysis proceeded. The author is unaware if any attempts have been made in estimating similarly the parameters of various other processes developed in recent years.

In the kernel mixtures, usually the normal distribution with mean 0 and variance σ^2 is the preferred choice of the kernel. But this may be extended to include other types of functions. James (2006) explores this path in his paper.

1.16 Bivariate Processes

Bivariate extensions of the prior processes are not so easy. In the non-Bayesian context, several attempts have been made to extend the Kaplan-Meier (1958) PL estimator to the case of bivariate survival function but they encountered problems. In some cases the estimators of a distribution function (or a survival function) obtained are not proper distribution functions (or survival functions), while in other cases, no explicit or simple forms are possible. Dabrowska (1988) constructed an analogue of the PL estimator which is consistent but is not a proper survival function as it assigns negative mass to certain regions. For other efforts, see the references in Dabrowska (1988). However there is some hope in the Bayesian approach.

Ferguson's (1973) Dirichlet process was defined on an arbitrary space of probability measures. This made it easy in extending the Dirichlet process to higher dimensions in a straight forward manner. See for example, Ferguson (1973, 1974), Dalal and Phadia (1983), Phadia and Susarla (1983), among others, who assign a Dirichlet process prior for an unknown bivariate distribution function defined on $R \times R^+$ or $R^+ \times R^+$ in addressing some estimation problems. In dealing with the estimation of a survival function, Tsai (1986) follows a slightly different path. He places a Dirichlet process prior with parameter α^*, on $(\mathcal{R}^*, \mathcal{B}^*)$, where $\mathcal{R}^* = R^+ \times \{0, 1\}$ and $\mathcal{B}^* = \mathcal{B} \times \{\phi, \{0\}, \{1\}, \{0, 1\}\}$, $\mathcal{B}$ is a Borel field on R^+ and α^* is a non-null finite measure on $(\mathcal{R}^*, \mathcal{B}^*)$. Salinas-Torres et al. (2002) generalize Tsai's approach by taking the second coordinate with values in $\{1, \ldots, k\}$. In the context of survival data, Pruiit (1988) shows that the Bayes estimator of a bivariate survival function with Dirichlet prior could be inconsistent. This point was further discussed in Ghosh et al. (2006).

On the other hand, all the processes belonging to the class of processes neutral to the right are defined on the real line and their extension to the bivariate case is difficult and remained unexplored. However, there is a renewed interest and several attempts have been made in recent years. See for example, Walker and Muliere (2003), Ghosh et al. (2006), Bulla et al. (2007, 2009), and Phadia (2007). Recall that in the univariate case the Bayesian estimator with respect to the Dirichlet process prior and more generally, with respect to the neutral to the right processes, converges to the PL estimator. Recognizing this fact, Ghosh et al. (2006) approach this problem by developing a natural generalization of the beta process to the bivariate case and derive an estimator using an approach which is labeled as “essential likelihood approach” in view of it not using full likelihood.

Bulla et al. (2007) approach the same problem from a different angle. They use a reinforced process derived from the Generalized Polya urn scheme in constructing a bivariate prior on the space of distribution functions defined on the product space $\{1, 2, \ldots\} \times \{1, 2, \ldots\}$. Thus this approach may be suitable when the Bayesian prediction of future behavior of a bivariate observation based on the past observations is of interest. They extend their approach to the estimation of a multivariate survival function in Bulla et al. (2009). Yang et al. (2008) use mixtures of Polya trees in nonparametric Bayesian estimation of a bivariate density based on interval censored data.

Walker and Muliere (2003) considered a different model. Suppose we have data from two distributions which are known to be close but otherwise unknown. Their closeness is modeled by defining a parameter $\rho = \operatorname{corr}(F_1(A), F_2(A)) \geq 0$ for every set A in the domain. They describe a bivariate Dirichlet process model for $\varphi(F_1, F_2)$ in which marginal distributions for F_1 and F_2 are taken to be the same Dirichlet distribution and show how to find their posterior distributions. Their prior utilizes the Dirichlet-multinomial point process introduced by Lo (1986) (Sect. 1.15). The difficulty in describing the posterior completely is pointed out.

In contrast, it is relatively easy to construct bivariate tailfree and Polya tree processes. Mauldin et al. (1992) constructed one such process in terms of a prior guess of the unknown distribution. On the other hand, taking a cue from Lavine (1992), Mauldin et al. (1992) and Ferguson (1974), Phadia (2007) proposed a two-dimensional extension of Ferguson’s (1974) dyadic tailfree process and showed that given a random sample, the posterior distribution is also tailfree. It is then used in deriving bivariate estimators of a distribution (survival) function which are included in later sections. This extension is presented here.

1.16.1 Bivariate Tailfree Process

Recall the definition of a tailfree process presented in Sect. 1.10. The distribution of a random probability P on $(R, \mathcal{B})$ is said to be tailfree with respect to a sequence of nested partitions $\{\pi_m\}$ if there exists a family of nonnegative random variables $\{V_{m,B}; m = 1, 2, \ldots, B \in \pi_m\}$ such that (i) the families $\{V_{1,B}; B \in \pi_1\}, \{V_{2,B}; B \in \pi_2\}, \ldots$ are independent, and (ii) for every $m = 1, 2, \ldots,$ if $B_j \in \pi_j$, $j = 1, 2, \ldots, m$ is such that $B_1 \supset B_2 \supset \ldots \supset B_m$, then $P(B_m) = \prod_{j=1}^{m} V_{j,B_j}$. Thus to describe a bivariate tailfree process we need two things: a sequence of nested partitions $\Pi = \{\pi_m\}$ and a set of variables V_i’s. The construction is similar to Ferguson’s (1974) dyadic tailfree process of Sect. 1.10 except that each set in π_m is split into four instead of two at the $(m + 1)$-th level, and deals with Dirichlet rather than beta distributions. For analytical convenience a unit square $\mathfrak{X} = (0, 1] \times (0, 1]$ is taken in Phadia (2007) and a sequence of partitions is defined as follows.

The unit square $(0, 1] \times (0, 1]$ is denoted by B_0. It is subdivided into four symmetric subsquares B_1, B_2, B_3, B_4 of size $1/2$ and they are identified with suffixes 1, 2, 3 and 4 as those starting from the bottom left end side and moving in a clock-wise direction. Thus $B_1 = (0, \frac{1}{2}] \times (0, \frac{1}{2}]$, $B_2 = (0, \frac{1}{2}] \times (\frac{1}{2}, 1]$, $B_3 = (\frac{1}{2}, 1] \times (\frac{1}{2}, 1]$ and $B_4 = (\frac{1}{2}, 1] \times (0, \frac{1}{2}]$. Each subsquare is further divided into four symmetric subsquares, $B_{11}, \dots, B_{14}, \dots, B_{41}, \dots, B_{44}$, each of size 1/4 and the process is continued. Now let $\pi_0 = \{B_0\}$, $\pi_1 = \{B_1, B_2, B_3, B_4\}$, $\pi_2 = \{B_{11}, \dots, B_{14}, \dots, B_{41}, \dots, B_{44}\}$, $\dots$, $\pi_m = \{B_{c_1c_2\dots c_m}$, where $c_i = 1, 2, 3, 4$ for $i = 1, 2, \dots, m\}$, $m = 1, 2, \dots$. π_m will have 4^{m-1} 4-tuple subsquares of size 2^{-m}. Note that $B_{c_1} \supset B_{c_1c_2} \supset B_{c_1c_2c_3} \supset \dots$. Thus $\Pi = \{\pi_m; m = 0, 1, \dots\}$ forms a sequence of nested partitions.

The subsquares $B_{c_1c_2\dots c_m}$ may be identified by their bottom left end corners taken to be dyadic rationals (r, s), which can be expressed in terms of binary expansion of $\sum_{j=1}^{m} \epsilon_j.2^{-j}$ with $\epsilon_j = 0$ or 1 as $(.e_1e_2\dots e_m, .e'_1e'_2\dots e'_m)$, where e_i and e'_i take values 0 or 1. For example, the bottom left end corner of B_{32} is $(\frac{1}{2}, \frac{3}{4}) = (.10, .11)$. To further identify the 4 subsquares $B_{c_1c_2\dots c_m1}$, $B_{c_1c_2\dots c_m2}$, $B_{c_1c_2\dots c_m3}$, and $B_{c_1c_2\dots c_m4}$ of $B_{c_1c_2\dots c_m}$ with their bottom left end corners $(.e_1e_2\dots e_m\cdot, .e'_1e'_2\dots e'_m\cdot)$, we place at the blank places $(0, 0)$ for $c_{m+1} = 1$, $(0, 1)$ for $c_{m+1} = 2$, $(1, 1)$, for $c_{m+1} = 3$ and $(1, 0)$ for $c_{m+1} = 4$. For example: B_{231} is the set corresponding to a square with bottom left end corner $(.010, .100)$.

The collection of these squares forms a dense set in $(0, 1] \times (0, 1]$. It can be identified as follows. Let $E = \{1, 2, 3, 4\}$ and E_k be the set of sequences of numbers $i \in E$, of length k denoted by $\underline{c}_k = c_1c_2\dots c_k$. Let $E^* = \bigcup_k E_k$ be the set of all sequences of 1, 2, 3 and 4 of finite lengths. We shall denote the elements of E^* by $\underline{\epsilon}$. Thus π_n is a partition consisting of sets of the form $B_{\underline{\epsilon}}$ where $\underline{\epsilon} \in E_n$ and let $\{B_{\underline{\epsilon}1}, B_{\underline{\epsilon}2}, B_{\underline{\epsilon}3}, B_{\underline{\epsilon}4}\}$ be a further partition of $B_{\underline{\epsilon}}$. Thus $\bigcup_k E_k$ generates $\sigma((0, 1]^2)$.

Now we proceed to define the family of random variables $V_{m,B}$. Then the random probability P will be defined via these independent families. Set $\{V_{1,B} = Z_{c_1} = P(B_{\underline{c}_1}|B_0)$ for $B = B_{c_1} \in \pi_1, c_1 = 1, 2, 3, 4\}$, $\{V_{2,B} = Z_{\underline{c}_2} = P(B_{\underline{c}_2}|B_{\underline{c}_1})$ for $B = B_{\underline{c}_2} \in \pi_2, c_i = 1, 2, 3, 4, i = 1, 2\}$, $\dots$, $\{V_{m,B} = Z_{\underline{c}_m} = P(B_{\underline{c}_m}|B_{\underline{c}_{m-1}})$ for $B = B_{\underline{c}_m} \in \pi_m, c_i = 1, 2, 3, 4$ for $i = 1, 2, \dots, m\}, \dots$. We take these families to be independent between levels, $m = 1, 2, \dots$. This way for an arbitrary set $B \in \pi_m$, $P(B) = P(B_{c_1c_2\dots c_m})$, is the product of all the variables associated with the path in the tree from $[0, 1]^2$ to $B_{c_1c_2\dots c_m}$ so that $P(B_{c_1c_2\dots c_m}) = \prod_{i=1}^{m} Z_{c_1c_2\dots c_i}$, $c_i \in \{1, 2, 3, 4\}$, $i = 1, 2, \dots, m$. For example, if $B_{2143} \in \pi_4$, $P(B_{2143}) = Z_2Z_{21}Z_{214}Z_{2143}$. The random probability P is defined through the joint distribution of $P(B_{c_1c_2\dots c_m})$ by assigning suitable distributions to Z's. Since the sets $B_{\underline{c}_m}$ are decreasing to an empty set $\emptyset$, we should therefore have $P(B_{\underline{c}_m\underbrace{ii\dots i}})$ going to 0 for $i \in \{1, 2, 3, 4\}$. Thus the choice of distributions of Z's should be such that $P(B_{\underline{c}_m\underbrace{ii\dots i}\cdot\cdot}) = P(B_{\underline{c}_m}) \times \prod Z_{\underline{c}_m\underbrace{ii\dots i}\cdot\cdot} \xrightarrow{\text{a.s.}} 0$.

When P is extended to be defined over the algebra of sets generated by the squares, P will be σ-additive. Finally, it is extended in the usual manner to a unique probability defined on the class of Borel sets on $(0, 1] \times (0, 1]$. This will yield a random probability P on $((0, 1]^2, \sigma((0, 1]^2))$. The distribution of P will be tail-free with respect to the sequence of partitions, Π. Now Theorem 1.17 is applicable.

Thus if $X_1, X_2, \ldots, X_n$ is a sample from P, then the posterior distribution of P given $X_1, X_2, \ldots, X_n$ is also tailfree w.r.t. $\{\pi_m\}$. The unit square may be replaced by $(0, T] \times (0, T]$ for a finite T.

Recall that in the definition of a Polya tree prior (Lavine 1992), all (between as well as within partitions) the pairs of variables Z's are assumed to be independent each having a beta distribution. Similarly, in the bivariate case we may assume the 4^{m-1}, 4-tuple vectors $(Z_{\underline{c}_{m-1,1}}, \ldots, Z_{\underline{c}_{m-1,4}})$ at level m to be mutually independent each having a Dirichlet distribution with parameters $(\alpha_{\underline{c}_{m-1,1}}, \ldots, \alpha_{\underline{c}_{m-1,4}})$ for some nonnegative real numbers α's yielding a *bivariate Polya tree*. Although we take here the Dirichlet distribution in place of beta distribution, for any specific values of $c_1 c_2 \ldots c_m$, $Z_{\underline{c}_m}$ will have a beta distribution. For example, $Z_{2143} \sim Be(\alpha_{2143}, \alpha_{2141} + \alpha_{2142} + \alpha_{2144})$. To ensure that $P(B_{\underline{c}_m}) \prod Z_{\underline{c}_m \underbrace{ii\ldots i}..} \xrightarrow{\text{a.s.}} 0$, we place the following condition on α's.

$$\sum_{j=0}^{\infty} \frac{\gamma_{\underline{c}_m \underbrace{ii\ldots i}_{j}} - \alpha_{\underline{c}_m \underbrace{ii\ldots i}_{j} l}}{\gamma_{\underline{c}_m \underbrace{ii\ldots i}_{j}}} = \infty, \quad i, l \in \{1, 2, 3, 4\} \quad \text{and} \tag{1.147}$$

$$\gamma_{\underline{c}_m \underbrace{ii\ldots i}_{j}} = \sum_{l=1}^{4} \alpha_{\underline{c}_m \underbrace{ii\ldots i}_{j} l}.$$

The posterior distributions of the Z-vectors is again independent each with a Dirichlet distribution and the parameters α's of the posterior distributions get updated.

$$\begin{aligned}(Z_{\underline{c}_m 1}, Z_{\underline{c}_m 2}, Z_{\underline{c}_m 3}, Z_{\underline{c}_m 4}) \mid X_1, X_2, \ldots, X_n \\ \sim D(\alpha_{\underline{c}_m 1} + N_{\underline{c}_m 1}, \ldots, \alpha_{\underline{c}_m 4} + N_{\underline{c}_m 4}),\end{aligned} \tag{1.148}$$

where $N_{\underline{c}_m i}$ = the number of X_j's that fall in the set $B_{\underline{c}_m i}$.

By appropriate choice of all the parameters $\{\alpha_{c_1}\}, \{\alpha_{c_1 c_2}\}, \ldots$, it is possible, like in the univariate case, to obtain tailfree processes that are discrete, that are continuous singular, or that are absolutely continuous with probability one (Ferguson 1974).

It is possible to define mixtures of bivariate Polya trees in the same way as was defined in the univariate case by Lavine (1992). This probably opens up the possibility of further applications of Polya trees in modeling certain type of data.

Chapter 2
Inference Based on Complete Data

2.1 Introduction

In the statistical problem of nonparametric Bayesian analysis we have a random probability $\mathcal{P}$ belonging to Π and having a particular prior distribution. Given $\mathcal{P} = P$, we also have a random sample $X_1, \ldots, X_n$, which are iid P taking values in χ. Based on the sample, our objective is to estimate a function $\phi(P)$ of P, with respect to a certain loss function. Most of the applications presented in this chapter use the Dirichlet process prior or its variants—Dirichlet Invariant process and mixtures of Dirichlet processes. In Chap. 3 while dealing with censored data, we use other priors such as the neutral to the right processes which are more suited to such data, but obviously they are applicable in the uncensored data case as well.

In this chapter, first we will deal primarily with estimation problems and thereafter we will present hypothesis testing and other applications briefly. We will consider the distribution function (CDF) or its functionals. Since the Dirichlet process prior is conjugate, the strategy will be to obtain first the Bayes estimator of ϕ for the no sample problem and then to update the parameter(s) of the prior to obtain Bayes estimator for any sample size n. Through out this chapter we assume that we have a random sample $X_1, \ldots, X_n$ from an unknown distribution function F (corresponding to P) defined on the real line. In the case of two sample problem, we will have a second sample $Y_1, \ldots, Y_n$ from another distribution function, say, G. Both samples will be assumed to be independent. The loss functions used are a weighted (integral) squared error loss L_1 for the distribution function and a squared error loss L_2 for its functionals, where

$$L_1(F, \widehat{F}) = \int \big(F(t) - \widehat{F}(t)\big)^2 dW(t); \qquad L_2(\varphi, \widehat{\varphi}) = (\varphi - \widehat{\varphi})^2, \tag{2.1}$$

with W being a given weight function or a finite measure on $(R, \mathcal{B})$.

Through out this and the next chapter, we will denote the samples by bold letters, such as $\mathbf{X} = (X_1, \ldots, X_n)$, the sample distribution by $\widehat{F}_n(t)$ and the Bayes estimator with respect to the Dirichlet prior $\mathcal{D}(\alpha)$, by $\widehat{F}_\alpha$. Additionally, we let

E.G. Phadia, *Prior Processes and Their Applications*,
DOI 10.1007/978-3-642-39280-1_2, © Springer-Verlag Berlin Heidelberg 2013

$\overline{\alpha}(\cdot) = \alpha(\cdot)/\alpha(R)$, $M = \alpha(R)$ and $p_n = \alpha(R)/(\alpha(R) + n)$. In some applications, we use $\mathcal{X}$ instead of R.

The topics in this chapter are organized as follows:

1. Estimation of a distribution function.
2. Tolerance region and Confidence bands.
3. Estimation of functionals of a distribution function.
4. Other applications
5. Bivariate distribution function.
6. Estimation of a function of P.
7. Two sample problems
8. Hypothesis Testing

Under these headings, applications to sequential estimation, empirical Bayes, linear Bayes, minimax estimation, bioassay, and other applications will be presented. The beauty of the Dirichlet process is that most of the results are in closed forms. Also, once the no-sample problem is solved, all that is needed to solve the problem for any sample size is to update the parameter of the Dirichlet process. This strategy is used repeatedly. Needless to say that many of the problems discussed here could also be solved by using other priors, such as processes neutral to the right, Polya trees, beta-Stacy, etc., although closed form of the results may not be guaranteed. Since the Dirichlet process is inadequate in handling problems such as density estimation, there has been recently intense activity in using Polya trees and mixtures of Polya trees in such problems. A brief discussion on these efforts is included in Sect. 2.5.4.

2.2 Estimation of a Distribution Function

In this section the Bayesian estimation of a distribution function with respect to the Dirichlet process and related prior processes is presented. Also included are the empirical Bayes, sequential and minimax estimation procedures.

2.2.1 Estimation of a CDF

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$, F defined on $(R, \mathcal{B})$. The objective is to estimate F based on $\mathbf{X}$ under the loss function L_1 and prior $\mathcal{D}(\alpha)$. For each t, $F(t) \sim Be(\alpha(-\infty, t], \alpha(t, \infty))$. The risk is given by $\mathcal{E}(L(F, \widehat{F})) = \int \mathcal{E}(F(t) - \widehat{F}(t))^2 dW(t)$. The Bayes estimate of F for the no-sample problem is the posterior mean $\widehat{F}(t) = \mathcal{E}(F(t)) = F_0(t) = \alpha(-\infty, t]/\alpha(R)$, where the expectation is taken with respect to $\mathcal{D}(\alpha)$. By Theorem 1.1 of Sect. 1.2, we have $F|\mathbf{X} \in \mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{X_i})$. Therefore, for a sample of size n, the expectation is taken with respect to $\mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{X_i})$, and the Bayes estimator is $\widehat{F}(t) = \mathcal{E}(F(t) \mid X_1, \ldots, X_n)$ obtained as (Ferguson 1973)

$$\widehat{F}_\alpha(t) = \widehat{F}(t \mid X_1, \ldots, X_n) = \frac{\alpha(-\infty, t] + \sum_{i=1}^n \delta_{X_i}(-\infty, t]}{\alpha(R) + n}$$
$$= p_n \cdot F_0(t) + (1 - p_n) \cdot \widehat{F}_n(t), \quad \text{say}, \tag{2.2}$$

where $\widehat{F}_n(t) = \frac{1}{n} \cdot \sum_{i=1}^n \delta_{X_i}(-\infty, t]$, the empirical distribution function of the sample. Thus the Bayes rule $\widehat{F}_\alpha$ may be interpreted as a mixture of the prior guess F_0 and the empirical distribution function with respective weights, p_n and $1 - p_n$. At the same time, F_0 can be interpreted as the 'center' around which the Bayes estimate resides. Robustness of this estimator is discussed in Hannum and Hollander (1983a, 1983b).

Remark 2.1 $M = \alpha(R)$ may be interpreted as a precision parameter or the prior sample size (Ferguson 1973). As $\alpha(R) \to \infty$, $\widehat{F}_\alpha$ reduces to the prior guess F_0 at F. On the other hand, if $\alpha(R) \to 0$, the Bayes estimator reduces to the sample distribution function and hence it could be said that it corresponds to providing no information. However, Sethuraman and Tiwari (1982) take issue with this interpretation. For finite $\alpha_0, \alpha_r, r \geq 1$ on R, they show that if, along with the sequence $\alpha_r(R) \to 0$, we have $\text{Sup}_A |\overline{\alpha}_r(A) - \overline{\alpha}_0(A)| \to 0$ as $r \to \infty$, $A \in \mathcal{B}$, then $\mathcal{D}(\alpha_r) \to \delta_{Y_0}$, where Y_0 has the distribution $\overline{\alpha}_0$. This means that the information about P is that it is a probability measure concentrated at a particular point in R, and the point is selected according to $\overline{\alpha}_0$. This is a definite information about P and it's discreteness.

2.2.2 Estimation of a Symmetric CDF

In the previous section, F was an arbitrary distribution function. Suppose now we wish to estimate F which is symmetric about a known point η. This suggests that the space of all distribution functions be restricted to the symmetric distributions only. So assume F to be distributed according to the Dirichlet Invariant process (Sect. 1.3), that is, $F \in \mathcal{DGI}(\alpha)$, $\mathcal{G} = \{e, g\}$ with $e(x) = x$, $g(x) = 2\eta - x$. Then the Bayes estimate of F under the loss function L_1 is (Dalal 1979a)

$$\widehat{F}_{\alpha\eta}(t \mid X_1, \ldots, X_n) = \frac{\alpha(-\infty, t] + (1/2)\sum_{i=1}^n (\delta_{X_i}(-\infty, t] + \delta_{2\eta - X_i}(-\infty, t])}{\alpha(R) + n}$$
$$= p_n \cdot F_0(t) + (1 - p_n) \cdot \widehat{F}_{sn}(t), \tag{2.3}$$

where $\widehat{F}_{sn}(t)$ is η-symmetrized version of the empirical distribution. This is an analog of the Bayes estimator $\widehat{F}_\alpha$.

The Bayes estimator of F with respect to certain other prior processes will be presented when dealing with censored data in Chap. 3.

2.2.3 Estimation of a CDF with MDP Prior

Let $G(\theta)$ stand for a random distribution function selected from a mixture of Dirichlet processes (Sect. 1.4) with index space $U = R$, parameter space $\Theta = R$, observation space also R and mixing distribution H. That is $G \in \int_R D(\alpha_u)dH(u)$ and let $\theta_1, \dots, \theta_n \overset{iid}{\sim} G$, and given θ_i let $X_{i1}, \dots, X_{im_i}$ be a sample of size m_i from $F_{\theta_i}(x)$, $i = 1, \dots, n$.

The Bayes estimate of G under the L_1 loss function is given by $\widehat{G} = \mathcal{E}(G|\theta_1, \dots, \theta_n)$ if θ_i's are observed directly, and $\widehat{G} = \mathcal{E}(G|X_{i1}, \dots, X_{im_i})$ if X_{ij}'s are observed. One can use the formula given under property 4 of Sect. 1.4 to evaluate the former, and the formula given under property 5 to evaluate the latter. Antoniak (1974) has illustrated computational procedures by taking examples of small sample of size $n = 2$. As an example, he takes the transition measure $\alpha_u(\cdot)/\alpha(R) = N(u, \sigma^2)$, mixing distribution $H = N(0, \rho^2)$ and sampling distribution $F_\theta = N(\theta, \tau^2)$, and obtains the expression for $\widehat{G}$ in a closed form. For larger sample size, the evaluations are difficult. However, computational algorithms are developed for handling such problems in Kuo (1986b), and are described in many publications that have appeared since her paper. See for example books by Dey et al. (1998) and Ibrahim et al. (2001).

In an effort to compromise between parametric and purely nonparametric models, Doss (1994) investigates prior distributions for F which give most of their mass to a "small neighborhood" of an "entire" parametric family. In other words, he considers the situation where a parametric family H_θ, $\theta \in \Theta \subset R^p$ is specified. Thus, a prior on F is placed as follows. First choose θ according to some prior υ, then choose F from $\mathcal{D}(\alpha(R)H_\theta)$ with specified $\alpha(R) > 0$. This leads to the mixture of Dirichlet processes priors, $F \in \int \mathcal{D}_{\alpha(R)H_\theta}\upsilon(d\theta)$. While this formulation encounters the same computational difficulties, it allows him to consider a more general set up when instead of exact values of X_i ($\sim F$), it is only known that $X_i \in A_i \subset R$. Thus we may have $A_i = \{x_i\}$ if X_i is an exact observation, and $A_i = (c_i, \infty)$ if X_i is censored on the right by c_i. The task is to obtain the posterior distribution of F given the data. Doss develops an algorithm for generating a random distribution function from this conditional posterior distribution. Further details may be found in his paper.

2.2.4 Empirical Bayes Estimation

In the Sect. 2.2.2 we derived the Bayesian estimator of F assuming a Dirichlet process prior with parameter α. It was assumed there that α is known via a known prior guess F_0 of F, and the total mass M. If this is not the case, we need to estimate F_0 or M or both. This can be done via the *empirical Bayes* (EB) approach which is described now. The efficacy of the empirical Bayes estimator is judged by a criterion called '*asymptotic optimality*': An empirical Bayes estimator is said to be *asymptotically optimal* relative to a class of Dirichlet process priors if the Bayes risk of the

EB estimator given α converges to the Bayes risk of the Bayes estimator for all α. This being a weak criterion, generally, the rate of convergence is also indicated.

Since $\mathcal{E}[F(t)] = F_0(t)$ and $\text{var}(F(t)) = F_0(t)(1 - F_0(t))/(M+1)$, the parameter α is then expressed as $\alpha(\cdot) = MF_0(\cdot)$, which provides interpretation of M as a 'precision' or 'accuracy' or 'uncertainty' parameter and specification of F_0 implies the random distribution function is centered around F_0. For this reason, it is felt that the empirical Bayes method, where the sample data is used for identifying F_0, is better rather than specifying some arbitrary F_0, whose validation may or may not be ascertained.

In the empirical Bayes framework, we are currently at the $(n+1)$-th stage of an experiment, and information is available not only from the current stage, but also from the n previous stages. Thus we have a sequence of pairs $(P_i, \mathbf{X}_i)$, $i = 1, 2, \ldots, n+1$ of independent random elements, where P_i's are probability measures on $(R, \mathcal{B})$ having a common Dirichlet process prior $\mathcal{D}(\alpha)$. Given $P_i = P$, $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{im_i})$ is a random sample of size m_i from P. The task is to estimate the distribution function corresponding to P at the $(n+1)$-th stage or its functional. The strategy is to use the information provided by the previous n stages in estimating the parameters of the prior at the $(n+1)$-th stage. This approach will be used in estimating the distribution function, the mean, and in general, any estimable parameters of degree 2 or 3.

Remark 2.2 In many hierarchical modeling, the parameters at intermediate stages are assumed to have certain distributions with some hyper parameters. It is fine if there are valid justifications for such assignments. However, in absence of such information it is judged that the empirical Bayes methods may offer better solution since here the observed data itself is used to provide information on unknown parameters.

Empirical Bayes Estimation of a CDF Let $F_1, F_2, \ldots, F_{n+1}$ be $n+1$ distribution functions on the real line, and for $i = 1, 2, \ldots, n+1$, let $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{im_i})$ be a sample of size m_i from F_i. We assume each F_i to have a common Dirichlet process prior, $\mathcal{D}(\alpha)$. Our prior information is incorporated through F_0 and M. As before $\widehat{F_j}(t)$ is the sample distribution function of $\mathbf{X}_j$ and $p_j = \alpha(R)/(\alpha(R) + m_j)$, $j = 1, \ldots, n+1$. Consider the estimation of $F_{n+1}(t)$ based on $\mathbf{X}_1, \ldots, \mathbf{X}_{n+1}$. The Bayes estimator of F_{n+1} at the $(n+1)$-th stage with respect to the prior $\mathcal{D}(\alpha)$ and the loss function

$$L(F_{n+1}, \widetilde{F}_{n+1}) = \int \big(F_{n+1}(t) - \widetilde{F}_{n+1}(t)\big)^2 dW(t) \tag{2.4}$$

is from Sect. 2.2.1 (suppressing the dependence on α),

$$\widetilde{F}_{n+1}(t) = p_{n+1}F_0 + (1 - p_{n+1})\widehat{F}_{n+1}(t). \tag{2.5}$$

The Bayes risk of $\widetilde{F}_{n+1}(t)$ with respect to $D(\alpha)$, denoted by

$$r(\alpha) = \mathcal{E}_{\mathcal{D}(\alpha)}\mathcal{E}_{F_{n+1}}\big[L(F_{n+1}, \widetilde{F}_{n+1})\big],$$

is (Korwar and Hollander 1976)

$$r(\alpha) = r(\widetilde{F}_{n+1}, \alpha) = \frac{p_{n+1}}{\alpha(R)+1} \int F_0(t)\big(1 - F_0(t)\big) dW(t)$$
$$= \frac{p_{n+1}}{\alpha(R)+1} \sigma^2, \tag{2.6}$$

where $\sigma^2 = \int x^2 dF_0(x) - (\int x dF_0(x))^2$ is the variance of F_0.

If F_0 and M are known, $r(\alpha)$ can be evaluated completely. In the EB approach, we are able to estimate these parameters from the previous data and adjust this estimator, resulting in what is known as an *empirical Bayes* estimator.

Korwar and Hollander (1976), considered the case of α when M is known but F_0 is unknown. They estimated $F_0(t)$ by the average of first n sample distributions, $\frac{1}{n}\sum_{j=1}^{n} \widehat{F}_j(t)$, substituted this in (2.5) and proposed the following empirical Bayes estimator of F_{n+1}:

$$\overline{F}_{n+1}(t) = p_{n+1} \frac{1}{n} \sum_{j=1}^{n} \widehat{F}_j(t) + (1 - p_{n+1}) \widehat{F}_{n+1}(t). \tag{2.7}$$

They evaluated the Bayes risk of $\overline{F}_{n+1}(t)$ as

$$r(\overline{F}_{n+1}, \alpha) = \mathcal{E}\big[L(F_{n+1}, \overline{F}_{n+1})\big] = r(\alpha)\left[1 + \frac{p_{n+1}}{n^2} \sum_{j=1}^{n} \frac{1}{1 - p_j}\right], \tag{2.8}$$

where the expectation is taken with respect to $\mathcal{D}(\alpha)$ as well as $\mathbf{X}_1, \ldots, \mathbf{X}_{n+1}$. When the samples are of same size as m, $r(\overline{F}_{n+1}, \alpha)$ reduces to $r(\alpha)[1 + \alpha(R)/mn]$. Clearly, as $n \to \infty$, $r(\overline{F}_{n+1}, \alpha) \to r(\alpha)$ for any α. Thus they concluded that the estimator is asymptotically optimal and established the rate of convergence $O(n^{-1})$. Zehnwirth (1981) relaxed the assumption of M known in the case of equal sample size and estimated M in a clever way (to be described below) by the F-ratio statistic $\mathcal{F}_n$ in one-way analysis of variance based on $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and showed that the resulting estimator of F_{n+1} is also asymptotically optimal with the same rate of convergence, $O(n^{-1})$.

Note that the estimators of F_0 proposed by Korwar and Hollander and Zehnwirth were based only on the past data, but not the current data. Ghosh et al. (1989) modified these estimators to include the current data as well. Thus, it gives greater weight to the current data in estimating $F_{n+1}(t)$ than that in the Hollander and Korwar and Zehnwirth estimators, and yields smaller risk than those estimators.

When $\alpha(R)$ is known, their proposed empirical Bayes estimator of $F_{n+1}(t)$ turns out to be

$$\widetilde{F}^*_{n+1}(t) = p_{n+1} \widehat{F}_0(t) + (1 - p_{n+1}) \widehat{F}_{n+1}(t), \tag{2.9}$$

where $\widehat{F}_0(t) = \sum_{j=1}^{n+1}(1-p_j)\widehat{F}_j(t)/\sum_{j=1}^{n+1}(1-p_j)$, and the Bayes risk is

$$r\big(\widetilde{F}^*_{n+1},\alpha\big) = r(\alpha)\left[1+\frac{p_{n+1}}{\sum_{j=1}^{n+1}(1-p_j)}\right], \tag{2.10}$$

which converges to $r(\alpha)$ as $n \to \infty$, and hence it is asymptotically optimal. Comparing the risks of estimators with and without the use of the current data, it can be verified that $r(\widetilde{F}^*_{n+1},\alpha) - r(\alpha) \le r(\overline{F}_{n+1},\alpha) - r(\alpha)$ and hence an improvement is achieved by using the estimator $\widetilde{F}^*_{n+1}(t)$ over $\overline{F}_{n+1}$. In fact if the sample sizes are equal, $r(\widetilde{F}^*_{n+1},\alpha) - r(\alpha) = (n/(n+1))[r(\overline{F}_{n+1},\alpha) - r(\alpha)]$.

Observe that $\widetilde{F}^*_{n+1}(t)$ is a linear combination $\sum_{j=1}^{n+1} w^*_j \widehat{F}_j(t)$, with $w^*_j = p_{n+1}(1-p_j)/(\sum_{j=1}^{n+1}(1-p_j))$, $j = 1,\dots,n$, and $w^*_{n+1} = p_{n+1}(1-p_{n+1})/(\sum_{j=1}^{n+1}(1-p_j)) + (1-p_{n+1})$. Clearly $\sum_{j=1}^{n+1} w^*_j = 1$. This gives a clue for them to show that indeed the Bayes risk of $\widetilde{F}^*_{n+1}$ is smaller than the Bayes risk of any other estimator of the form $\sum_{j=1}^{n+1} w_j \widehat{F}_j$ with $\sum_{j=1}^{n+1} w_j = 1$. By taking different choices of w_j we can see that this class includes the following estimators. The choice of $w_j = p_m/n$ $(j = 1,\dots,n)$ and $w_{n+1} = 1 - p_m$, with $m_1 = \ldots = m_n = m$ and $p_m = \alpha(R)/(\alpha(R)+m)$, leads to Korwar and Hollander (1976) estimator $\overline{F}_{n+1}(t)$. Another possible choice of $w_j = 1/(n+1)$ for $j = 1,\dots,n+1$ which leads to the estimator $\sum_{j=1}^{n+1} \widehat{F}_j/(n+1)$ of F_{n+1}. Also, the usual MLE estimator of F_{n+1} is $\widehat{F}_{n+1}$ which is obtained when $w_{n+1} = 1$, and $w_1 = \ldots = w_n = 0$.

When $\alpha(R)$ is unknown, Zehnwirth (1981) proposed an estimator for $\alpha(R)$ based on a one-way ANOVA table using the past data $\mathbf{X}_1,\dots,\mathbf{X}_n$ for equal sample size m at each stage and proved it's consistency

$$m/(1-\mathcal{F}_n) \to \alpha(R) \quad \text{in probability as } n \to \infty. \tag{2.11}$$

Ghosh et al. (1989) provide an improvement over his estimator by including $\mathbf{X}_{n+1}$ as well in the $\mathcal{F}_n$ statistic. Let $\overline{X}_j = \sum_{i=1}^{m_j} X_{ji}/m_j$ be the mean of the sample values at j-th stage and $\overline{X} = \sum_{j=1}^{n+1} m_j \overline{X}_j / \sum_{j=1}^{n+1} m_j$ denote the overall mean. Define

$$\begin{aligned} MSW &= \sum_{j=1}^{n+1}\sum_{i=1}^{m_j}(X_{ji}-\overline{X}_j)^2 \Big/ \sum_{j=1}^{n+1}(m_j-1) \quad \text{and} \\ MSB &= \sum_{j=1}^{n+1}(\overline{X}_j-\overline{X})^2 \Big/ n, \end{aligned} \tag{2.12}$$

the usual within and between mean squares, respectively. Simple evaluations involving the Dirichlet process yield (see Ghosh et al. 1989)

$$E(MSW) = \big(\alpha(R)/\big(\alpha(R)+1\big)\big)\sigma^2 \tag{2.13}$$

$$E(MSB) = \big(\alpha(R)/\big(\alpha(R)+1\big)\big)\sigma^2 + \xi.\sigma^2/\big(n\big(\alpha(R)+1\big)\big), \tag{2.14}$$

where $\xi = \sum_{j=1}^{n+1} m_j - \sum_{j=1}^{n+1} m_j^2 / \sum_{j=1}^{n+1} m_j$. They proposed the following estimator of $\alpha(R)$.

$$\widehat{\alpha}^{-1}(R) = \max\{0, (MSB/MSW - 1)n/\xi\}, \tag{2.15}$$

which is shown to be strongly consistent under some mild conditions. (Note that the Zehnwirth's (1981) estimator of $\alpha(R)$ is based only on $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and had assumed $m_1 = \ldots = m_{n+1}$.) Substituting this estimator of $\alpha(R)$ in $p_j = (1 + m_j \alpha^{-1}(R))^{-1}$ they revise the estimate for F_0 as

$$\widetilde{F}_0(t) = \begin{cases} \sum_{j=1}^{n+1}(1-\widehat{p}_j)\widehat{F}_j(t)/\sum_{j=1}^{n+1}(1-\widehat{p}_j), & \text{if } \widehat{\alpha}^{-1}(R) \neq 0 \\ \sum_{j=1}^{n+1} \widehat{F}_j(t)/(n+1), & \text{if } \widehat{\alpha}^{-1}(R) = 0. \end{cases} \tag{2.16}$$

Finally for the case $\alpha(R)$ unknown, Ghosh et al. (1989) utilizing these estimators proposed $\widehat{\widehat{F}}_{n+1}$ as an improved empirical Bayes estimator of F_{n+1}, where

$$\widehat{\widehat{F}}_{n+1}(t) = \widehat{p}_{n+1}\widetilde{F}_0(t) + (1-\widehat{p}_{n+1})\widehat{F}_{n+1}(t), \tag{2.17}$$

and proved the asymptotic optimality of this estimator.

2.2.5 Sequential Estimation of a CDF

Ferguson (1982) derives the sequential estimator of F under the loss function L_1 and prior $\mathcal{D}(\alpha)$, with $F_0 = \overline{\alpha}$ as a specified distribution function on R. Then as noted earlier, $\mathcal{E}(F) = F_0$ and the posterior distribution of F, given the sample $X_1, \ldots, X_n$ from F, is $\mathcal{D}((M+n)\widehat{F}_\alpha)$, where $\widehat{F}_\alpha$, as before, is the Bayes estimator of F under L_1, and the minimum Bayes risk is

$$\int \mathrm{Var}\big(F(x)|\mathbf{X}\big)dW(x) = \big(1/(M+n+1)\big)\int_\alpha \widehat{F}_\alpha(x)\big(1-\widehat{F}_\alpha(x)\big)dW(x). \tag{2.18}$$

In sequential estimation we need a stopping rule and a terminal estimator. It is enough to find a stopping rule, since once we have the stopping rule, the terminal Bayes estimator is $\widehat{F}_\alpha$ itself. Ferguson discusses the k-stage look ahead rule which, at each stage stops or continues according to whether the rule is optimal among those taking at most k more observations stops or continues. There is a positive cost $c > 0$ to look for each additional observation. After observing $X_1, \ldots, X_n$, the 1-stage look ahead rule that he develops calls for stopping after the first n observations for which

$$\int_\alpha \widehat{F}_\alpha(1-\widehat{F}_\alpha)dW \leq c(M+n+1)^2. \tag{2.19}$$

Clearly the left hand side is bounded above by $W(R)/4$ and the right hand side increases with n. Ferguson argues that the 1-stage look-ahead rule eventually calls for stopping and bounds on the maximum sample size can be found. He provides justification for the optimality of this rule.

Ferguson also discusses the sequential estimation of the mean $\mu = \int x dF(x)$ under the squared error loss function L_2. Let $\mu_0 = \int x dF_0(x)$, the prior estimate of the mean and $\overline{X}_n = (1/n)\sum_1^n X_i$. The minimum conditional Bayes risk is given by $\operatorname{Var}(\mu|X_1,\ldots,X_n) = \sigma_n^2/(M+n+1)$, where σ_n^2 is the variance of the distribution $\widehat{F}_n$, which can be expressed as

$$\begin{aligned}\sigma_n^2 &= \int (x-\mu_n)^2 d\widehat{F}_n(x)\\ &= \left(M\sigma_0^2 + ns_n^2 + \frac{Mn}{M+n}(\overline{x}_n-\mu_0)^2\right)\Big/(M+n), \end{aligned} \tag{2.20}$$

with σ_0^2 as the variance of F_0 and $ns_n^2 = \sum_1^n (x_i - \overline{x}_n)^2$. Then his 1-stage rule is to stop after the first n observations for which $\sigma_n^2 \le c(M+n+1)^2$. He also provides further some modified stopping rules and discusses their usage. His paper may be referred to for more details.

Sequential approach from the Bayesian point of view is also used by Hall (1976, 1977) in treating search problems with random overlook probabilities having a Dirichlet or a mixture of Dirichlet processes. Clayton and Berry (1985) treat one-armed bandit problem and Clayton (1985) a sequential testing problem for the mean of a population.

2.2.6 Minimax Estimation of a CDF

One of the first non-Bayesian application of the Dirichlet process was contained in Phadia (1971), where a sequence of Dirichlet process priors was used in deriving the minimax estimator of an unknown F based on a sample of size n. The technique was first to find an equalizer rule given by

$$\widehat{F}_{mx}(t) = \frac{\sqrt{n}/2 + \sum_{i=1}^n \delta_{X_i}(-\infty, t]}{\sqrt{n}+n}.$$

Then a sequence of least favorable priors $\mathcal{D}(\alpha_k)$ was defined, where α_k was taken to be a finite measure giving equal weight $\sqrt{n}/2$ to points $\pm k$, k a nonnegative real number. Then it was shown that the Bayes risk of the Bayes estimator with respect to $\mathcal{D}(\alpha_k)$ converges to the risk of the above equalizer rule as $k \to \infty$. Thus it was established that the above estimator was minimax under the L_1 loss (minimax estimators for other loss functions were also obtained and in particular it was shown that the sample distribution function was minimax under a weighted quadratic loss function). However, at Ferguson's suggestion the results were simplified by taking a sequence of beta distribution priors (Phadia 1973).

2.3 Tolerance Region and Confidence Bands

In this section we first present the Tolerance region discussed by Ferguson in his 1973 paper, and then the construction of a confidence band as proposed by Breth (1978).

2.3.1 Tolerance Region

Ferguson (1973) treats the problem of deriving a tolerance region from a decision theoretic approach. Suppose we want to estimate the q-th quantile t_q of an unknown distribution F on the real line by an upper tolerance point a under the loss function

$$L(p,a) = pP\big((-\infty,a]\big) + qI_{(a,\infty)}(t_q), \tag{2.21}$$

where p is a constant, $0 < p < 1$. If t_q is known exactly, L is minimized by choosing $a = t_q$. But if t_q is not known precisely, then we need to minimize the Bayes risk with respect to the $\mathcal{D}(\alpha)$ given by

$$\mathcal{E}\big(L(p,a)\big) = pu + q\int_0^q \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)} z^{uM-1}(1-z)^{(1-u)M-1}dz, \tag{2.22}$$

where u represents $F_0(a)$, and $M = \alpha(R)$ as before. Let $u = f(p,q,M)$ denote the point at which the minimum occurs. Then the Bayes rule for the no sample problem is given by $a = f(p,q,\alpha(R))$-th quantile of F_0. For a sample $X_1,\dots,X_n$ of size n, the Bayes rule therefore is given by

$$\widehat{a}_n(\mathbf{X}) = f\big(p,q,\alpha(R)+n\big)\text{-th quantile of } \widehat{F}_n. \tag{2.23}$$

2.3.2 Confidence Bands

In the classical theory, confidence bands (F_L, F_U) for an unknown distribution function F are constructed for a given confidence level $1-\nu$, such that $\mathcal{P}(F_L \le F \le F_U) = 1-\nu$. Here F is considered to be fixed while F_L and F_U are random, they being functions of ordered sample values. In the Bayesian context, it is the other way around—F is considered to be random and F_L and F_U are fixed and determined in terms of the prior and posterior probabilities. Breth (1978) treats this problem.

Definition 2.3 (Breth) Suppose $F \in \mathcal{D}(\alpha)$. Then if $\mathcal{P}\{F_L(t) \le F(t) \le F_U(t)$ for all $t\} = \nu_1$ (ν_2) is a prior (posterior) probability, the functions $F_L(t)$ and $F_U(t)$ constitute the boundaries for a fixed region within which the random distribution function lies with prior (posterior) probability ν_1 (ν_2). $(F_L(t), F_U(t))$ are defined to be a pair of Bayesian confidence bands for the random distribution function F with prior (posterior) probability ν_1 (ν_2).

Let m be a fixed positive integer and for $i = 1, 2, \ldots, m$ define u_i and v_i such that $u_i < v_i$ for all i and $0 = u_0 \leq u_1 \leq \ldots \leq u_m < 1$, $0 < v_1 \leq v_2 \leq \ldots \leq v_{m+1} = 1$. Further, let $I(x) = 1$ if $x \geq 0$ and 0 otherwise, and $J(x) = 1$ if $x > 0$ and 0 otherwise. For $-\infty = t_0 < t_1 < t_2 < \ldots < t_m < t_{m+1} = \infty$, define $F_L(x) = \sum_{i=1}^{m}(u_i - u_{i-1})I(x - t_i)$ and $F_U(x) = v_1 + \sum_{i=1}^{m}(v_{i+1} - v_i)J(x - t_i)$. Also, for $a > 0$, let $\alpha(R) = a + 1 > 0$ and $\alpha(t)/\alpha(R)$ be a distribution function.

It is clear that $\mathcal{P}\{F_L(t) \leq F(t) \leq F_U(t) \text{ for all } t\} = \mathcal{P}\{u_j \leq F(t_j) \leq v_j \text{ for } j = 1, \ldots, m\}$. Therefore, to be able to calculate the probabilities of this type, it suffices to be able to calculate general rectangular probabilities (Steck 1971) over the ordered Dirichlet distribution, since $(F(t_1), \ldots, F(t_m)) \sim D(a_1, \ldots, a_m; a_{m+1})$ with $a_j = \alpha(t_j) - \alpha(t_{j-1})$, $j = 1, \ldots, m$. To calculate the boundaries with respect to the posterior probability, replace α by $\alpha^* = \alpha + n\widehat{F}_n$, where $\widehat{F}_n$ is the sample distribution function for the sample of size n.

It should be noted that as in the classical theory, there are many pairs of Bayesian confidence bands for F with the same probability content $1 - \nu$, say. In practice, a particular pair must be chosen to express quantitative confidence in F.

Breth (1978) uses recursive methods for computing $\mathcal{P}\{u_j \leq F(t_j) \leq v_j \text{ for all } j\}$ for fixed numbers $\{u_j\}$, $\{v_j\}$ and $\{t_j\}$ when F is a Dirichlet process. For details on calculations that are needed in practical applications, one can refer to his paper. In a follow up paper he (Breth 1979) discusses construction of Bayesian confidence intervals for quantiles and the mean, and also treats Bayesian tolerance intervals. The complexity in numerical calculation is evident. If α is not stipulated a priori, it can be estimated (see Korwar and Hollander 1973).

In this connection it is worth mentioning that in non-Bayesian context, Phadia (1974) constructed the best invariant one and two-sided confidence bands for an unknown continuous distribution function. They were invariant under the group $\mathcal{G}$ of transformations $g_\phi(y_1, \ldots, y_n) = (\phi(y_1), \ldots, \phi(y_n))$, where ϕ is a continuous, strictly increasing function from R onto R. The confidence bands were step functions taking jumps at the ordered sample values. For a given confidence level, the values of jumps were calculated as a minimization problem using Steck's result.

Simulation Method Neath and Bodden (1997) also constructed $(1 - \gamma)100\ \%$ Bayesian confidence bands F_L and F_U by using a simulation method. Let P be a random probability measure having a mixture of Dirichlet processes as prior distribution. In other words, $\theta \sim G$, $P|\theta \in \mathcal{D}(\alpha_\theta)$. Let F be a distribution function corresponding to P. Given a random sample from F, first a value of θ is obtained from the posterior distribution $G_{\mathbf{X}}$ of θ, given $\mathbf{X}$. The posterior distribution of F given the data and θ is $\mathcal{D}(\alpha_\theta + \sum_1^n \delta_{x_i})$. However, this distribution is analytically intractable. Therefore, in constructing the confidence bands, they treat simulated sample of distribution functions $F_1, \ldots, F_N$ as the actual distributions and choose F_L and F_U such that

$$\frac{1}{N}\sum_{i=1}^{N} I\left\{F_L(t) \leq F_i(t) \leq F_U(t), t \in R\right\} \geq 1 - \gamma. \tag{2.24}$$

In choosing the 'best' bounds, the following two criteria are used.

$$\begin{aligned}&\text{(i)}\quad \min\Big\{\max_t\big[F_U(t)-F_L(t)\big]\Big\}\quad\text{and}\\ &\text{(ii)}\quad \min\Big\{\int\big[F_U(t)-F_L(t)\big]dW(t)\Big\}.\end{aligned}\tag{2.25}$$

The minimum is taken over all functions F_L and F_U such that (2.24) is satisfied. For the process of choosing F_L and F_U, they give two algorithms and discuss their implementation. They also provide a numerical example to illustrate the procedure.

Bayesian Bootstrap Method Hjort (1985) uses a Bayesian bootstrap method to construct confidence intervals for a function $\theta(F)$ of an unknown F. Let $F \in \mathcal{D}(\alpha)$ and $X_1,\ldots,X_n \overset{iid}{\sim} F$. Then $F|\mathbf{X} \in \mathcal{D}(\alpha + \sum_1^n \delta_{x_i})$ which can be written as $\mathcal{D}(MF_0 + n\widehat{F}_n)$ with $F_0 = \alpha/M$, $M = \alpha(R)$. Let $G(t) = P(\theta(F) \le t|\widehat{F}_n)$. We need to find θ_L and θ_U such that $\mathcal{P}(\theta_L \le \theta(F) \le \theta_U) = 1 - 2\upsilon$, say. Thus $G^{-1}(\upsilon)$ and $G^{-1}(1-\upsilon)$ are the natural choices for θ_L and θ_U, respectively with $G^{-1}(p) = \inf\{t : G(t) \ge p\}$.

2.4 Estimation of Functionals of a CDF

In this section we discuss various applications in which Bayesian estimators of certain functionals such as the mean, median, variance, etc. are derived using the Dirichlet process priors.

2.4.1 *Estimation of the Mean*

Ferguson (1973) considered the Bayesian estimation of the mean $\mu = \int xdP(x)$ with respect to the Dirichlet process prior and under the squared error loss L_2. It is assumed that α has a finite first moment. The Bayes rule for the no-sample problem is the mean of μ, say, μ_0 which, by property 1 of Sect. 1.2.2, is $\widehat{\mu} = \mathcal{E}_{\mathcal{D}(\alpha)} \int xdP(x) = \int xd\alpha(x)/\alpha(R) = \mu_0$. The Bayes rule for a sample of size n therefore is obtained by updating the parameter α to $\alpha + \sum_{i=1}^n \delta_{X_i}$ and is given by

$$\begin{aligned}\widehat{\mu}_{\alpha n}(\mathbf{X}) &= \big(\alpha(R)+n\big)^{-1}\int xd\left(\alpha(x)+\sum_{i=1}^n \delta_{X_i}(x)\right)\\ &= p_n\mu_0 + (1-p_n)\overline{X}_n,\end{aligned}\tag{2.26}$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean. The Bayes estimator thus is, like $\widehat{F}_\alpha$, a convex combination of the prior guess at μ, namely μ_0, and the sample mean.

As $\alpha(R) \to 0$, $\widehat{\mu}_{\alpha n} \to \overline{X}_n$, and as $\alpha(R) \to \infty$, $\widehat{\mu}_n \to \mu_0$. The Bayes risk of $\widehat{\mu}_n$ is $r(\alpha) = p_n\sigma^2/(\alpha(R) + n)$. Alternatively, the estimator can also be obtained by taking $g(x) = x$ in property 9 of Sect. 1.2.2. More generally, let Z be a measurable real valued function defined on $(R, \mathcal{B})$ and $\theta = \int ZdP$. If $P \in \mathcal{D}(\alpha)$ and $\theta_0 = \int Zd\alpha/\alpha(R) < \infty$, then the Bayes estimator of θ under the loss L_2 is given by

$$\widehat{\theta}_{\alpha n}(\mathbf{X}) = p_n\theta_0 + (1 - p_n)\frac{1}{n}\sum_{i=1}^{n} Z(X_i). \tag{2.27}$$

Yamato (1984) showed that the mean μ is distributed symmetrically about a constant θ if the measure α is symmetric about θ and $\int |x|d\alpha(x) < \infty$.

If α and $\alpha(R)$ are unknown, the empirical Bayes method can be used.

Empirical Bayes Estimation of the Mean This can be dealt with in the same manner as the distribution function in Sect. 2.2.4 and the same notations will be used here as well. The Bayes estimator with respect to $\mathcal{D}(\alpha)$ at the $(n + 1)$-th stage is given by

$$\widehat{\mu}_\alpha = p_{n+1}\mu_0 + (1 - p_{n+1})\sum_{i=1}^{m_{n+1}} X_{n+1,i}/m_{n+1}. \tag{2.28}$$

The Bayes risk of $\widehat{\mu}_\alpha$ is given by $r(\alpha) = p_{n+1}\sigma^2/(\alpha(R) + 1)$. For the empirical Bayes approach, μ_0 is estimated from the first n samples by Korwar and Hollander (1976) and the resulting estimator $\widehat{\mu}_n$ has the Bayes risk as $r(\widehat{\mu}_n, \alpha) = (1 + \alpha(R)/\sum_{i=1}^{n+1} m_i)r(\alpha)$. Ghosh et al. (1989) estimates μ_0 from the past as well as current sample data as $\widehat{\mu}_0 = \sum_{j=1}^{n+1}(1 - p_j)\overline{X}_j/\sum_{j=1}^{n+1}(1 - p_j)$ and plugs in $\widehat{\mu}_\alpha$. The resulting estimator is $\widehat{\mu}_{n+1}$ and it's Bayes risk is

$$r(\widehat{\mu}_{n+1}, \alpha) = r(\alpha) + p_{n+1}^2\sigma^2 \Big/ \sum_{j=1}^{n+1}(1 - p_j). \tag{2.29}$$

They have shown that $\widehat{\mu}_{n+1}$ is asymptotically optimal and has a smaller Bayes risk than the estimator proposed by Korwar and Hollander (1976). Again, if $\alpha(R)$ is unknown, it can be estimated as indicated in Sect. 2.2.4.

In the context of a finite population of size N, Binder (1982) considered the task of Bayes estimation of the population mean $\sum_{i=1}^{N} X_i/N$, where $X_1, \ldots, X_N$ are population values, by assuming that there is a super population P with prior $\mathcal{D}(\alpha)$, and given P these values are iid P.

Ghosh et al. (1989) have also considered the empirical Bayes estimation of the finite population distribution function. Tiwari and Lahiri (1989) have treated the Bayes and empirical Bayes estimation of variances from stratified samples and studied the risk performance of the empirical Bayes estimators.

2.4.2 Estimation of a Variance

Consider now the task of estimating the variance of an unknown probability distribution P. If α has a finite second moment, then the variance of P is defined by

$$\operatorname{Var} P = \int x^2 dP(x) - \left(\int x dP(x)\right)^2, \tag{2.30}$$

which is a random variable. Ferguson (1973) obtained the Bayes estimator under the squared error loss L_2 assuming the Dirichlet process prior. The Bayes estimator of $\operatorname{Var} P$ for the no-sample problem is the posterior mean

$$\begin{aligned}\mathcal{E}\operatorname{Var} P &= \mathcal{E}\int x^2 dP(x) - \mathcal{E}\left(\int x dP(x)\right)^2 = \left(\sigma_0^2 + \mu_0^2\right) - \left(\frac{\sigma_0^2}{\alpha(R)+1} + \mu_0^2\right) \\ &= \frac{\alpha(R)}{\alpha(R)+1}\sigma_0^2,\end{aligned} \tag{2.31}$$

where μ_0 is as defined above in Sect. 2.4.1 and $\sigma_0^2 = \int x^2 d\alpha(x)/\alpha(R) - \mu_0^2$ is the variance of F_0.

For a sample of size n, the Bayes rule is therefore obtained by replacing the parameter α by $\alpha + \sum_{i=1}^n \delta_{X_i}$. After some simplification and rearrangement, we get the Bayes estimator of $\operatorname{Var} P$ as

$$\begin{aligned}\widehat{\sigma}_n^2(\mathbf{X}) &= \frac{\alpha(R)+n}{\alpha(R)+n+1}\operatorname{Var}(\widehat{F}_\alpha) \\ &= \frac{\alpha(R)+n}{\alpha(R)+n+1}\left(p_n\sigma_0^2 + (1-p_n)s_n^2 + p_n(1-p_n)(\mu_0 - \overline{X}_n)^2\right) \\ &= \frac{\alpha(R)+n}{\alpha(R)+n+1} \\ &\quad \times \left(p_n\sigma_0^2 + (1-p_n)\left(p_n\frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2 + (1-p_n)s_n^2\right)\right),\end{aligned} \tag{2.32}$$

where $s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. The last equality expresses $\widehat{\sigma}_n^2$ as a mixture of three different estimates of the variance, as noted by Ferguson.

If the prior sample size $\alpha(R) \to 0$, keeping F_0 fixed, $\widehat{\sigma}_n^2$ converges to the estimate $\frac{1}{n+1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. This estimate is the best invariant or minimax estimator of the variance of a normal distribution under the loss $(\operatorname{Var} P - \widehat{\sigma}^2)^2/(\operatorname{Var} P)^2$.

2.4.3 Estimation of the Median

Next consider the problem of estimation of the median θ defined as $\theta = \operatorname{med} P$. Ferguson (1973) derived the Bayes estimator under the absolute error loss, $L(\theta, \widehat{\theta}) =$

$|\theta - \widehat{\theta}|$. θ is unique with probability one and thus a well defined random variable. Under this loss function, any median of the distribution of θ is a Bayes estimator of θ. For the Dirichlet process prior with parameter α, Ferguson points out that any median of the distribution of θ is a median of the expectation of P, and conversely,

$$\text{med(dist. med}\,P) = \text{med}\,\mathcal{E}P. \tag{2.33}$$

Thus any number t satisfying

$$\frac{\alpha((-\infty, t))}{\alpha(R)} \leq \frac{1}{2} \leq \frac{\alpha((-\infty, t])}{\alpha(R)} \tag{2.34}$$

is a Bayes estimate of θ with respect to the prior $\mathcal{D}(\alpha)$ and absolute error loss. With $F_0(t) = \alpha((-\infty, t])/\alpha(R)$, the Bayes estimate for the no-sample problem is $\widehat{\theta} = \textit{median of } F_0$ and for the sample of size n, it is

$$\widehat{\theta}_{\alpha n} = \textit{median of } \widehat{F}_\alpha, \tag{2.35}$$

where $\widehat{F}_\alpha$ is the Bayes estimate of F derived in Sect. 2.2.1.

Doss (1985a, 1985b) also considers the problem of estimating the median but in a different nonparametric Bayesian framework. Let $X_1, \ldots, X_n$ be a random sample with distribution F_θ, where $F_\theta(x) = F(x - \theta)$ for some F that has median 0. F is assumed to be unknown and the problem is to estimate θ. Rather than placing a prior on F, he chooses F_- and F_+ from $\mathcal{D}(\alpha_-)$ and $\mathcal{D}(\alpha_+)$, respectively, and defines $F(t) = (F_-(t) + F_+(t))/2$, where α_- and α_+ are the restriction of α to $(-\infty, 0)$ and $(0, \infty)$, respectively. Then, F is a random distribution function such that $F(0) = \frac{1}{2}$ (but not symmetric, although $E(F(t)) = F_0$). In fact F so defined is a mixture of two Dirichlet processes. Let $\mathcal{D}^*(\alpha)$ denote its distribution.

Let $\alpha = MF_0$, where F_0 is a distribution function with median zero and for simplicity, no mass at zero. He places a prior on the pair (F, θ) by assuming F and θ independent, $F \in \mathcal{D}^*(\alpha)$ and θ having an arbitrary distribution υ. Given θ and F, let $\mathbf{X}$ be a sample from $F(x - \theta)$. Assume that F_0 has continuous density f_0. Then, Doss obtains the marginal posterior distribution of θ given $\mathbf{X}$ as

$$d\upsilon(\theta|\mathbf{X}) = \kappa(\mathbf{X}) \prod^{*}\big[f_0(X_i - \theta)\big]\Psi(\mathbf{X}, \theta)d\upsilon(\theta), \tag{2.36}$$

where $\Psi^{-1}(\mathbf{X}, \theta) = \Gamma(M/2 + n\widehat{F}_n(\theta))\Gamma(M/2 + n(1 - \widehat{F}_n(\theta)))$, $\widehat{F}_n$ the sample distribution function, $\prod^*$ represents the product taken over the distinct X_i and $\kappa(\mathbf{X})$ is a normalizing constant.

Using the posterior distribution one can find the Bayes estimate of θ. Doss states that the estimator is essentially a convex combination of the maximum likelihood estimator with respect to F_0 and the sample median, with mixing weights depending on the sample values. He also shows that the Bayes estimator is consistent only if the true distribution of X_j is discrete. He also derives the posterior distribution of θ in the case of F being a 'neutral to the right type' distribution discussed in Sect. 1.5.

2.4.4 Estimation of the q-th Quantile

Ferguson (1973) extends the estimation of the median to the q^{th} quantile of P, denoted by t_q: $P((-\infty, t_q)) \leq q \leq P((-\infty, t_q])$, for $0 < q < 1$. The q^{th} quantile of $P \in \mathcal{D}(\alpha)$ is unique with probability 1, so that t_q is a well defined random variable. He considers the following loss function,

$$\begin{aligned} L(t_q, \widehat{t_q}) &= p(t_q - \widehat{t_q}) \quad \text{if } t_q \geq \widehat{t_q} \\ &= (1-p)(t_q - \widehat{t_q}) \quad \text{if } t_q < \widehat{t_q}, \end{aligned} \tag{2.37}$$

for some p, $0 < p < 1$. For this loss, any p^{th} quantile of the distribution of t_q is a Bayes estimator of t_q. The distribution of t_q is

$$\begin{aligned} \mathcal{P}\{t_q \leq t\} &= \mathcal{P}\big\{F(t) > q\big\} \\ &= \int_q^1 \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)} z^{uM-1}(1-z)^{(1-u)M-1} dz, \end{aligned} \tag{2.38}$$

where $M = \alpha(R)$ and $u = \alpha((-\infty, t])/\alpha(R) = F_0(t)$. Setting this expression equal to p and solving the resulting equation for t, Ferguson obtains the p^{th} quantile of t_q. For fixed p, q, and M, let this equation define a function $u(p, q, M)$. The Bayes estimate of t_q for the no-sample problem is the u^{th} quantile of F_0,

$$\widehat{t_q} = u\big(p, q, \alpha(R)\big)\text{-th quantile of } F_0, \tag{2.39}$$

and for the sample of size n, it is

$$\widehat{t_q}(\mathbf{X}) = u\big(p, q, \big(\alpha(R) + n\big)\big)\text{-th quantile of } \widehat{F_\alpha}. \tag{2.40}$$

If p and q are both $\frac{1}{2}$, this reduces to the estimate of the median, since $u(\frac{1}{2}, \frac{1}{2}, M) = \frac{1}{2}$ for all M.

Doss (1985a, 1985b) extends his results of estimating the median to the estimation of quantiles as well, and discusses their properties.

2.4.5 Estimation of a Location Parameter

Dalal (1979b) considered the following model for sample observations. Let $X = \eta + \varepsilon$, where η is the location parameter and ε, the error term. Assume that η and ε are independent. The objective is to estimate η based on a random sample $Y_1, \ldots, Y_n$ from a η-symmetric distribution function F_η. That is F_η is assumed to be symmetric about η, but otherwise η and F_η are unknown. If $\varepsilon \sim G$ and $G \in \mathcal{D}(MF_0)$, where F_0 could be a standard normal distribution, then $\mathcal{E}(G) = F_0$ and hence the errors are generated by a distribution in the neighborhood of F_0. With M large the neighborhood becomes concentrated around F_0. Thus Dalal argues that the model can be interpreted from a robustness perspective as well. Let η be distributed according to a prior distribution υ, the group of transformations $\mathcal{G} = \{e, g\}$ with $e(x) = x$, $g(x) = 2\eta - x$, and α be a η-symmetric non-null finite measure on $(R, \mathcal{B})$. Given η, Dalal (1979b) assumes F_η to be distributed according to the Dirichlet Invariant process, $\mathcal{DGI}(\alpha)$, and obtains a Bayes estimate $\widehat{\eta}(\mathbf{y}) = \mathcal{E}_{\eta|\mathbf{y}}(\eta)$ of η, where the expectation is taken with respect to the conditional distribution $\upsilon(\cdot|\mathbf{y})$ of η given $\mathbf{y}$ averaged over F_η. However, $\widehat{\eta}(\mathbf{y})$ is not in a closed form and he encounters computational difficulties which is illustrated by an example consisting of 2 observations.

Let $\alpha = MF_0$, and assume that F_0 has a density f_0, and that we have a sample of size one, $Y_1 \sim F_\eta$ with $F_\eta \in \mathcal{DGI}(\alpha)$. Then $\mathcal{E}(F_\eta) = F_0$ and the marginal conditional distribution of Y_1 given η is F_0. Since υ is prior distribution of η, the conditional density of $\eta|Y_1 = y_1 \sim f_0(y_1)/\int f_0(x)d\upsilon(x)$.

If we have a second observation y_2, then we run into difficulty since the distribution of $Y_2|y_1, \eta, F_\eta \sim F_\eta$. But $F_\eta|y_1, \eta \in \mathcal{DGI}(\alpha + \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{2\eta-y_1})$ which results in a distribution of $Y_2|y_1, \eta$ as a combination of continuous and discrete parts with point discrete masses at y_1 and $2\eta - y_1$. Thus the evaluation of the posterior distribution of $\eta|y_2, y_1$ gets complicated (see Dalal 1979b). The above argument is extended to the case of n observations and shown that if υ is absolutely continuous, the posterior distribution of $\eta|\mathbf{y}$ is a mixture of absolutely continuous and discrete probabilities. The mixing weights depend upon not only the distinct observations but also on their multiplicities. The discrete component concentrates its mass on the points $(y_i + y_j)/2$, $i \neq j$. However, the computational techniques lately developed should make it easy to compute the posterior distribution.

This and other aspects of Bayesian estimation of a location parameter are discussed in his paper in detail.

Doss (1985a) also discusses this model, but instead of errors drawn from a symmetric distribution, he takes them to be drawn from an F which has median 0, but otherwise unknown, and it is desired to estimate η. He places priors on the pair (F, η) and computes the marginal posterior distribution of η and takes the mean of the distribution as the estimate of η. In a follow up paper (Doss 1985b) he discusses consistency issues and shows that the Bayes estimates are consistent if the distribution of errors is discrete, otherwise they can be inconsistent.

2.4.6 Estimation of $P(Z > X + Y)$

Zalkikar et al. (1986) considered the problem of estimation of the parameter

$$\Delta(F) = P(Z > X + Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(x + y)dF(x)dF(y), \tag{2.41}$$

where it is assumed that the random variables X, Y and Z are independent from the distribution function F, and $S = 1 - F$. This problem is encountered in reliability theory where it is desired to test whether new is better than used. Assume $F \in \mathcal{D}(\alpha)$ and the squared error loss L_2. Based on a random sample $\mathbf{X} = (X_1, \ldots, X_m)$ from F, they derived the Bayes estimator as follows.

$$\widehat{\Delta}(F) = \frac{M+m}{(M+m)^{(3)}} \Bigg[2(M+m+1) + \widehat{F}_\alpha(0-) \\ + (M+m) \int_{-\infty}^{\infty} S_\alpha(2y) d\widehat{F}_\alpha(y) + (M+m)^2 \Delta(\widehat{F}_\alpha) \Bigg], \tag{2.42}$$

where $a^{(k)} = a(a+1)\ldots(a+k-1)$, $S_\alpha = 1 - \widehat{F}_\alpha$.

When $M \to 0$, the estimator reduces to an estimator which is asymptotically equivalent to a U-statistic,

$$U_m = \frac{1}{m(m-1)(m-2)} \sum I[X_i > X_j + X_k], \tag{2.43}$$

where the summation is taken over all $m(m-1)(m-2)$ distinct triplets (i, j, k), $1 \leq i, j, k \leq m$.

By using the earlier mentioned technique, they obtain an empirical Bayes estimate at the $(n+1)$-th stage utilizing the past as well as current data, which is also shown to be asymptotically optimal with the rate of convergence $O(n^{-1})$.

2.5 Other Applications

There are many other applications that have not been covered here. For example, Lo (1987) has studied the Bayesian bootstrap estimation of a functional $\theta(P; X_1, \ldots, X_n)$, where the variables $X_1, \ldots, X_n$ are iid P, with P having the prior $\mathcal{D}(\alpha)$. He has provided large sample Bayesian bootstrap probability intervals for the mean, variance and confidence bands for the distribution function, the smoothed density and smoothed rate function. In a second paper (Lo 1988), he also considered a Bayesian bootstrap for a finite population.

Dirichlet process priors have also been used for bandits problem by Clayton and Berry (1985).

We do however present a few interesting applications.

2.5.1 Bayes Empirical Bayes Estimation

In a general empirical Bayes setting, we have n unobservable independent random variables θ_i, $i = 1, 2, \ldots, n$ from an unknown distribution G, and associated with each θ_i, we have a random variable X_i chosen independently from a distribution with density function $f_i(x|\theta_i)$, $i = 1, 2, \ldots, n$. The problem is to estimate θ_i's or G itself. A common procedure is to obtain first an estimator G_n of G from the data $X_1, \ldots, X_n$, and then estimate θ_i as the Bayes estimate with respect to the prior G_n. In the Bayesian approach to the empirical Bayes problem, G itself is to be considered random with a prior distribution. Berry and Christensen (1979) followed this route assuming the Dirichlet prior $\mathcal{D}(\alpha)$ for G. Antoniak (1974) had shown that the posterior distribution of G is a mixture of Dirichlet processes with parameter $\alpha + \sum_{i=1}^n \delta_{\theta_i}$ and mixing distribution $H(\boldsymbol{\theta}|\mathbf{X})$. Thus, the posterior distribution of G given $\mathbf{X}$ in symbols is

$$G|\mathbf{X} \in \int \mathcal{D}\left(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}\right) dH(\boldsymbol{\theta}|\mathbf{X}). \tag{2.44}$$

If we have an unconditional marginal distribution of $\boldsymbol{\theta}$, then $dH(\boldsymbol{\theta}|\mathbf{X})$ can be expressed as

$$dH(\boldsymbol{\theta}|\mathbf{X}) = \prod_{j=1}^{n} f_j(x_j|\theta_j) dH(\boldsymbol{\theta}) \Bigg/ \left[\int \prod_{j=1}^{n} f_j(x_j|\theta_j) dH(\boldsymbol{\theta})\right]. \tag{2.45}$$

However, even in the simple case where $f_i(x|\theta)$ is a binomial distribution with parameter θ, Berry and Christensen (1979) found it difficult to evaluate and recommended some approximations. By using a lemma of Lo (1984), Kuo (1986a, 1986b) was able to express the Bayes estimator of θ_i under the loss $\sum_{i=1}^n (\theta_i - \widehat{\theta})^2$ in a concise form as a ratio of two n-dimensional integrals as follows.

$$\widehat{\theta_i} = \mathcal{E}(\theta_i|\mathbf{X}) = \frac{\int \cdots \int_{R^n} (\theta_i \prod_{i=1}^n f_i(x_i|\theta_i)) \prod_{i=1}^n (\alpha + \sum_{j=1}^{i-1} \delta_{\theta_j})(d\theta_i)}{\int \cdots \int_{R^n} (\prod_{i=1}^n f_i(x_i|\theta_i)) \prod_{i=1}^n (\alpha + \sum_{j=1}^{i-1} \delta_{\theta_j})(d\theta_i)} \tag{2.46}$$

for all $i = 1, 2, \ldots, n$. Still it is hard to evaluate these integrals. She overcomes this problem by decomposing each of the multi-dimensional integrals as a weighted average of products of one dimensional integrals and approximating each of the weighted averages by an importance sampling Monte Carlo method. She illustrates the computation in detail with a numerical example.

This model has been discussed in Escobar and West (1995) and Escobar (1994). Lavine (1994) generalizes the approach by using a Polya tree prior for G and shows how the posterior distribution can be computed via the Gibbs sampler and demonstrates the advantages of using mixtures of Polya trees over mixtures of Dirichlet processes.

2.5.2 Bioassay Problem

The goal of the bioassay problem is to assess the dose-response relationship in a population. In particular, one is interested in the estimation of the distribution of tolerance level to a drug administered to subjects at various dose levels. In order to determine an effective dose, one needs to collect data at different dose levels and their effect on the subject in mitigating the condition for which the drug is administered. The impact of the drug on subjects is represented by a CDF, $F(t)$, defined on $[0, \infty)$ and represents the proportion of the population that would respond to dose t. This distribution is often known as dose-response curve in the field of bioassay.

Suppose a stimulus is administered to n_j subjects at dose level t_j with positive response in r_j subjects, $j = 1, 2, \ldots, L$. Let $F(t)$ represent the probability of getting positive response at dose level t. Thus r_j, $j = 1, 2, \ldots, L$ are independent, each being a binomial random variable with parameters n_j and $F(t_j)$. Based on such quantal response data, the task is to estimate the response curve F nonparametrically from a Bayesian approach. This problem was first considered by Kraft and van Eeden (1964) who use a dyadic tailfree process as prior. The computations are difficult and were illustrated in the case of only three dose levels in their paper. Ramsey (1972) uses a Dirichlet process prior and obtains the modal estimates of F by maximizing the finite dimensional joint density of the posterior distribution which is not a Dirichlet.

Ferguson and Phadia (1979) noted that the bioassay problem may be considered as a censored sampling problem in which bioassay positive responses are observations censored on the left (since they could have responded to the drug at t_i^- but were observed at t_i only), and non-responses (failures) are observations censored on the right. Thus if all positive responses were considered as the real observations, they can be taken care of by updating the parameter of the Dirichlet prior. They showed (see Sect. 3.2.8) that the application of Ramsey's formulas when all observations are failures and Dirichlet process updated for real observations, yield the modal estimate of F (expression (3.22)). They also noted that in the case of all failures, Ramsey's modal estimate has a simple closed form (Ramsey's estimator was not) and is essentially given by the Susarla-Van Ryzin estimator of the survival function $S = 1 - F$.

Antoniak (1974) also assumes the Dirichlet process prior and worked out an exact solution in the case of two dose levels and showed that the posterior distribution leads to a mixture of Dirichlet processes. For example if there is only one dose at t_1, $F(t_1)$ has a beta distribution, $Be(\alpha(0, t_1], \alpha(t_1, \infty))$ and the posterior distribution would be $Be(\alpha(0, t_1] + r_1, \alpha(t_1, \infty) + n_1 - r_1)$, and therefore, the Bayes estimator under the integrated squared error loss will be the mean of this distribution, $\widehat{F}(t_1) = (\alpha(0, t_1] + r_1)/(\alpha(0, \infty) + n_1)$. This is deceptively simple. For two dose levels at $t_1 < t_2$ it starts to get complicated. Antoniak worked out the details and produced the following estimator.

$$\widehat{F}(t_1) = \sum_{i=0}^{r_2} \sum_{j=0}^{n_1 - r_1} a_{ij} \frac{\beta_1 + r_1 + i}{M + n_1 + n_2} \tag{2.47}$$

$$\widehat{F}(t_2) = \sum_{i=0}^{r_2} \sum_{j=0}^{n_1 - r_1} b_{ij} \frac{\beta_1 + \beta_2 + n_1 + r_2 - j}{M + n_1 + n_2} \tag{2.48}$$

and for other values of t, $\widehat{F}(t)$ is obtained by the linear interpolation. Here

$$a_{ij} = b_{ij} \Big/ \sum_{i=0}^{r_2} \sum_{j=0}^{n_1-r_1} b_{ij} \quad \text{and} \tag{2.49}$$

$$\begin{aligned} b_{ij} = {} & \binom{n_1 - r_1}{j}\binom{r_2}{i} \\ & \times \Gamma(\beta_1 + r_1 + i)\Gamma(\beta_2 + n_1 - r_1 + r_2 - i - j)\Gamma(\beta_3 + n_2 - r_2 + j) \\ & \Big/ \big(\Gamma(\beta_1)\Gamma(\beta_2)\Gamma(\beta_3)\big), \end{aligned} \tag{2.50}$$

with $\beta_1 = \alpha(0, t_1]$, $\beta_2 = \alpha(0, t_2]$ and $\beta_3 = \alpha(t_2, \infty)$. For the general case, the expressions are complicated and involve multiple integrals.

Bhattacharya (1981) develops procedures to compute finite dimensional distributions of the posterior distribution of a Dirichlet prior. Taking a lead from Ferguson and Phadia (1979), Ammann (1984) writes $F(t) = 1 - \exp(-H(t))$ and assumes H to be a process with independent increments with no deterministic component. He then derives the posterior distribution of $H(t)$ in terms of Laplace transforms. However, the expressions are no simpler.

In view of these difficulties, Kuo (1988) proposed a linear Bayes estimate of F which is a Bayes rule in the space generated by $r_1, \ldots, r_L$ and 1. She derives the estimator by point-wise minimization of the loss function $\int (F - \widehat{F})^2 dW$ at each dose level. At any point t which is not a dose level, the estimate is defined by the linear interpolation of estimates at the two adjacent dose levels. Her result is as follows.

Let $\operatorname{cov}(\mathbf{r})$ denote the covariance matrix of $r_1, \ldots, r_L$, and let $D(i, t_j)$ denote the covariance matrix with the i-th column replaced by the column $(\operatorname{cov}(r_1, F(t_j)), \ldots, \operatorname{cov}(r_L, F(t_j)))^T$. Also, let $M = \alpha[0, \infty)$, $F_0(t) = \alpha(t)/M$ and C be a class of decision rules which are linear combinations of $r_1, \ldots, r_L$ and 1. Then with $F \in \mathcal{D}(\alpha)$, the Bayes rule in this class at each dose level t_j, $j = 1, 2, \ldots, L$ is given by

$$\widehat{F}(t_j) = F_0(t_j) + \sum_{i=1}^{L} n_i \widehat{\lambda}_i(j)\big[r_i/n_i - F_0(t_i)\big], \tag{2.51}$$

and at t, $t_j < t < t_{j+1}$

$$\widehat{F}(t) = \frac{F_0(t_{j+1}) - F_0(t)}{F_0(t_{j+1}) - F_0(t_j)}\widehat{F}(t_j) + \frac{F_0(t) - F_0(t_j)}{F_0(t_{j+1}) - F_0(t_j)}\widehat{F}(t_{j+1}) \tag{2.52}$$

where $\widehat{\lambda}_i(j) = |D(i, t_j)|/|\operatorname{cov}(r)|$.

Kuo also shows that $\widehat{F}(t_j)$ is an asymptotically unbiased and consistent estimator of $F(t_j)$. As $M \to 0$, $\widehat{F}(t_j) \to r_j/n_j$ and as $M \to \infty$, $\widehat{F}(t_j) \to F_0(t_j)$. She points out that at times the estimator may not be monotone and if monotonicity is essential, one can use the pool-adjacent-violators algorithm (pp. 13–18 in Barlow et al. 1972) for obtaining the desired result.

In the case of M and F_0 unknown, empirical Bayes method of Sect. 2.2.4 may be used.

2.5.3 A Regression Problem

In the bioassay problem, the objective was to estimate the dose-response curve. Antoniak (1974) points out that a similar problem that arise in regression problems can also be handled in the same way. Let G be a distribution function on [0, 1] and assume that $G \in \mathcal{D}(\alpha)$. At chosen points $0 = t_0 < t_1 < \ldots < t_k \leq 1$, assume that we have samples $\mathbf{X}_l = (X_{l1}, \ldots, X_{lm_l})$ from $F(x|G(t_l))$, $l = 1, 2, \ldots, k$, and based on these samples, our aim is to make inferences about the parameters $G(t_l)$. Since G has a Dirichlet process prior, the joint distribution of $(G(t_1), G(t_2) - G(t_1), \ldots, 1 - G(t_k))$ is a Dirichlet distribution with parameters $(\alpha(t_1), \alpha(t_2) - \alpha(t_1), \ldots, \alpha(1) - \alpha(t_k))$. He points out that the observations for different values of l will not be generally independent and thus the calculations become complex. He illustrates them by taking an example with $k = 2$. Note that in bioassay problems, at each value of l, the observations available were from a binomial distribution, where as in the regression problem, they arise from some known distribution.

Consider the general linear model $Z = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\mathbf{X}$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of regression coefficients, and ϵ is the error term. Traditionally the practice is to assume the error term to be distributed as a parametric distribution, typically normal distribution with mean zero. The nonparametric Bayesian approach is to assume the error term having an unknown distribution, and a prior is placed on the unknown distribution (see Antoniak 1974 for example) centered around a base distribution which may be taken as normal with mean zero. There are several papers along this line using different priors. Since the base of a Polya tree prior includes absolutely continuous distributions, it is found to be favorable over the Dirichlet process.

Lavine (1994) considers the model $Y_i = \varphi(X_i, \beta) + \epsilon_i$, where φ is a known function, X_i is a known vector of covariates, β is an unknown vector of regression parameters with prior density f and the ϵ_i are independent with unknown distribution P. Assuming $P|\beta \sim PT(\Pi_\beta, \mathcal{A}_\beta)$, he derives the posterior distribution of β and shows that the posterior distribution of $P|\beta$ is $PT(\Pi_\beta, \mathcal{A}_\beta | Y_1 - \varphi(X_1, \beta), \ldots, Y_n - \varphi(X_n, \beta))$.

Walker and Mallick (1997b) use a finite Polya tree prior for the error distribution in a hierarchical generalized linear model centered around a known base probability measure (by taking partitions to coincide with the percentiles of the corresponding distribution function) and find this approach to be more appropriate than a parametric approach. They extend this approach to an accelerated failure time model (Walker and Mallick 1999) where the error distribution is assumed to have a Polya tree prior and show how to implement MCMC procedure with application to survival data. Procedure to simulate a random probability measure P from $\mathcal{PT}(\Pi, A)$ is also indicated in their paper. This is done by first generating a finite set of beta random variables and defining the random measure P_M by $P(B_{\epsilon_1 \ldots \epsilon_M})$ for each $\epsilon_1 \ldots \epsilon_M$ according to (1.108). Then one of the 2^M sets is picked according to the random weights $P(B_{\epsilon_1 \ldots \epsilon_M})$ and then a uniform random variate is taken from this set. If one of the set chosen happens to be an extreme set, then the random variate is chosen according to the base measure G_0 restricted to this set. α's are chosen such that they increase rapidly down towards level M. See their paper for details.

Hanson and Johnson (2002) argue that in practice it may be difficult to specify a single centering/base distribution G_0. Therefore, they recommend modeling the error distribution in a linear model as a mixture of Polya trees. A mixture of Polya tree distribution G is specified by allowing parameters of the centering distribution G_0 and/or the family of real numbers α's to be random. That is, $G|U, C \sim PT(\Pi_u, A_c)$, $U \sim f_u(u)$, $C \sim f_C(c)$. They consider mixtures of Polya trees in which the partition is constructed by a parametric family of probability distributions with variance U. The effect of taking mixtures is to smooth out the partitions of a simple Polya tree. Hanson (2006) further justify the efficiency of using mixtures of Polya trees alternative to using parametric models and provide computational strategies to carry out the analysis and illustrate them by discussing several examples.

Kalbfleisch (1978), Wild and Kalbfleisch (1981) and Hjort (1990) cast the regression problem in terms of the Cox model to accommodate covariates in survival data analysis. Kalbfleisch (1978) used a gamma process as prior for the unknown distribution function, Wild and Kalbfleisch (1981) extended the work of Ferguson and Phadia (1979) in which the neutral to the right process was used, and Hjort (1990) uses a beta process as prior for the cumulative hazard function. Their work is summarized in Sect. 3.7.

2.5.4 Estimation of a Density Function

The nonparametric Bayesian density function estimation may be viewed as an application of the mixtures of Dirichlet processes.

Let $X_1, \ldots, X_n$ be a sample of size n from a density function $f(x)$ with respect to some finite measure on R. Based on $\mathbf{X} = (X_1, \ldots, X_n)$, consider the problem of estimating $f(x)$ at some fixed point x, or some functional of $f(x)$, such as the mean $\int xf(x)dx$. For the Bayesian treatment, we need to assign a prior on the space of all density functions and be able to handle the posterior distribution analytically. In order that the posterior distribution is manageable, it would be preferable to find a conjugate family of priors. This is known to be difficult. Lo (1984, 1986) approaches this problem by using a kernel representation of the density function, and assigning a Dirichlet prior to G. His results are presented here.

Let G be a distribution function on R and α a finite measure on $(R, \mathcal{B})$. Let $K(x, u)$ represent a kernel defined on $(\mathcal{X} \times R)$ into R^+ such that for each $u \in R$, $\int_{\mathcal{X}} K(x, u)dx = 1$ and for each $x \in \mathcal{X}$, $\int_R K(x, u)\alpha(du) < \infty$. (Lo takes $\mathcal{X}$ and R to be Borel subsets of Euclidean spaces.) The posterior distribution of $G|\mathbf{X}$ has been obtained by Antoniak (1974) as indicated earlier. For each $G \in \mathcal{F}$, define

$f(x|G) = \int_R K(x,u)G(du)$, then $f(\cdot|G)$ is a kernel representation of the density function f and G is known as a mixing distribution. Lo defines a prior distribution for random f by letting G to be a random distribution with Dirichlet process prior $\mathcal{D}(\alpha)$. This way the broad support for the prior on the space of G is extended to the broad support for the prior on the space of all density functions. Since $G \in \mathcal{D}(\alpha)$, it can be seen that for each $x \in \mathcal{X}$, the marginal density of X is $f_0(x) = \int_{\mathcal{F}} f(x|G)\mathcal{D}_\alpha(dG) = \int_R K(x,u)\alpha(du)/\alpha(R)$. Now the posterior distribution of G given the data $\mathbf{X}$ can be seen to be

$$\mathcal{P}(G \in B|\mathbf{X}) = \frac{\int_B \prod_{i=1}^n \int_R K(x_i,u_i)G(du_i)\mathcal{D}_\alpha(dG)}{\int_{\mathcal{F}} \prod_{i=1}^n \int_R K(x_i,u_i)G(du_i)\mathcal{D}_\alpha(dG)}, \tag{2.53}$$

for all $B \in \mathcal{F}$. By repeated application of his lemma (interchanging the order of integration),

$$\int_{\mathcal{F}}\int_R h(u,G)G(du)\mathcal{D}_\alpha(dG) = \int_R\int_{\mathcal{F}} h(u,G)\mathcal{D}_{\alpha+\delta_u}(dG)\alpha(du)/\alpha(R), \tag{2.54}$$

he shows that

$$\mathcal{P}(G \in B|\mathbf{X}) = \frac{\int_{R^n} \mathcal{D}_{\alpha+\sum\delta_{u_i}}(B)\mu_{n,k,\alpha}(d\mathbf{u})}{\int_{R^n} \mu_{n,k,\alpha}(d\mathbf{u})}, \tag{2.55}$$

where

$$\mu_{n,K,\alpha}(C) = \int_C \prod_{i=1}^n K(x_i,u_i) \prod_{i=1}^n \left(\alpha + \sum_{j=1}^{i-1} \delta_{u_j}\right)(du_i) \tag{2.56}$$

for $C \in \mathcal{B}^n$, $d\mathbf{u} = \prod_{i=1}^n du_i$ and $\mathbf{u} \in R^n$. For any measurable function g, this leads to

$$\mathcal{E}\big(g(G)|\mathbf{X}\big) = \frac{\int_{R^n} g(G)\mathcal{D}_{\alpha+\sum\delta_{u_i}}(dG)\mu_{n,k,\alpha}(d\mathbf{u})}{\int_{R^n} \mu_{n,k,\alpha}(d\mathbf{u})}. \tag{2.57}$$

Now, by taking $g(G) = f(x|G)$ and simplifying, the posterior expectation $\widehat{f}(x|G)$ of $f(x|G)$ is derived as

$$\widehat{f_\alpha}(x|G) = \mathcal{E}\big(f(x|G)|\mathbf{X}\big) = p_n f_0(x) + (1-p_n)\widehat{f_n}(x), \tag{2.58}$$

which is a convex combination of prior guess $f_0(x)$ defined above, and a quantity $\widehat{f_n}(x)$, to be defined below, which mirrors the sample distribution function, but is complicated.

Let $N(\underline{P})$ denote the number of cells in the partition $\underline{P}$ of $\{1,2,\ldots,m\}$; C_i the i-th cell of $\underline{P}$ with m_i elements in it, $i = 1,\ldots,N(\underline{P})$; $g_i(u)$, $i = 1,\ldots,m$ are m positive or α-integrable functions;

$$\varphi(\underline{P}) = \prod_{i=1}^{N(\underline{P})} \left\{(m_i-1)! \int_R \prod_{l \in C_i} g_l(u)\alpha(du)\right\} \tag{2.59}$$

and finally, $w(\underline{P}) = \varphi(\underline{P})/\sum_{\underline{P}} \varphi(\underline{P})$. Then $\widehat{f_n}(x)$ is given by

$$\widehat{f_n}(x) = \frac{1}{n} \sum_{\underline{P}} w(\underline{P}) \sum_{i=1}^{N(\underline{P})} m_i \left\{ \frac{\int_R K(x,u) \prod_{l \in C_i} K(x_l, u)\alpha(du)}{\int_R \prod_{l \in C_i} K(x_l, u)\alpha(du)} \right\}, \tag{2.60}$$

where the summation is taken over all partitions $\underline{P}$ of $\{1, 2, \ldots, m\}$. $\widehat{f}$ serves as a Bayes estimate under the loss function $L(f, \widehat{f}) = \int |f(x|G) - \widehat{f}(x|G)|^2 W(dx)$, where W is a weight function.

Lo discusses the choice of the kernel K and the parameter α of the prior, and gives several examples of $K(x, u)$ and α and computes the Bayes estimators. His examples of kernels include histogram, normal with location and/or scale parameters, symmetric and unimodal densities, decreasing densities, etc. For example, if K is chosen to reflect the histogram model, the estimator reduces to the usual Bayes estimates of cell probabilities. Kuo's (1986a, 1986b) Monte Carlo method may be adapted to carry out the calculations. Details may be found in his paper. Lavine (1992) uses mixtures of Polya trees in density estimation.

Ghorai and Susarla (1982) considered an empirical Bayes approach to the above problem. Assuming $\alpha(R)$ to be known, they obtained an estimator of $f_0(x) = \int_R K(x,u)\alpha(du)/\alpha(R)$ based on previous n copies and substituted in the Bayesian estimator $\widehat{f}(x|G)$ at the $(n+1)$-th stage. Under certain conditions, they prove the asymptotic optimality of the resulting estimator.

Ferguson (1983) considered a different formulation of the density function. He modeled it as a countable mixtures of normal densities: $f(x) = \sum_{i=1}^{\infty} p_i h(x|\mu_i, \sigma_i)$ where $h(x|\mu, \sigma)$ is the normal density with mean μ and variance σ^2. This formulation has countably infinite number of parameters, $(p_1, p_2, \ldots, \mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$. Since the interest is in estimating $f(x)$ at a point x, and not in estimating the parameters themselves, it can be written as $f(x) = \int h(x|\mu, \sigma) dG(\mu, \sigma)$, where G is the probability measure on the half plane $\{(\mu, \sigma) : \sigma > 0\}$ that gives weight p_i to the point (μ_i, σ_i), $i = 1, 2, \ldots$. While Lo assumes a Dirichlet process prior for the unknown G, Ferguson defines a prior via the Sethuraman representation of G. He defines the prior distribution for the parameter vector $(p_1, p_2, \ldots, \mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$ as follows: vectors $(p_1, p_2, \ldots)$ and $(\mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$ are independent; $p_1, p_2, \ldots.$ are the weights with parameter M in Sethuraman representation; and $\xi_i = (\mu_i, \sigma_i)$ are iid with common gamma-normal conjugate prior for the two-parameter normal distribution. This shows that G is a Dirichlet process with parameter $\alpha = MG_0$, where $G_0 = \mathcal{E}(G)$ is the conjugate prior for (μ, σ^2), and its infinite sum representation is $G = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}$ where as usual $(p_1, p_2, \ldots)$ and $(\xi_1, \xi_2, \ldots)$ are independent and $\xi_i \overset{iid}{\sim} G_0$. Now given a sample $x_1, \ldots, x_n$ of size n from a distribution with density $f(x) = \int h(x|\xi) dG(\xi)$, the posterior distribution of G given $x_1, \ldots, x_n$ has been obtained by Antoniak (1974) as mixture of Dirichlet processes

$$G|x_1, \ldots, x_n \sim \int \cdots \int \mathcal{D}(\alpha + nG_n) dH(\xi_1, \ldots, \xi_n | x_1, \ldots, x_n)$$

with $nG_n = \sum_{i=1}^{n} \delta_{\xi_i}$. $H(\xi_1, \ldots, \xi_n | x_1, \ldots, x_n)$ is the posterior distribution of $\xi_1, \ldots, \xi_n$ given $x_1, \ldots, x_n$. Since $\mathcal{E}(\mathcal{D}(\alpha + nG_n)) = (MG_0 + nG_n)/(M + n)$,

$$\mathcal{E}\big(G(\xi)|x_1, \ldots, x_n\big) \\ = p_n G_0(\xi) + (1 - p_n) \int \ldots \int G_n(\xi) dH(\xi_1, \ldots, \xi_n | x_1, \ldots, x_n) \qquad (2.61)$$

and

$$\widehat{f}(x) = \mathcal{E}\big(f(x)|x_1, \ldots, x_n\big) = p_n f_0(x) + (1 - p_n)\widehat{f_n}(x), \qquad (2.62)$$

where: $p_n = M/(M + n)$ as before, $f_0(x) = \mathcal{E}(f(x)) = \sum_{i=1}^{\infty} \mathcal{E}(p_i)\mathcal{E}h(x|(\mu_i, \sigma_i)) = \mathcal{E}h(x|\mu, \sigma)$ and $\widehat{f_n}(x)$ is given by

$$\widehat{f_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \int \ldots \int h(x|\xi_i) dH(\xi_1, \ldots, \xi_n | x_1, \ldots, x_n). \qquad (2.63)$$

Following Lo, $\widehat{f_n}(x)$ can be written as a ratio $h(x, x_1, \ldots, x_n)/h(x_1, \ldots, x_n)$, where

$$h(x_1, \ldots, x_n) = \frac{1}{M^{(n)}} \int \ldots \int \left(\prod_{n=1}^{n} h(x_i|\xi_i)\right) \prod_{n=1}^{n} d\left(MG_0 + \sum_{j=1}^{i-1} \delta_{\xi_j}\right)(\xi_i), \quad (2.64)$$

and computations are carried out by Kuo's (1986a, 1986b) Monte Carlo method.

Normal mixtures also turn up in Escobar (1994) and Escobar and West (1995). Escobar's set up is as follows. Let $Y_i|\mu_i \sim N(\mu_i, 1)$, $\mu_i|G \overset{iid}{\sim} G$, μ_i and G are unknown. In contrast to Ferguson's and Lo's objectives, his objective is to estimate μ_i's (with the variance being known to be 1) based on observed Y_i's using the nonparametric Bayesian approach. When G is known the Bayesian estimator is the posterior mean

$$\mathcal{E}(\mu_i|Y_i) = \frac{\int \mu_i \phi(Y_i - \mu_i) dG(\mu_i)}{\int \phi(Y_i - \mu_i) dG(\mu_i)}, \qquad (2.65)$$

where ϕ is the density of the standard normal distribution function. When G is unknown, empirical Bayes methods are typically used. Instead Escobar uses a Dirichlet process prior for G. Antoniak has shown that if Dirichlet process prior is used for G, then the posterior distribution of μ_i is a mixture of Dirichlet process. Thus it was computationally difficult. Kuo (1986b) and Lo (1984) developed Monte Carlo integration algorithms, but Escobar points out that they are inefficient since they do not sample values conditionally based on the data. He introduces a new Gibbs sampler like method that remedied this problem.

Escobar and West (1995) describe a normal mixture model, similar to Ferguson's (1983), in terms of the predictive distribution of a future observation. For their model, given (μ_i, σ_i^2), we have a random sample, say $Y_1, \ldots, Y_n$, such that $Y_i|(\mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ and the objective is to find the predictive distribution of next observation Y_{n+1} which is a mixture of normals,

$Y_{n+1}|Y_1,\ldots,Y_n \sim N(\mu_{n+1},\sigma_{n+1}^2)$. A usual practice is to put a parametric prior on vector $\boldsymbol{\nu} = (\mu_1,\ldots,\mu_n,\sigma_1^2,\ldots,\sigma_n^2)$. Ferguson models the common prior for $\nu_i = (\mu_i,\sigma_i^2)$ as a Dirichlet process prior. Thus the data is considered as coming from a Dirichlet mixture of normals in contrast to Antoniak where the Dirichlet process processes were mixed with respect to a parametric distribution $H(\theta)$, $\alpha_\theta \sim H(\theta)$. A particular case of $(\mu_i,\sigma_i^2) = (\mu_i,\sigma^2)$ has been studied (see West 1990, 1992) in which the distribution of μ_i's is modeled as Dirichlet process with a normal base measure.

In view of the discreteness of Dirichlet process prior which induces multiplicities of observations, $\nu_{n+1}|\nu_1,\ldots,\nu_n$ will have distribution of the form given in property 19 of Sect. 1.2. Then they proceed on the line of Ferguson, derive the conditional distribution of $Y_{n+1}|\nu_1,\ldots,\nu_n$ which is a mixture of a Student's t-distribution and n normals $N(\mu_i,\sigma_i^2)$, and then it is shown that the unconditional predictive distribution is given by $Y_{n+1}|Y_1,\ldots,Y_n \sim \int \mathcal{P}(Y_{n+1}|\boldsymbol{\nu})d\mathcal{P}(\boldsymbol{\nu}|Y_1,\ldots,Y_n)$. Since the evaluation of $\mathcal{P}(\boldsymbol{\nu}|Y_1,\ldots,Y_n)$ is difficult even in small samples, they use Monte Carlo approximation using extensions of the iterative technique developed by Escobar (1994).

2.5.5 Estimation of the Rank of X_1 Among $X_1,\ldots,X_n$

Let $X_1,\ldots,X_n \overset{iid}{\sim} F$. The problem of estimating the rank order G of X_1 among $X_1,\ldots,X_n$ based on the knowledge of r $(< n)$ observed values of $X_1,\ldots,X_r$ was considered from a Bayesian point of view by Campbell and Hollander (1978). WLOG assume that $x_1,\ldots,x_r$ are the first r values in the sample. Let K, L and M denote the number of observations among $X_1,\ldots,X_n$ that are less than, equal to and greater than X_1, respectively. Then the rank order G of X_1 is taken as the average value of the ranks that would be assigned to the L values tied at X_1, in the ascending order, i.e. $G = \frac{1}{L}\sum_{i=1}^{L}(K+i) = K + (L+1)/2$.

Let K', L' and M' be defined respectively, as the corresponding numbers of observations among $x_1,\ldots,x_r$. Then the rank order G' of x_1 among $x_1,\ldots,x_r$ is given by $G' = K' + (L'+1)/2$. Given $x_1,\ldots,x_r$, the problem is to estimate G which is clearly a function of K, L and M. Assuming $F \in D(\alpha)$, Campbell and Hollander obtained the posterior mean,

$$\widehat{G} = \mathcal{E}(G|x_1,\ldots,x_r) = G' + (n-r)\left\{\alpha'\big((-\infty,x_1)\big) + \frac{1}{2}\alpha'\big(\{x_1\}\big)/\alpha'(R)\right\}, \quad (2.66)$$

where $\alpha' = \alpha + \sum_{i=1}^{r}\delta_{x_i}$. $\widehat{G}$ depends on $x_1,\ldots,x_r$ only through G' and x_1. In comparison, the non-Bayesian estimators are given by $G_F = G' + (n-r)F(x_1)$ in the case of a known continuous function F and $G_U = G' + (n-1)G'/(n-r)$, when F is unknown.

2.6 Bivariate Distribution Function

Ferguson's (1973) definition of the Dirichlet process on an arbitrary space of probability measures makes it amenable for its extension to higher dimensions in a straight forward manner. In presenting the applications of Dirichlet process in bivariate situation, we will be concerned with the distribution and survival functions defined on $R^2 = R \times R$ and a finite non-null measure α on $(R^2, \mathcal{B}^2)$ where $\mathcal{B}^2$ represents the σ-field of Borel subsets of R^2.

Let P be a random probability measure on $(R^2, \mathcal{B}^2)$ and $F(x, y)$ be the corresponding bivariate distribution function. Assume that we have a random sample $(\mathbf{X}, \mathbf{Y}) = (X_1, Y_1), \ldots, (X_n, Y_n)$ from $F(x, y)$. Then the Bayesian estimators are presented first for the distribution function F and then for its functionals.

2.6.1 *Estimation of F w.r.t. the Dirichlet Process Prior*

For the Bayesian estimation of $F(x, y)$, we assume that F has a Dirichlet process prior with parameter α. As in the univariate case, we take the weighted loss function $L(F, \widehat{F}) = \int_{R^2} (F - \widehat{F})^2 dW$, where W now is a nonnegative weight function on R^2. The Bayesian estimator of $F(x, y)$ with respect to the Dirichlet process prior and the loss function L, is a direct extension of Ferguson's Bayesian estimator in one-dimension, and is given by

$$\begin{aligned}\widehat{F}_\alpha(x, y) &= \frac{\alpha((-\infty, x] \times (-\infty, y]) + \sum_{i=1}^n \delta_{(X_i, Y_i)}((-\infty, x] \times (-\infty, y])}{\alpha(R^2) + n} \\ &= p_n \frac{\alpha((-\infty, x] \times (-\infty, y])}{\alpha(R^2)} \\ &\quad + (1 - p_n) \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}\big((-\infty, x] \times (-\infty, y]\big). \qquad (2.67)\end{aligned}$$

Empirical Bayes estimation of $F(x, y)$ when $\alpha(\cdot)$ is unknown but $\alpha(R^2)$ is known can be carried out as in the univariate case. Also, following Zehnwirth's (1981) lead, an estimator for unknown $\alpha(R^2)$ was developed in Dalal and Phadia (1983) and was used when $\alpha(R^2)$ is assumed to be unknown.

2.6.2 *Estimation of F w.r.t. a Tailfree Process Prior*

In Chap. 1, the tailfree processes were introduced and their properties as well as the bivariate extension (Phadia 2007) were discussed. Here the Bayes estimator of F with respect to the bivariate tailfree process prior is derived under the weighted loss function. If x and y are binary rationals, then the estimate can be written as a finite sum; if either x or y is not a binary rational, then the estimate involves an infinite sum.

In view of the conjugacy property of tailfree processes, it is sufficient to derive the estimate for the no-sample problem. Then for a sample of size n, all we have to do is to update the parameters. Consider for example, $(x, y) = (\frac{1}{2}, \frac{3}{4})$. Following the notation of Sect. 1.16,

$$\begin{aligned}\widehat{F}\left(\frac{1}{2}, \frac{3}{4}\right) &= \mathcal{E}\left[F\left(\frac{1}{2}, \frac{3}{4}\right)\right]\\ &= \mathcal{E}\big[P(B_1) + P(B_{21}) + P(B_{23})\big]\\ &= \mathcal{E}[Z_1 + Z_2 Z_{21} + Z_2 Z_{23}]\\ &= \mathcal{E}[Z_1 + Z_{21} + Z_{23}]\\ &= \frac{\alpha_1}{\gamma_1} + \frac{\alpha_2}{\gamma_2}\frac{\alpha_{21}}{\gamma_{21}} + \frac{\alpha_2}{\gamma_2}\frac{\alpha_{23}}{\gamma_{23}}.\end{aligned}$$

On the other hand if $(x, y) = (\frac{1}{3}, \frac{1}{2})$, say, then

$$\begin{aligned}\widehat{F}\left(\frac{1}{3}, \frac{1}{2}\right) &= \mathcal{E}\left[F\left(\frac{1}{3}, \frac{1}{2}\right)\right]\\ &= \mathcal{E}\left[P(B_{11} \cup B_{12}) + P\left(\bigcup_{i=1}^{2}\bigcup_{j=1}^{2}(B_{13ij} \cup B_{14ij}) + \dots\right)\right]\\ &= \mathcal{E}\left[Z_{11} + Z_{12} + \sum_{i=1}^{2}\sum_{j=1}^{2}(Z_{13ij} + Z_{14ij}) + \dots\right]\\ &= \frac{\alpha_{11}}{\gamma_{11}} + \frac{\alpha_{12}}{\gamma_{12}} + \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{\alpha_{13ij}}{\gamma_{13ij}} + \frac{\alpha_{14ij}}{\gamma_{14ij}}\right) + \dots\end{aligned}$$

Now given a sample $\mathbf{X}$, all we have to do is to update the α's.

2.6.3 Estimation of a Covariance

The covariance of P is defined for $(x, y) \in R^2$ by the formula

$$\operatorname{Cov} P = \int xy\,dP - \int x\,dP \int y\,dP. \tag{2.68}$$

Assuming the squared error loss L_2 and $P \in \mathcal{D}(\alpha)$, Ferguson (1973) derived its Bayesian estimator. For the no-sample problem we have,

$$\mathcal{E}(\operatorname{Cov} P) = \frac{\alpha(R^2)}{\alpha(R^2) + 1}\sigma_{12}, \tag{2.69}$$

where σ_{12} is the covariance of $\mathcal{E}(P)$ given by $\sigma_{12} = [\int xyd\alpha(x,y) - \mu_1\mu_2]/\alpha(R^2)$, $\mu_1 = \int xd\alpha(x,y)/\alpha(R^2)$, and $\mu_2 = \int yd\alpha(x,y)/\alpha(R^2)$. Now for the sample of size n, we update α and obtain the Bayes estimate as

$$\widehat{\text{Cov}\, P}_\alpha = \frac{\alpha(R^2)+n}{\alpha(R^2)+n+1} \times \big(p_n\sigma_{12} + (1-p_n)s_{12} + p_n(1-p_n)(\mu_1 - \overline{X}_n)(\mu_2 - \overline{Y}_n)\big), \quad (2.70)$$

where $s_{12} = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$ is the sample covariance. This is again a mixture of three relevant quantities.

2.6.4 Estimation of the Concordance Coefficient

The problem of estimation of concordance coefficient in a bivariate distribution was treated in Dalal and Phadia (1983). Let (X,Y) and (X',Y') be two independent observations from a joint distribution function $F(x,y)$. A quantity of interest is $\Delta = P\{(X-X')(Y-Y') > 0\}$, which is related to Kendall's $\tau = \mathcal{E}(sign)(X-X')(Y-Y')$, by the equation $\tau = 2\Delta - 1$. It is used as a measure of the dependence between X and Y as well as a measure of the degree of concordance among observations from $F(x,y)$. Let

$$T_1 = \big\{(x,y,x',y') : (x-x')(y-y') > 0\big\} \quad \text{and}$$
$$T_2 = \big\{(x,y,x',y') : (x-x')(y-y') = 0\big\}. \quad (2.71)$$

Since F is allowed to be discrete, a slight modification of Δ, namely,

$$\Delta = P_F\big\{(X-X')(Y-Y') > 0\big\} + \frac{1}{2}\cdot P_F\big\{(X-X')(Y-Y') = 0\big\} \quad (2.72)$$

is preferred. The rationale is that the tied pairs are evenly distributed among concordants $(X-X')(Y-Y') > 0$ and discordant $(X-X')(Y-Y') < 0$. When X and Y are independent, $\Delta = 0$, and its estimator serves as a statistic to test the hypothesis of independence of X and Y. Now,

$$\Delta = \Delta_F = \int \Big(I_{T_1} + \frac{1}{2}\cdot I_{T_2}\Big) d\big(F(x,y)F(x',y')\big). \quad (2.73)$$

Assuming $F \in \mathcal{D}(\alpha)$, and α defined on $(R^2, \mathcal{B}^2)$, the Bayes estimator of Δ for the no sample problem is given by

$$\widehat{\Delta}_{\alpha 0} = \mathcal{E}_{\mathcal{D}(\alpha)}(\Delta_F) = \int \Big(I_{T_1} + \frac{1}{2}\cdot I_{T_2}\Big) d\mathcal{E}_{\mathcal{D}(\alpha)}\big(F(x,y)F(x',y')\big). \quad (2.74)$$

Let $\alpha = MQ$ and let G be a CDF corresponding to the measure Q. Applying Theorem 4 of Ferguson (1973) in evaluating $\mathcal{E}_{\mathcal{D}(\alpha)}(F(x, y)F(x', y'))$ and simplifying we get,

$$\widehat{\Delta}_{\alpha 0} = \frac{M}{M+1}\Delta_G + \frac{1}{2(M+1)}, \tag{2.75}$$

where $\Delta_G = P_G[(X - X')(Y - Y') > 0] + \frac{1}{2}P_G[(X - X')(Y - Y') = 0]$.

When X and Y are independent, $\Delta_G = \frac{1}{2}$, and therefore, $\widehat{\Delta}_{\alpha 0} = \frac{1}{2}$ also.

Now for the case of n observations, $(X_1, Y_1), \ldots, (X_n, Y_n) \sim F(x, y)$, the posterior distribution of F given the data is again a Dirichlet process with the parameter α updated as $\alpha + \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$, which can be rewritten as

$$\begin{aligned} \alpha + \sum_{i=1}^{n} \delta_{(X_i,Y_i)} &= (M+n)\left[\frac{M}{M+n}Q + \frac{1}{M+n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}\right] \\ &= (M+n)Q^*, \quad \text{say.} \end{aligned} \tag{2.76}$$

If G^* is a CDF corresponding to Q^*, then $G^* = p_n G + (1 - p_n)\widehat{G}_n$, where $\widehat{G}_n$ is the empirical CDF based on the n observations (x_i, y_i), $i = 1, 2, \ldots, n$, and $p_n = M/(M + n)$. Hence the Bayes estimator is given by,

$$\begin{aligned} \widehat{\Delta}_{\alpha n} &= \frac{M+n}{M+n+1}\int\left(I_{T_1} + \frac{1}{2}\cdot I_{T_2}\right)d\big(p_n G + (1-p_n)\widehat{G}_n\big)\big(p_n G + (1-p_n)\widehat{G}_n\big) \\ &\quad + \frac{1}{2(M+n+1)} \\ &= \frac{M+n}{M+n+1}\big[p_n^2\Delta_G + 2p_n(1-p_n)\Delta(G, \widehat{G}_n) + (1-p_n)^2\Delta_{\widehat{G}_n}\big] \\ &\quad + \frac{1}{2(M+n+1)}, \end{aligned} \tag{2.77}$$

where

$$\Delta_{\widehat{G}_n} = \frac{1}{n^2}\sum_{i,j=1}^{n}\left(I_{[(x-x')(y-y')>0]} + \frac{1}{2}I_{[(x-x')(y-y')=0]}\right), \tag{2.78}$$

and $\Delta(G, \widehat{G}_n) = \frac{1}{n}\sum_{i=1}^{n}\Delta_G(x_i, y_i)$ with

$$\Delta_G(x_i, y_i) = \left\{P_G\big[(X - x_i)(Y - y_i) > 0\big] + \frac{1}{2}P_G\big[(X - x_i)(Y - y_i) = 0\big]\right\}. \tag{2.79}$$

Here Δ_G and $\Delta_{\widehat{G}_n}$ can be interpreted as the natural estimates of the coefficient of concordance for the idealized model, and the sample and a single observation, respectively; whereas $\Delta_G(x_i, y_i)$ is the theoretical concordance probability of the pair (x_i, y_i).

The authors evaluated explicitly the Bayesian estimator for two interesting models, namely the bivariate normal and Gumbel's bivariate exponential distribution.

They extended the above result to the empirical Bayes estimate of Δ with M known, and used Zehnwirth's (1981) technique to estimate M, when M is unknown. In both cases, they showed that the estimates are asymptotically optimal with rate of convergence $O(n^{-1})$. The details can be found in their paper.

2.7 Estimation of a Function of *P*

The examples of Sects. 2.4 and 2.6 can be generalized to any measurable function $\phi(P)$ of P. Let $\mathfrak{X}^k$ denote the product space. A real valued function $\varphi : \Pi \to R$ is said to be *estimable* with kernel h if there exists a statistics $h(X_1, \dots, X_k)$ such that $\phi(P) = \int_{\mathfrak{X}^k} h(x_1, \dots, x_k) \prod_{i=1}^k dP(x_i)$. The degree of an estimable parameter $\phi(P)$ is the least sample size for which there is such an h (p. 151 in Zacks 1971). The Bayes and empirical Bayes estimation of estimable parameters of degree 1 and 2 under loss function L_2 and with respect to the Dirichlet and Dirichlet Invariant processes as priors were investigated by Yamato (1977a, 1977b), Tiwari (1981) and Tiwari and Zalkikar (1985). Their results are as follows.

Dirichlet Process Prior Based on a random sample $\mathbf{X}$ from P, the Bayesian estimator $\widehat{\phi}$ of ϕ under L_2 loss is given by the posterior mean $\mathcal{E}(\phi(P) \mid X_1, \dots, X_n)$. In particular, suppose $\phi(P) = \phi_h(P)$ and $P \in \mathcal{D}(\alpha)$, where

$$\phi_h(P) = \int_{\mathfrak{X}^k} h(x_1, \dots, x_k) dP(x_1) \dots dP(x_k), \tag{2.80}$$

and h is a symmetric measurable function from χ^k into R satisfying

$$\int_{\mathfrak{X}^k} \big|h(x_1, \dots, x_k)\big| d\overline{\alpha}(x_1) \dots d\overline{\alpha}(x_m) < \infty, \tag{2.81}$$

where as before, $\overline{\alpha}(\cdot) = \alpha(\cdot)/\alpha(R)$. Under a further assumption concerning the second moment of h with respect to $\overline{\alpha}^m$, $m \leq k$, namely

$$\int_{\mathfrak{X}^m} \big|h(x_1, \dots, x_1, x_2, \dots, x_2, \dots, x_m, \dots, x_m)\big|^2 d\overline{\alpha}(x_1) \dots d\overline{\alpha}(x_m) < \infty, \tag{2.82}$$

for all possible combinations of arguments $(x_1, \dots, x_1, x_2, \dots, x_2, \dots, x_m, \dots, x_m)$, $m \leq k$, from all distinct ($m = k$) to all identical ($m = 1$), the Bayes estimator of $\phi_h(P)$ with respect to $\mathcal{D}(\alpha)$ for the no sample problem is $\widehat{\phi}^0_{h,\alpha} = \mathcal{E}_{\mathcal{D}(\alpha)}(\phi_h(P))$, and for the sample $X_1, \dots, X_n$ it is

$$\widehat{\phi}^n_{h,\alpha} = \mathcal{E}_{\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{x_i})}\big(\phi_h(P)\big) = \widehat{\phi}^0_{h,\alpha + \sum_{i=1}^n \delta_{X_i}}. \tag{2.83}$$

Thus using this expression and Property 9 of Sect. 1.2, Yamato (1977a, 1977b) and Tiwari (1981) derived the following result. Based on a sample $X_1, \dots, X_n$, the

Bayes estimator of $\phi_h(P)$ with respect to the prior $\mathcal{D}(\alpha)$ and loss L_2 is given by

$$\begin{aligned}\widehat{\phi}^n_{h,\alpha} = &\sum_{C(\sum im_i=k)} \frac{k![\alpha(\mathfrak{X})+n]^{\sum m_i}}{\prod_{i=1}^k [i^{m_i}(m_i)!][\alpha(\mathfrak{X})+n]^{(k)}} \\ &\times \int_{\mathfrak{X}^{\sum m_i}} h(x_{11},\dots,x_{1m_1},x_{21},\dots,x_{2m_2},\dots,x_{k1},\dots,x_{km_k}) \\ &\times \prod_{i=1}^k \prod_{j=1}^{m_i} d\widehat{F}_\alpha(x_{ij}), \end{aligned} \tag{2.84}$$

where $\widehat{F}_\alpha(\cdot) = p_n\overline{\alpha}(\cdot) + (1-p_n)\widehat{F}_n(\cdot)$ is the Bayes estimator of F corresponding to P. Sethuraman and Tiwari (1982) showed that $\widehat{\phi}^n_{h,\alpha} \to \widehat{\phi}_{h,\sum_{i=1}^n \delta_{x_i}}$ as $\alpha(\mathfrak{X}) \to 0$.

Also, if $h(x_1,\dots,x_k)$ is such that it vanishes whenever two coordinates are equal, then

$$\widehat{\phi}_{h,\sum_{i=1}^n \delta_{x_i}} = \frac{n(n-1)\dots(n-k+1)}{n^{(k)}} U_{h,n}, \tag{2.85}$$

where $U_{h,n}$ is the usual U-statistic based on the sample $X_1,\dots,X_n$. Yamato (1977b) has proved that the asymptotic distribution of $\widehat{\phi}^0_{h,\sum_{i=1}^n \delta_{x_i}}$ is the same as that of $U_{h,n}$.

Using the above result and based on a sample $\mathbf{X}$, the Bayes estimators with respect to $\mathcal{D}(\alpha)$ of estimable functions of degree 1 and 2, namely $\phi_1(P) = \int h(x)dP(x)$ and $\phi_2(P) = \int h(x,y)dP(x)dP(y)$ are obtained in Tiwari and Zalkikar (1985) as

$$\widehat{\phi}_1(P) = p_n \int h(x)d\overline{\alpha}(x) + \frac{(1-p_n)}{n}\sum_{i=1}^n h(X_i) \tag{2.86}$$

and

$$\begin{aligned}\widehat{\phi}_2(P) = \frac{M+n}{M+n+1}\Bigg\{ & p_n^2 \int h(x,y)d\overline{\alpha}(x)d\overline{\alpha}(y) \\ & + \frac{2p_n(1-p_n)}{n}\sum_{i=1}^n \int h(x,x_i)d\overline{\alpha}(x) \\ & + \frac{(1-p_n)^2}{n^2}\sum_{i\neq j} h(X_i,X_j)\Bigg\}. \end{aligned} \tag{2.87}$$

From these two expressions, the Bayes estimators of parameters such as the mean, variance, covariance and the probability that X is stochastically smaller than Y can be derived. Explicit expressions were given earlier.

Tiwari and Zalkikar also extended Dalal and Phadia's (1983) result for the Bayes and empirical Bayes estimators of the concordance coefficient to a general parameter of degree 2, namely,

$$\varsigma = \int h\big(x, y; x', y'\big) dP(x, y) dP\big(x', y'\big), \tag{2.88}$$

where $h(x, y; x', y')$ is a real valued function defined on $(R^4, \mathcal{B}^4)$, where $\mathcal{B}^4$ stands for the corresponding Borel sets of R^4. The Bayes estimator of ς with respect to the Dirichlet process prior defined on $(R^2, \mathcal{B}^2)$ is given by

$$\begin{aligned} \widehat{\varsigma}_\alpha = {} & \frac{M+m}{M+m+1}\big[p_m^2 \varsigma_{\overline{\alpha}} + 2p_m(1-p_m)\varsigma(\overline{\alpha}, F_m) + (1-p_m)^2 \varsigma_{F_m}\big] \\ & + \frac{1}{M+m+1}\left\{ p_m \int h(x, y; x, y) d\overline{\alpha}(x, y) \right. \\ & \left. + \frac{(1-p_m)}{m} \sum_{i=1}^{m} h(X_i, Y_i; X_i, Y_i) \right\}, \end{aligned} \tag{2.89}$$

where $\varsigma_{\overline{\alpha}} = \int h(x, y; x', y') d\overline{\alpha}(x, y) d\overline{\alpha}(x', y')$, $\varsigma_{F_m} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} h(X_i, Y_i; X_j, Y_j)$, $\varsigma(\overline{\alpha}, F_m) = \frac{1}{m} \sum_{i=1}^{m} \varsigma_{\overline{\alpha}}(X_i, Y_i)$, and $\varsigma_{\overline{\alpha}}(x, y) = \int h(x, y; x', y') d\overline{\alpha}(x', y')$.

Note that by taking $h(x, y; x', y') = I_{T_1} + \frac{1}{2} I_{T_2}$ where

$$T_1 = \big\{\big(x, y; x', y'\big) : \big(x - x'\big)\big(y - y'\big) > 0\big\} \tag{2.90}$$

and

$$T_2 = \big\{\big(x, y; x', y'\big) : \big(x - x'\big)\big(y - y'\big) = 0\big\}, \tag{2.91}$$

the results of Dalal and Phadia (1983) can be obtained.

Dirichlet Invariant Process Prior Yamato (1986, 1987) carried out similar estimation procedures using the Dirichlet Invariant process with parameter α and under the same loss L_2. Let $\alpha = MQ$ and $M = \alpha(\mathfrak{X})$ and assume the same finite group $\mathcal{G} = \{g_1, \ldots, g_k\}$ of transformations as used in Dalal (1979a).

In particular, if we take $\mathcal{G} = \{e, g\}$ with $e(x) = x$, $g(x) = 2\eta - x$, for $x \in R$ and η a constant, and $h(x) = I[x \leq t]$, then $F^* = P((-\infty, t])$ and its Bayes estimate yields

$$\widehat{F}_\alpha^*(t) = p_n F_0(t) + (1 - p_n)\widehat{F}_n^*(t), \tag{2.92}$$

where, $F_0(t) = Q((-\infty, t])$ and $\widehat{F}_n^*(t)$ is η-symmetrized version of the empirical distribution,

$$\widehat{F}_n^*(t) = \frac{1}{2n} \sum_{i=1}^{n} \delta_{X_i}\big((-\infty, t]\big) + \delta_{2\eta - X_i}\big((-\infty, t]\big). \tag{2.93}$$

$\widehat{F}_\alpha^*$ is identical to the one obtained by Dalal (1979a).

In the second paper, Yamato (1987) using the alternative definition of the Dirichlet Invariant process generalizes the above treatment to an arbitrary degree s of estimable parameters in one sample case. As an example of this result, the Bayes estimate of ϕ_1 under L_2 loss is obtained as

$$\widehat{\phi}^*_{1\alpha} = p_n \int h(x) dQ(x) + \frac{(1-p_n)}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} h(g_j X_i), \tag{2.94}$$

wherein $\frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} h(g_j X_i)$ is the $\mathcal{G}$-invariant U-statistic based on kernel h.

Similarly, the Bayesian estimator for an estimable parameter of degree 2, φ_2 is obtained. Assume that

$$\begin{aligned} &\int_{\mathfrak{X}^2} h(x,y) dQ(x) dQ(y) < \infty, \\ &\qquad \int_{\mathfrak{X}} h(x, gx) dQ(x) < \infty \quad \text{for any } g \in \mathcal{G}, \end{aligned} \tag{2.95}$$

and let $X_1, \ldots, X_n$ be a sample from P, $P \in \mathcal{DGI}(\alpha)$. Then the Bayes estimate of φ_2 under L_2 loss is (Yamato 1986, 1987)

$$\begin{aligned} \widehat{\phi}^*_{2\alpha} = \frac{M+n}{M+n+1} \Bigg[& p_n^2 \int_{\mathfrak{X}^2} h(x,y) dQ(x) dQ(y) \\ & + \frac{2p_n(1-p_n)}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} \int_{\mathfrak{X}} h(x, g_j X_i) dQ(x) \\ & + \frac{(1-p_n)^2}{n^2k^2} \sum_{i_1,i_2} \sum_{j_1,j_2} h(g_{j_1} X_{i_1}, g_{j_2} X_{i_2}) \Bigg] \\ & + \frac{1}{k(M+n+1)} \Bigg[p_n \sum_{j=1}^{k} \int_{\mathfrak{X}} h(x, g_j x) dQ(x) \\ & + \frac{(1-p_n)}{nk} \sum_{i} \sum_{j_1,j_2} h(g_{j_1} X_i, g_{j_2} X_i) \Bigg]. \end{aligned} \tag{2.96}$$

If we let M go to zero, the above estimator reduces to

$$\widehat{\phi}^{**}_2 = \frac{1}{n(n+1)k^2} \sum_{j_1,j_2} \Bigg[\sum_{i_1,i_2} h(g_{j_1} X_{i_1}, g_{j_2} X_{i_2}) + \sum_{i} h(g_{j_1} X_i, g_{j_2} X_i) \Bigg], \tag{2.97}$$

and if we replace Dirichlet Invariant with Dirichlet process, clearly the estimator reduces to

$$\widehat{\varphi}_{2D}^{**} = \frac{1}{n(n+1)}\left[\sum_{i_1,i_2} h(X_{i_1}, X_{i_2}) + \sum_i h(X_i, X_i)\right]. \tag{2.98}$$

For illustrative purposes, Yamato takes several different forms of $h(x, y)$ and derives the Bayes estimates of the resulting parameters. For example if we take $h(x, y) = |x - y|$ and $\mathfrak{X} = R$, then $\theta = \int_{R^2} |x - y| dP(x)dP(y)$ is the coefficient of mean difference of the distribution P. On the other hand if $h(x, y) = (x_1 - y_1)(x_2 - y_2)/2$ with $x = (x_1, x_2)$ and $y = (y_1, y_2)$, then

$$\theta = \int_{R^2} x_1 x_2 dP(x_1, x_2) - \int_R x_1 dP(x_1, x_2) \int_R x_2 dP(x_1, x_2) \tag{2.99}$$

is the covariance of the distribution P.

In another example, he takes $h(x, y) = \psi((x_1 - y_1)(x_2 - y_2))$ with $x = (x_1, x_2)$ and $y = (y_1, y_2)$, and $\psi = I[t > 0]$. Then $\theta = 2P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - 1$ is a measure of the correlation between (X_1, Y_1) and (X_2, Y_2) or concordance. Taking $\mathcal{G} = \{e, g\}$ with $e(x_1, x_2) = (x_1, x_2)$, $g(x_1, x_2) = (x_2, x_1)$, for $(x_1, x_2) \in R^2$, he derives the Bayes estimate. When $M \to 0$, this estimator reduces to the non-Bayesian estimator (Randles and Wolf 1979), namely

$$\begin{aligned}\widehat{\theta} = \frac{1}{n(n+1)}\Big[&\#\text{ of }\left\{\text{pairs }(i, j) : (X_i - X_j)(Y_i - Y_j) > 0, 1 \le i < j \le n\right\}\\ &+ \#\text{ of }\left\{\text{pairs }(i, j) : (X_i - Y_j)(Y_i - X_j) > 0, 1 \le i < j \le n\right\}\\ &+ \#\{i : X_i = Y_i, 1 \le i \le n\} - n\Big].\end{aligned} \tag{2.100}$$

Empirical Bayes Estimation of $\boldsymbol{\phi(P)}$ Earlier empirical Bayes estimation results derived for $F(t)$ by Korwar and Hollander (1976), Hollander and Korwar (1976), and for $P(X \le Y)$ by Phadia and Susarla (1979) under $\mathcal{D}(\alpha)$ prior were reported. Tiwari and Zalkikar (1985) generalize these results by replacing the indicator function of the sets $(-\infty, x]$ and $[X \le Y]$ by arbitrary measurable functions $h(x)$ and $h(x, y)$. Specifically, the empirical Bayes estimation of estimable parameters of degree one and two of an unknown probability measure on $(R, \mathcal{B})$ is treated, and asymptotically optimal results with rate of convergence $O(n^{-1})$ of these estimators were established. In proving these results they used the Sethuraman (1994) representation for the Dirichlet process.

The Bayesian estimator of ϕ_1 based on a sample $\mathbf{X}_{n+1}$ of size m at the $(n+1)$-th stage was obtained earlier as

$$\widehat{\phi}_{1\alpha} = p_m \phi_0 + (1 - p_m) U_{n+1}, \tag{2.101}$$

where $\phi_0 = \int h d\overline{\alpha}$, $p_m = M/(M + nm)$ and $U_{n+1} = \frac{1}{m}\sum_{j=1}^m h(X_{n+1,j})$.

To estimate ϕ_1 at the $(n + 1)$-th stage on the basis of $(\mathbf{X}_1, \ldots, \mathbf{X}_{n+1})$, we may use the techniques of Sect. 2.2.4 to estimate first ϕ_0 from the previous n copies and M by Zehnwirth's approach. Substituting these estimates, the empirical Bayes estimator of ϕ_1 at the $(n+1)$-th stage is given by

$$\widehat{\phi}_{\alpha 1,n+1} = \widehat{p}_m \sum_{i=1}^{n} \frac{U_i}{n} + (1-\widehat{p}_m)U_{n+1}, \tag{2.102}$$

where, for the samples $\mathbf{X}_i$, $i = 1, 2, \ldots, n$, $U_i = \frac{1}{m}\sum_{j=1}^{m} h(X_{ij})$, $\widehat{p}_m = \widehat{M}_n/(\widehat{M}_n + m)$, $\widehat{M}_n = \max(0, m(\mathcal{F}_n - 1)^{-1})$, and $\mathcal{F}_n$ is the F-ratio statistics in one-way ANOVA table based on the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. For this estimator, they also established the asymptotic optimality relative to α with rate of convergence $O(n^{-1})$.

Similarly they consider the empirical Bayes estimation of ϕ_2 with $h(x, y) = 0$ whenever $x = y$. If M is known, the empirical Bayes estimator of ϕ_2 at the $(n+1)$-th stage is

$$\widehat{\phi}_{\alpha 2,n+1}(P) = \frac{M+m}{M+m+1}\left[p_m^2 \frac{M+1}{M} \sum_{i=1}^{n} \frac{U_{2i}}{n} + 2p_m(1-p_m) \sum_{k=1}^{n}\sum_{i=1}^{n} \frac{U_{n+1,i,k}}{mn} \right.$$
$$\left. + (1-p_m)^2 \sum_{1\le j\ne k\le m} \frac{h(X_{n+1,j}, X_{n+1,k})}{m^2}\right], \tag{2.103}$$

where for the i-th sample, $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,m})$, $i = 1, 2, \ldots, n$ and,

$$U_{2i} = \frac{1}{m(m-1)} \sum_{1\le j\ne k\le m} h(X_{i,j}, X_{i,k}), \tag{2.104}$$

$$U_{n+1,i,k} = \frac{1}{m}\sum_{j=1}^{m} h(X_{n+1,k}, X_{i,j}), \quad i = 1, 2, \ldots, n. \tag{2.105}$$

Under the assumption that $\int h^2(x, y) d\overline{\alpha}(x) d\overline{\alpha}(y)$ exists and is finite, they show that the sequence $\{\widehat{\phi}_{2,n+1}\}$ is asymptotically optimal relative to α with the rate of convergence $O(n^{-1})$.

Ghosh (1985) also considered the empirical Bayes estimation of $\phi_2(P)$ and computed exact Bayes risk for Bayes and empirical Bayes estimators. He showed that Dalal and Phadia (1983) result for the estimation of concordance coefficient can be obtained as a special case of his result.

Again for the case when M is unknown, Tiwari and Zalkikar (1985) use Zehnwirth's estimate for M and established similar result (see their paper for details). They also obtained the empirical Bayes estimator for ς and proved its asymptotic optimality with rate of convergence $O(n^{-1})$.

Remark 2.4 Asymptotic optimality of the empirical Bayes estimators of variance and the mean deviation about the mean of P can be derived from the above result by taking $h(x, y) = \frac{1}{2}(x-y)^2$ and $|x-y|$, respectively.

Similar empirical Bayes treatment is also given in Ghosh et al. (1989), where the main idea was to use past as well as the current data in estimating the parameters of the Dirichlet process prior. They show that by doing this, we get improved estimators in terms of smaller risks.

2.8 Two-Sample Problems

Suppose we have two independent samples, $X_1, \dots, X_{n_1}$ from F and $Y_1, \dots, Y_{n_2}$ from G. In this section we consider the Bayesian estimation of certain functionals of F and G with respect to the Dirichlet priors $\mathcal{D}(\alpha_1)$ and $\mathcal{D}(\alpha_2)$, respectively.

2.8.1 Estimation of $P(X \le Y)$

Ferguson (1973) derived the Bayesian estimator of $\Delta = P(X \le Y) = \int F dG$ under the squared error loss L_2. Let $F \in \mathcal{D}(\alpha_1)$ and independently, $G \in \mathcal{D}(\alpha_2)$. Then for the no-sample problem the estimate of Δ is given by $\Delta_0 = \mathcal{E}(\Delta) = \int F_0 dG_0$ where $F_0 = \mathcal{E}(F)$ and $G_0 = \mathcal{E}(G)$, and the expectation is taken with respect to the Dirichlet priors. Given the samples $X_1, \dots, X_{n_1} \sim F$ and $Y_1, \dots, Y_{n_2} \sim G$, we update the estimate Δ_0 and obtain the Bayesian estimate as $\widehat{\Delta} = \int \widehat{F}_{\alpha_1} d\widehat{G}_{\alpha_2}$, where $\widehat{F}_{\alpha_1}$ and $\widehat{G}_{\alpha_2}$ are Bayes estimators of F and G, respectively as obtained in Sect. 2.2.1. Simplifying further we get

$$\widehat{\Delta}_{\alpha_1\alpha_2}(\mathbf{X}, \mathbf{Y}) = p_{1n_1} p_{2n_2} \Delta_0 + p_{1n_1}(1 - p_{2n_2}) \frac{1}{n_2} \sum_{i=1}^{n_2} F_0(Y_i)$$
$$+ (1 - p_{1n_1}) p_{2n_2} \frac{1}{n_1} \sum_{i=1}^{n_1} G_0(X_i -) + (1 - p_{1n_1})(1 - p_{2n_2}) \frac{1}{n_1 n_2} U, \tag{2.106}$$

where $p_{1n_1} = \alpha_1(R)/(\alpha_1(R) + n_1)$; $p_{2n_2} = \alpha_2(R)/(\alpha_2(R) + n_2)$, and

$$U = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} I_{(-\infty, Y_j]}(X_i)$$

is the Mann-Whitney statistic. When both $\alpha_1(R)$ and $\alpha_2(R)$ tend to zero, $\widehat{\Delta}_{\alpha_1\alpha_2}$ reduces to the usual nonparametric estimate $(1/(n_1 n_2))U$.

Hollander and Korwar (1976) extend this estimator to the empirical Bayes estimator. Assume that α_1 and α_2 are unknown except for $\alpha_1(R)$ and $\alpha_2(R)$ which are specified, and that we have n copies of data available from the first n stages and are required to estimate Δ at the $(n + 1)$-th stage. As in one sample case, they estimate α_1 and α_2 from the first n-stage data $\mathbf{X}_i = (X_{i1}, \dots, X_{in_1})$ and $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_2})$

for $i = 1, 2, \ldots, n$ and propose the following estimator

$$\begin{aligned}\widehat{\Delta}_{\alpha_1\alpha_2 n}(\mathbf{X},\mathbf{Y}) &= p_{1n_1}p_{2n_2}\frac{1}{n^2 n_1 n_2}\sum_{j=1}^{n}\sum_{k=1}^{n_2}\sum_{i=1}^{n}\sum_{l=1}^{n_1} I_{(-\infty,Y_{jk}]}(X_{il})\\ &\quad + p_{1n_1}(1-p_{2n_2})\frac{1}{nn_1n_2}\sum_{k=1}^{n_2}\sum_{i=1}^{n}\sum_{l=1}^{n_1} I_{(-\infty,Y_{n+1k}]}(X_{il})\\ &\quad + (1-p_{1n_1})p_{2n_2}\frac{1}{n_1}\sum_{l=1}^{n_1}\left\{1-\frac{1}{nn_2}\sum_{j=1}^{n}\sum_{k=1}^{n_2} I_{(-\infty,X_{n+1l}]}(Y_{jk})\right\}\\ &\quad + (1-p_{1n_1})(1-p_{2n_2})\frac{1}{n_1n_2}\sum_{l=1}^{n_1}\sum_{k=1}^{n_2} I_{(-\infty,Y_{n+1k}]}(X_{n+1l}).\end{aligned} \tag{2.107}$$

Finally, they show that $\widehat{\Delta}_{\alpha_1\alpha_2 n}$ is asymptotically optimal with respect to α_1 and α_2. Clearly, when $\alpha_1(R)$ and $\alpha_2(R)$ are also unknown, they could be estimated as indicated earlier, and the above estimator may be adjusted accordingly.

2.8.2 Estimation of the Difference Between Two CDFs

A measure of the difference between two distributions functions F and G, is defined by

$$d(F,G) = \int \big(F(t)-G(t)\big)^2 d\left(\frac{F(t)+G(t)}{2}\right), \tag{2.108}$$

which is somewhat difficult to handle. However, if the distributions are continuous on R, then it can be written as

$$d(F,G) = \frac{4}{3} - \left[\int G(t)dF^2(t) + \int F(t)dG^2(t)\right]. \tag{2.109}$$

Based on two independent samples, $X_1, \ldots, X_{n_1}$ from F and $Y_1, \ldots, Y_{n_2}$ from G, Yamato (1975) considered the problem of Bayesian estimation of $d(F,G)$ under the squared error loss $L(d(F,G), \widehat{d}(F,G)) = (d(F,G) - \widehat{d}(F,G))^2$. In order to use the latter version of the definition, he defines linearized Dirichlet process as priors for F and G which are assumed to be continuous. Following Doksum (1972), he defines a linearized Dirichlet process as follows. For reals $a < b$, consider the partition π of (a,b), $a = t_1 < t_2 < \ldots < t_k = b$ and denote the norm of the partition as $\|\Delta\pi\| = \max_{1\le i\le k-1}|t_{i+1}-t_i|$. Let α be a finite measure on $(R,\mathcal{B})$ with support (a,b). Let H_0 be a realization of the Dirichlet process with parameter α such that $H_0(a) = 0$ and $H_0(b) = 1$ with probability one. Given the partition π, the joint distribution of the corresponding increments of the distribution function has a Dirichlet distribution. With this formulation, he defines a linearized Dirichlet process as follows:

Definition 2.5 (Yamato) H is said to be a linearized Dirichlet process with parameter α and partition π, when H is linear between the points $(t_1, H_0(t_1)), \ldots, (t_k, H_0(t_k))$ and $H_0(t_i)$, $i = 1, 2, \ldots, k$ are the realization of the Dirichlet process with parameter α having support (a, b) and partition π, with $a = t_1$ and $b = t_k$.

Assume F and G as independent linearized Dirichlet processes with parameters α_1 and α_2, respectively, and partition π. Then under the squared error loss, the Bayes estimate is given by the posterior mean,

$$\mathcal{E}\big[d(F, G)|X_1, \ldots, X_{n_1} \text{ and } Y_1, \ldots, Y_{n_2}\big]. \tag{2.110}$$

To evaluate this expectation, he defines (pseudo Bayesian estimators)

$$\begin{aligned} \widehat{F}_{\alpha_1 n_1}(t) &= p_{n_1} F_0(t) + (1 - p_{n_1})\widehat{F}_{n_1}(t), \\ \widehat{G}_{\alpha_2 n_2}(t) &= p_{n_2} G_0(t) + (1 - p_{n_2})\widehat{G}_{n_2}(t), \end{aligned} \tag{2.111}$$

on the interval (a, b), with $\widehat{F}_{\alpha_1 n_1}(t) = \widehat{G}_{\alpha_2 n_2}(t) = 0$ for $t \leq a$, and $\widehat{F}_{\alpha_1 n_1}(t) = \widehat{G}_{\alpha_2 n_2}(t) = 1$ for $t \geq b$, with probability one, where $p_{n_1} = \alpha_1(R)/(\alpha_1(R) + n_1)$, $p_{n_2} = \alpha_2(R)/(\alpha_2(R) + n_2)$, $\widehat{F}_{n_1}$ and $\widehat{G}_{n_2}$ are the empirical distribution functions of the samples $\mathbf{X}$ and $\mathbf{Y}$, respectively, $F_0(t) = \alpha_1(t)/\alpha_1(R)$ and $G_0(t) = \alpha_2(t)/\alpha_2(R)$. Denoting by $\widehat{F}_{\alpha_1 n_1, \Delta}$ and $\widehat{G}_{\alpha_2 n_2, \Delta}$, the linearized versions of $\widehat{F}_{\alpha_1 n_1}$ and $\widehat{G}_{\alpha_2 n_2}$, respectively on the partition π, he evaluates the above expectation obtaining the Bayesian estimator of $d(F, G)$ on the interval (a, b) as

$$\begin{aligned} \widehat{d}_{\alpha_1\alpha_2}(F, G) = \frac{4}{3} &- \frac{\alpha_1(R) + n_1}{\alpha_1(R) + n_1 + 1} \int_a^b \widehat{G}_{\alpha_2 n_2, \Delta}(t) d\widehat{F}^2_{\alpha_1 n_1, \Delta}(t) \\ &- \frac{1}{\alpha_1(R) + n_1 + 1} \left\{ \frac{2}{3} \int_a^b \widehat{G}_{\alpha_2 n_2, \Delta}(t) d\widehat{F}_{\alpha_1 n_1, \Delta}(t) \right. \\ &\left. + \frac{1}{3} \sum_1^{k-1} \widehat{G}_{\alpha_2 n_2}(t_{i+1}) \big[\widehat{F}_{\alpha_1 n_1}(t_{i+1}) - \widehat{F}_{\alpha_1 n_1}(t_i)\big] \right\} \\ &- \frac{\alpha_2(R) + n_2}{\alpha_2(R) + n_2 + 1} \int_a^b \widehat{F}_{\alpha_1 n_1, \Delta}(t) d\widehat{G}^2_{\alpha_2 n_2, \Delta}(t) \\ &- \frac{1}{\alpha_2(R) + n_2 + 1} \left\{ \frac{2}{3} \int_a^b \widehat{F}_{\alpha_1 n_1, \Delta}(t) d\widehat{G}_{\alpha_2 n_2, \Delta}(t) \right. \\ &\left. + \frac{1}{3} \sum_1^{k-1} \widehat{F}_{\alpha_1 n_1}(t_{i+1}) \big[\widehat{G}_{\alpha_2 n_2}(t_{i+1}) - \widehat{G}_{\alpha_2 n_2}(t_i)\big] \right\}. \end{aligned} \tag{2.112}$$

Finally, taking the limit $\|\Delta\pi\| \to 0$, the above estimator reduces to

$$\widehat{d}_{\alpha_1\alpha_2}(F,G) = \frac{4}{3} - \frac{1}{\alpha_1^* + 1}\left\{\int_a^b \widehat{G}_{\alpha_2 n_2}(t)d\widehat{F}_{\alpha_1 n_1}(t) + \alpha_1^* \int_a^b \widehat{G}_{\alpha_2 n_2}(t)d\widehat{F}^2_{\alpha_1 n_1}(t)\right\}$$
$$- \frac{1}{\alpha_2^* + 1}\left\{\int_a^b \widehat{F}_{\alpha_1 n_1}(t)d\widehat{G}_{\alpha_2 n_2}(t) + \alpha_2^* \int_a^b \widehat{F}_{\alpha_1 n_1}(t)d\widehat{G}^2_{\alpha_2 n_2}(t)\right\}, \tag{2.113}$$

where $\alpha_1^* = \alpha_1(R) + n_1$ and $\alpha_2^* = \alpha_2(R) + n_2$.

This estimator is derived on the basis of a particular prior distribution with the interval (a,b) as it's support. In general, when F and G are continuous, the author proposes an estimator of $d(F,G)$ as $\widehat{d}(F,G)$ with the range of integrals replaced in the above formula by $-\infty$ to ∞. By letting $\alpha_1(R)$ and $\alpha_2(R)$ tend to zero, we get a non-Bayesian estimator of $d(F,G)$.

It should be noted that the above formula (2.113) for the difference between two distribution is valid if and only if F and G are continuous. For this reason, the author used the linearized Dirichlet processes as priors in deriving the Bayes estimate, and passing through the limit of the Bayes estimate yielded the above estimator $\widehat{d}_{\alpha_1\alpha_2}(F,G)$ (with the integrals $\int_{-\infty}^{\infty}$). However, the author argues that if we define the difference as the above quantity regardless of the distribution functions being continuous or not, and assign Dirichlet priors to them, the direct computation will show that the resulting estimate is equal to the above estimate.

2.8.3 Estimation of the Distance Between Two CDFs

When F and G are continuous distribution functions, the horizontal distance between F and G is defined as $\Delta(x) = G^{-1}(F(x)) - x$, for a real number x. Hollander and Korwar (1982) consider a one-sample problem where G is assumed to be known and only a random sample of size n from F is available to estimate Δ. Although F is continuous, they assume $F \in \mathcal{D}(\alpha)$. Under the loss function L_1, the Bayes estimator for the no-sample problem is found by minimizing the integrand of

$$\mathcal{E}\big(L(\Delta,\widehat{\Delta})\big) = \int \mathcal{E}\big(\Delta(x) - \widehat{\Delta}(x)\big)^2 dW(x) \tag{2.114}$$

yielding $\widehat{\Delta}_0(x) = \mathcal{E}(\Delta(x)) = \mathcal{E}\{G^{-1}F(x)\} - x$. For a sample of size n from F, the Bayesian estimator is obtained simply by updating α.

If G is assumed to be an exponential distribution, $G(x) = 1 - e^{-\lambda x}$, $x > 0, \lambda > 0$, then $\widehat{\Delta}_0$ is

$$\widehat{\Delta}_0(x) = \frac{1}{\lambda \cdot B(\alpha',\beta')} \cdot \int_0^1 \sum_{j=1}^{\infty} \frac{y^{\alpha'+j-1}(1-y)^{\beta'-1}}{j} dy - x$$
$$= \frac{1}{\lambda} \cdot \sum_{j=1}^{\infty} \frac{B(\alpha'+j,\beta')}{j \cdot B(\alpha',\beta')} - x, \tag{2.115}$$

where $\alpha' = \alpha((-\infty, x])$, $\beta' = \alpha((x, \infty)) = \alpha(R) - \alpha'$. Now for a sample of size n from F, the Bayes estimator is the above expression $\widehat{\Delta}_0(x)$ with α' and β' replaced by $\alpha^* = \alpha' + \sum_{i=1}^{n} \delta_{X_i}$ and $\beta^* = \alpha(R) + n - \alpha^*$, respectively, their updated versions.

2.9 Hypothesis Testing

In applications of the Dirichlet process prior so far, we have discussed mainly the estimation of an unknown distribution function F or a parameter φ which is a function of the unknown probability measure P. Ferguson (1973) pointed out the difficulty of using the Dirichlet Process prior in hypothesis testing problems. However, Susarla and Phadia (1976) were able to show how such problems can be handled. The idea was to replace the usual 0–1 loss with a smoother loss function based on a known weight function W. Thus their approach to the problem of the hypothesis testing was from a decision theoretic point of view—a first as far as we know. Their method can be extended to treat multiple decision theoretic problems as well. This is described now.

2.9.1 Testing $H_0 : F \leq F_0$

Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random sample from the distribution function F. Let F_0 be a known distribution function. Consider the problem of testing hypothesis $H_0 : F \leq F_0$ against the alternative $H_1 : F \nleq F_0$ when the loss function L is given by

$$L(F, a_0) = \int (F - F_0)^+ dW \quad \text{and} \quad L(F, a_1) = \int (F - F_0)^- dW, \tag{2.116}$$

where $L(F, a_i)$ indicates the loss incurred when action a_i (deciding in favor of H_i) is taken for $i = 0, 1$, W is a weight function, $\mathrm{a}^+ = \max\{a, 0\}$ and $\mathrm{a}^- = -\min\{a, 0\}$ for any real number a. Assume $F \in \mathcal{D}(\alpha)$. Let $\delta(\mathbf{X}) = \mathcal{P}\{\text{taking action } a_0 \mid \mathbf{X}\}$. Then the Bayes risk of δ against $\mathcal{D}(\alpha)$ is

$$r_m(\delta, \alpha) = \int \mathcal{E}\big[L(F, a_0) - L(F, a_1) \mid \mathbf{X}\big]\delta(\mathbf{X}) dQ_m(\mathbf{X}) + \mathcal{E}\big[L(F, a_1)\big], \tag{2.117}$$

where Q_m is the unconditional distribution of $\mathbf{X}$ and the expectation is taken with respect to $\mathcal{D}(\alpha)$. Hence a Bayes rule against $\mathcal{D}(\alpha)$ which minimizes the above risk is given by $\delta_m(\mathbf{X}) = I[\Delta_m(\mathbf{X}) \leq 0]$ where $\Delta_m(\mathbf{X}) = \int \mathcal{E}[F(u) - F_0(u) \mid \mathbf{X}] dW(u)$ and the minimum Bayes risk is

$$r_m^*(\alpha) = \int_{[\Delta_m(\mathbf{X}) \leq 0]} \Delta_m(\mathbf{X}) dQ_m(\mathbf{X}) + \mathcal{E}\big[L(F, a_1)\big]. \tag{2.118}$$

If α is known, $\Delta_m(\mathbf{X})$ can be easily evaluated since for each u, $F(u)|\mathbf{X} = \mathbf{x} \sim Be(\alpha(-\infty, u] + \sum_{i=1}^{m} I[X_i \le u], \alpha(u, \infty) + \sum_{i=1}^{m} I[X_i > u])$.

When α is unknown, we can use the empirical Bayes method. Let $\alpha(R) = 1$ and assume the usual set up for the empirical Bayes estimation with sample size m_i at the i-th stage. Then an empirical Bayes rule at the $(n+1)$-th stage is given by

$$\xi_{n+1}(\mathbf{X}_{n+1}) = \mathcal{P}\{\text{accepting } a_0 \mid \mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{X}_{n+1}\}, \tag{2.119}$$

Let $\widehat{\Delta}_n(\mathbf{X}_{n+1})$ be an estimate based on $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ of

$$\Delta_{m_{n+1}}(\mathbf{X}_{n+1}) = \int \mathcal{E}\big[F_{n+1}(u) - F_0(u) \mid \mathbf{X}_{n+1}\big] dW(u)$$

given by

$$\widehat{\Delta}_n(\mathbf{x}_{n+1}) + \int F_0 dW$$
$$= \int \frac{\{\widehat{\alpha}(-\infty, u] + \sum_{i=1}^{m_{n+1}} I[X_{n+1,i} \le u]\} dW(u)}{(1 + m_{n+1})}, \tag{2.120}$$

where $\widehat{\alpha}(-\infty, u] = n^{-1} \sum_{j=1}^{n} m_j^{-1} \sum_{i=1}^{m_j} [X_{j,i} \le u]$. Let

$$\xi_n(\mathbf{x}_{n+1}) = I\big[\widehat{\Delta}_n(\mathbf{X}_{n+1}) \le 0\big],$$

and let $r_{n+1}(\xi_n)$ denote the risk of using ξ_n to decide about F_{n+1}. Then Susarla and Phadia (1976) have proved that $r_{n+1}(\xi_n) - r^*_{m_{n+1}}(\alpha) \le n^{-1/2}$.

When $\alpha(R)$ is unknown, they estimate it by a consistent estimator $\widehat{\alpha}(R) = (\log m_n)^{-1}\{\#$ of distinct observations in $\mathbf{X}_n\}$ (see property 16 of Sect. 1.2). Let ξ^*_n be the rule obtained by substituting this estimator in ξ_n with

$$\widehat{\Delta}_n(\mathbf{X}_{n+1}) + \int F_0 dW$$
$$= \int \frac{\{\widehat{\alpha}(R)\widehat{\alpha}(-\infty, u] + \sum_{i=1}^{m_{n+1}} [X_{n+1,i} \le u]\} dW(u)}{(\widehat{\alpha}(R) + m_{n+1})}. \tag{2.121}$$

Then they prove the following asymptotic result. Let α be nonatomic and $m_{n+1} \to \infty$ as $n \to \infty$. Then $r_{n+1}(\xi^*_n) - r^*_{m_{n+1}}(\alpha) = O((m_{n+1})^{-1}(\min\{\log m_n, n\})^{-1/2})$.

In addition, they have shown that some of these procedures are component-wise admissible and have also discussed the extension of their results to the multiple action problem.

2.9.2 Testing Positive Versus Nonpositive Dependence

In the bivariate distribution case, we come across the problem of testing positive dependence versus nonpositive dependence. Let $F(x, y)$ be a bivariate distribution function defined on $(R^2, \mathcal{B}^2)$ with marginal CDFs $F_X(x)$ and $F_Y(y)$, respectively. The objective is to test the following hypotheses.

$$\begin{aligned} H_0 &: F(x,y) \geq F_X(x)F_Y(y) \quad \text{for all } (x,y) \text{ in } R^2 \\ H_1 &: F(x,y) < F_X(x)F_Y(y) \quad \text{for all } (x,y) \text{ in } R^2, \end{aligned} \tag{2.122}$$

under the loss function

$$\begin{aligned} L(F,a_0) &= \int \big(F(x,y) - F_X(x)F_Y(y)\big)^- dW(x,y) \\ L(F,a_1) &= \int \big(F(x,y) - F_X(x)F_Y(y)\big)^+ dW(x,y), \end{aligned} \tag{2.123}$$

where the actions a_0 and a_1 are to accept H_0 and H_1, respectively, W is a known weight function on R^2. For given observations $(\mathbf{x}, \mathbf{y})$, denote by $\theta(\mathbf{x}, \mathbf{y})$ the probability of taking action a_0. Then Dalal and Phadia (1983) have shown that the Bayes rule against $\mathcal{D}(\alpha)$ is given by

$$\theta(\mathbf{x}, \mathbf{y}) = I_{[\Delta_n(\mathbf{x},\mathbf{y})]}, \tag{2.124}$$

where

$$\begin{aligned} \Delta_n(\mathbf{x}, \mathbf{y}) &= \mathcal{E}\big[L(F,a_0) - L(F,a_1) \mid (\mathbf{X}, \mathbf{Y})\big] \\ &= \int \big[\mathcal{E}\big(F\big(x', y'\big) - F_X\big(x'\big)F_Y\big(y'\big)\big) \mid (\mathbf{X}, \mathbf{Y})\big] dW\big(x', y'\big). \end{aligned} \tag{2.125}$$

Here the expectation is taken with respect to the posterior Dirichlet process with parameter $\alpha + \sum_{i=1}^n \delta_{(x_i, y_i)}$. Let $\alpha = MQ$, G_0 be a CDF corresponding to Q, $G^* = p_n G_0 + (1 - p_n)\widehat{G}_n$, where $\widehat{G}_n$ is the empirical CDF based on the n observations (x_i, y_i), $i = 1, 2, \ldots, n$, and $p_n = M/(M + n)$. Then the integrand can be evaluated as

$$\begin{aligned} &\frac{MG_0(x', y') + \sum_{i=1}^n \delta_{(x_i, y_i)}((-\infty, x'] \times (-\infty, y'])}{M + n} \\ &\quad - \frac{G^*(x', y') + MG_X^*(x')G_Y^*(y')}{M + n + 1}, \end{aligned} \tag{2.126}$$

and hence $\Delta_n(\mathbf{x}, \mathbf{y})$ can be evaluated. As in the case of estimating the concordance coefficient above, the empirical Bayes solution can be carried out here as well when α is not known, with M known or unknown.

Testing the Hypothesis $H_0 : F \leq G$ Against the Alternative $H_1 : F \nleq G$ An analog of the test discussed in Sect. 2.9.1 in a two-sample situation is to test the hypothesis $H_0 : F \leq G$ against the alternative $H_1 : F \nleq G$. This topic is covered more generally in Sect. 3.6 based on samples with right-censored observations. Its application to the uncensored data as a special case is obvious and therefore it will not be presented here.

2.9.3 A Selection Problem

Consider the following selection problem. We are given k samples, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik_i})$ distributed according to F_i, $i = 1, 2, \ldots, k$, and a sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ known to have come from one of the k distributions. The problem is to find from which one. Antoniak (1974) considered this problem and provided a Bayes solution. Let $\mathfrak{X}$ be a set of nonnegative integers and $\sigma(\mathfrak{X})$ be the corresponding σ-algebra generated by the singleton sets. Assume that for $i = 1, 2, \ldots, k$, $F_i \sim \mathcal{D}(\alpha_i)$. For technical reasons, each α_i is taken to be a discrete measure with the same support and defined on $\sigma(\mathfrak{X})$ with $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \ldots)$ and $\alpha_i(\{j\}) = \alpha_{ij}$, $|\alpha_i| = \sum_{j=0}^{\infty} \alpha_{ij}$. Let π_j be the prior probability that the sample $Y_1, \ldots, Y_n$ came from F_j, $j = 1, 2, \ldots, k$. Let $L(i, j)$ be the associated loss function in deciding $\mathbf{Y}$ as coming from F_i when in fact it is from F_j. The goal is to seek a non-randomized decision rule which minimizes the expected loss. First note that $F_i|\mathbf{X}_i \sim \mathcal{D}(\alpha_i^*)$, where $\alpha_i^* = (\alpha_{i0}^*, \alpha_{i1}^*, \ldots)$, with $\alpha_{ij}^* = \alpha_{ij} + m_{ij}$, and m_{ij} is the number of X_i's equal to j, $j = 0, 1, \ldots$. The Bayes risk r_i is given by

$$r_i(\boldsymbol{\pi}, \boldsymbol{\alpha}) = \sum_{j=1}^{k} L(i, j) P(j|Y_1, \ldots, Y_n) = \sum_{j=1}^{k} L(i, j) \frac{\pi_j P(\mathbf{Y}|j)}{\sum_{j=1}^{k} \pi_j P(\mathbf{Y}|j)}, \qquad (2.127)$$

where

$$P(\mathbf{Y}|j) = \prod_{l=0}^{\infty} \frac{\alpha_{jl}^{*(k_l)}}{\alpha_l^{*(n)}}, \qquad a^{(n)} = a(a+1)\ldots(a+n-1), n > 0, \qquad (2.128)$$

and k_l is the number of Y's equal to l. The Bayes decision rule selects s, where $r_s = \min r_i$. For the 0–1 loss and uniform prior $\pi_j = 1/k$, the Bayes decision rule is to choose s for which $P(\mathbf{Y}|s) = \max_j P(\mathbf{Y}|j)$.

Chapter 3
Inference Based on Incomplete Data

3.1 Introduction

Most common form of incomplete data is when the observations are censored on the right. Therefore, in this chapter we will be dealing mainly with the right censored data, although estimators based on other sampling schemes will also be presented. A typical problem in the analysis of right censored data may be described as follows. We have a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from an unknown distribution function F defined on $R^+ = (0, \infty)$, and let $Y_1, \ldots, Y_n$ be positive real numbers or random variables (to be specified later). We do not observe X_i's directly but only via $Z_i = \min(X_i, Y_i)$ and $\delta_i = I[X_i \leq Y_i]$, $i = 1, 2, \ldots, n$. Based on $(\mathbf{Z}, \boldsymbol{\delta}) = \{(Z_i, \delta_i)\}_{i=1}^n$, we are required to make various inferences about F or it's function, but mainly the survival function (SF), $S(t) = 1 - F(t)$. No distributional assumptions are made regarding F, and thus the procedures reported here may be considered as nonparametric. In the context of survival analysis, X_i's are known as "uncensored", "real" or "exact" observations while Y_i's are right "censoring" variables. δ_i indicates whether the i-th observation is real or censored.

This problem is encountered in many applications such as industrial, competing risks, life testing, life tables (Kaplan and Meier 1958), biomedical research, survival data analysis (Gross and Clark 1975), etc. However, its application and usefulness in the analysis of survival data arising in clinical research has received wide attention. In the non-Bayesian context, the problem was first considered by Kaplan and Meier (1958) who developed two estimators for estimating S. One of them, known as the product limit (PL) estimator,

$$\widehat{S}_{PL}(u) = \frac{N^+(u)}{n} \prod_{j=1}^{n} \left(\frac{N^+(Z_j) + \lambda_j}{N^+(Z_j)} \right)^{I[\delta_i = 0, Z_j \leq u]}, \tag{3.1}$$

with multiplicities λ_j at Z_j, $j = 1, \ldots, n$, and $N^+(u) = \sum_{i=1}^n I[Z_i > u]$, has been widely popular and studied extensively. It is also used in solving other problems encountered in estimation, prediction, in hypothesis testing, etc. It should be pointed out that unlike in the case of complete (uncensored) data, the estimators of S are in

E.G. Phadia, *Prior Processes and Their Applications*,
DOI 10.1007/978-3-642-39280-1_3, © Springer-Verlag Berlin Heidelberg 2013

the product form. This patterns is remarkably repeated consistently in the Bayesian estimation irrespective of the prior processes used.

We present here the Bayesian approach. Since the parameter of interest here is the distribution function F itself, it will be considered as random and a prior defined over the space of all distribution functions, $\mathcal{F}_+$, will be assigned to F.

Due to its analytical tractability, the Dirichlet process is used extensively as a prior for F in statistical inference problems. But unlike in the case of complete data, the posterior distribution given the right censored data is a mixture of Dirichlet processes. However, when viewed as a neutral to the right process, the Dirichlet process is structurally conjugate with respect to the right censored data. As such, it is shown that for the right censored data, neutral to the right processes are more suitable as priors. They cover the Dirichlet process and beta-Stacy process, among others, and is considered in more detail and relevant formulas are presented.

As in the previous chapter, the organization of material in this chapter is as follows. Estimation of the distribution (survival) function is of prime interest and therefore it is considered first in Sect. 3.2 assuming the Dirichlet process prior. Also estimation of the survival function under different censoring schemes is presented. In Sect. 3.3, the estimation of the survival function based on other priors is considered. In a slight digression, a linear Bayes estimation of the survival function is presented in Sect. 3.4. Various other estimation problems are included in Sect. 3.5. Section 3.6 deals with a hypothesis testing problem and finally, estimation of the survival function incorporating covariates primarily through the Cox model is presented in the last Sect. 3.7.

3.2 Estimation of a SF Based on DP Priors

In this section the pioneering work of Susarla and Van Ryzin (1976) in deriving the Bayesian estimator of a survival function with respect to the Dirichlet process is presented. Empirical Bayes and Bayesian estimators under various other sampling schemes are also considered. Interestingly they all have similar forms and are generalization of the non-Bayesian PL estimator.

3.2.1 Estimation Based on Right Censored Data

Based on the data $(\mathbf{Z}, \boldsymbol{\delta})$, we shall consider in this section the problem of Bayesian estimation of the survival function $S(t)$, under the loss function L_1 used earlier

$$L_1(S, \widehat{S}) = \int_0^\infty \big(S(t) - \widehat{S}(t)\big)^2 dW(t), \tag{3.2}$$

where $W(\cdot)$ is a known weight function on R^+. It is assumed that $X_1, \ldots, X_n$ are iid F, and $Y_1, \ldots, Y_n$ are independent with Y_i distributed as G_i, $i = 1, 2, \ldots, n$.

Assume also that $Y_1, \ldots, Y_n$ are independent of $(F, X_1, \ldots, X_n)$. The strategy employed by Susarla and Van Ryzin was first to take care of real observations by noting that the posterior distribution of F given these observations is again the Dirichlet process with updated parameter. Then the censored observations were dealt with the updated parameter.

Observe that $(\mathbf{Z}, \boldsymbol{\delta})$ are invariant for any permutation of observed pairs (δ_1, Z_1), $\ldots, (\delta_n, Z_n)$. Hence without loss of generality we can (and we shall) rearrange these observations as $(1, Z_1), \ldots, (1, Z_k), (0, Z_{k+1}), \ldots, (0, Z_n)$. Thus $Z_1, \ldots, Z_k$ are uncensored observations and $Z_{k+1}, \ldots, Z_n$ are censored observations. The uncensored observations are taken care of by replacing the parameter α by $\alpha_k = \alpha + \sum_{i=1}^{k} \delta_{Z_i}$. Among the censored observations, let $Z_{(k+1)}, \ldots, Z_{(m)}$ denote the distinct ordered observations with multiplicities λ_j at $Z_{(j)}$, $j = k+1, \ldots, m$, so that $\sum_{j=k+1}^{m} \lambda_j = n - k$. Then, the Bayes estimator of $S(u)$ given the data is the conditional expectation of $S(u)$ given $(0, Z_{k+1}), \ldots, (0, Z_n)$, where the expectation is now performed with respect to the posterior of F given $(1, Z_1), \ldots, (1, Z_k)$ which is $\mathcal{D}(\alpha_k)$. Thus, we have

$$\widehat{S}_\alpha(u) = \mathcal{E}_{\mathcal{D}(\alpha_k)}\left\{S(u) \mid (0, Z_{k+1}), \ldots, (0, Z_n)\right\}. \tag{3.3}$$

After establishing several intermediate steps for the moments of the conditional distribution of $S(u)$, Susarla and Van Ryzin derived the following Bayes estimator of the survival function $S(u)$ for u in the interval $Z_{(l)} \le u < Z_{(l+1)}$, $l = k, \ldots, m$, with $Z_{(k)} = 0$ and $Z_{(m+1)} = \infty$:

$$\widehat{S}_\alpha(u) = \frac{\alpha(u, \infty) + N^+(u)}{\alpha(R^+) + n} \prod_{j=k+1}^{l} \frac{\alpha[Z_{(j)}, \infty) + N^+(Z_{(j)}) + \lambda_j}{\alpha[Z_{(j)}, \infty) + N^+(Z_{(j)})}, \tag{3.4}$$

where $N^+(u)$ is the number of observations greater than u. Alternatively, it can be expressed in the form parallel to (3.1) as

$$\widehat{S}_\alpha(u) = \frac{\alpha(u, \infty) + N^+(u)}{\alpha(R^+) + n} \prod_{j=1}^{n} \left(\frac{\alpha[Z_j, \infty) + N^+(Z_j) + \lambda_j}{\alpha[Z_j, \infty) + N^+(Z_j)}\right)^{I[\delta_i = 0, Z_j \le u]}. \tag{3.5}$$

Several observations are in order here.

Remarks

1. This estimator is also in the product form and looks like the PL estimator of Kaplan and Meier (1958). But unlike the PL estimator it is defined every where.
2. Like the PL estimator it has jumps only at the uncensored observations.
3. Unlike the PL estimator, it is not constant between two uncensored observations but reflects the contribution of the prior information. It is a smoother version of the PL estimator.
4. As $\alpha(R^+) \to 0$, the estimator reduces to the PL estimator.

5. If there are no censored observations (i.e. all G_i's are degenerate at ∞) this estimator reduces to the Bayes estimator given by Ferguson (1973) restricted to R^+.

For a fixed u, the conditional distribution of $F(u)$ given $(\boldsymbol{\delta}, \mathbf{Z})$ is a mixture of beta distributions. Also, for a fixed u such that $Z_{(l)} \leq u < Z_{(l+1)}$, with $l = k, \dots, m$, $Z_{(k)} = 0$ and $Z_{(m+1)} = \infty$, the conditional p-th moment of $S(u)$ given the data is shown to be

$$\mathcal{E}\big[S^p(u) \mid (\boldsymbol{\delta}, \mathbf{Z})\big] = \prod_{s=0}^{p-1} \left\{ \frac{\alpha(u, \infty) + s + N^+(u)}{\alpha(R^+) + s + n} \times \prod_{j=k+1}^{l} \frac{\alpha[Z_{(j)}, \infty) + s + N^+(Z_{(j)}) + \lambda_j}{\alpha[Z_{(j)}, \infty) + s + N^+(Z_{(j)})} \right\}, \tag{3.6}$$

where the inside product is treated as one if $u < Z_{(k+1)}$.

This gives a clue that the conditional distribution of F given $(\boldsymbol{\delta}, \mathbf{Z})$ is a mixture of Dirichlet processes, as indicated by Susarla and Van Ryzin (1976).

In fact Blum and Susarla (1977) confirmed the above conjecture by proving that the posterior distribution of S, given the censored observations, is indeed a mixture of Dirichlet processes and indicated the transition and mixing measures.

Theorem 3.1 (Blum and Susarla) *Let G_j be absolutely continuous or discrete distribution for $j = 1, \dots, n$ and that $Y_1, \dots, Y_n$ are independent of $(F, X_1, \dots, X_n)$. Then the posterior distribution of P given $(\mathbf{Z}, \boldsymbol{\delta})$ is a mixture of Dirichlet processes with transition measure $\beta(\cdot) + \sum_{j=1}^{k-1} \mu_j(\cdot) + I_\cdot(u)$ and mixing measure μ_k, where $\beta(B) = \alpha(B) + \sum_{j=1}^{n-k} I_B(Z_j)$, $\beta([Z_{n-k+1}, \infty))\mu_1(B) = \beta(B \cap [Z_{n-k+1}, \infty))$ and*

$$\mu_l(B) = \frac{\beta(B \cap [Z_{n-k+l}, Z_{n-k+l-1}))}{\beta([Z_{n-k+1}, \infty)) + l - 1} \times \sum_{j=1}^{l-1} \frac{\beta(B \cap [Z_{n-k+j}, Z_{n-k+j-1}))}{\beta([Z_{n-k+1}, \infty)) + j - 1} \prod_{i=j}^{l-1} \frac{\beta([Z_{n-k+i}, \infty)) + i}{\beta([Z_{n-k+i+1}, \infty)) + i} \tag{3.7}$$

for $l = 2, \dots, k$.

If we assume that G_i's are identical to G, that is, $Y_i \sim G$, $i = 1, 2, \dots, n$, and that G is a fixed unknown continuous distribution on R^+, then in Eq. (3.5) $\lambda_j = 1$ for all j and the Bayes estimator $\widehat{S}_\alpha$ becomes

$$\widehat{S}_\alpha(u) = \frac{\alpha(u, \infty) + N^+(u)}{\alpha(R^+) + n} \prod_{j=1}^{n} \left(\frac{\alpha[Z_j, \infty) + N^+(Z_j) + 1}{\alpha[Z_j, \infty) + N^+(Z_j)} \right)^{I[\delta_i = 0, Z_j \leq u]}. \tag{3.8}$$

This representation of $\widehat{S}_\alpha$ is used in proving asymptotic properties of the Bayes estimator. It is shown (Susarla and Van Ryzin 1978b, 1980) that $\widehat{S}_\alpha$ is almost surely consistent with a convergence rate of $O(\log n/n^{1/2})$, and converges weakly to a mean zero Gaussian process. They also give an expression for the covariance matrix.

In the above formulation, X_i and Y_i are assumed to be independent. If they are allowed to be dependent, the marginal distribution of X may not be identifiable. Nevertheless, a Bayesian treatment of the problem is possible and has been carried out by Phadia and Susarla (1983) by assuming a Dirichlet process prior for the joint distribution of (X, Y). This will be further discussed in Sect. 3.5.3.

3.2.2 Empirical Bayes Estimation

Susarla and Van Ryzin (1978a) also considered the empirical Bayes approach in estimating a survival function. For simplicity we consider the case of sample size one only. The general case is straight forward. Thus, at each of $(n+1)$-stages we have one observation and α is unknown, but $\alpha(R^+)$ is known. As in the earlier sections for complete data, α is estimated from n previous stages by $\widehat{\alpha}_n$ and substituted in the Bayes estimator at the $(n+1)$-stage. Thus under the weighted squared error loss function, the empirical Bayes estimator of $S_{n+1}(u)$ is given by

$$\widehat{S}_{n+1}(u) = \begin{cases} \kappa(1+\widehat{\alpha}_n(u,\infty)) & \text{for } u < Z_{n+1} \\ \kappa\widehat{\alpha}_n(u,\infty)(\frac{\widehat{\alpha}_n(Z_{n+1},\infty)+I[\delta_{n+1}=0]}{\widehat{\alpha}_n(Z_{n+1},\infty)}) & \text{for } u \geq Z_{n+1}, \end{cases} \tag{3.9}$$

where $\kappa = (1+\alpha(R^+))^{-1}$ and

$$\frac{\widehat{\alpha}_n(u,\infty)}{\alpha(R^+)} = \frac{1}{nG_j(u)}\sum_{j=1}^{n} I[Z_j > u]. \tag{3.10}$$

Under some mild conditions, it is proved that the above estimator is asymptotically optimal with rate of convergence $O(n^{-1})$.

However, because of $G_j(u)$ in the denominator, their estimator was not monotone in that the estimator was increasing between two censored observations and hence was not a proper survival function. Phadia (1980) proposed a slightly modified estimator which did not have this undesirable property and at the same time, it was also asymptotically optimal with the same rate of convergence $O(n^{-1})$. His estimator is given by the same estimator $\widehat{S}_{n+1}(u)$ as above, but the estimator $\widehat{\alpha}_n$ is replaced by

$$\frac{\widehat{\alpha}_n(u,\infty)}{\alpha(R^+)} = \frac{N^+(u)}{n}\prod_{j=1}^{n}\left(\frac{N^+(Z_j)+1+c}{N^+(Z_j)+c}\right)^{[\delta_j=0, Z_j\leq u]}, \tag{3.11}$$

where c is a positive constant. By a suitable choice of c, a desired level of smoothness in the empirical Bayes estimator may be obtained. It was suggested that a value of $1/2$ to 5 for c to be reasonable, but may depend on some other optimal criterion.

3.2.3 Estimation Based on a Modified Censoring Scheme

In extending the results of Susarla-Van Ryzin (1976) to a more general class of priors, namely, the processes neutral to the right developed by Doksum (1974), Ferguson and Phadia (1979) considered a modified sampling scheme in which the censored observations were classified as "exclusive" if $X > x$ and "inclusive" if $X \geq x$. Assume that the observational data has three forms, m_1 real observations $X_1 = x_1, \ldots, X_{m_1} = x_{m_1}$, m_2 'exclusive censoring' $X_{m_1+1} > x_{m_1+1}, \ldots, X_{m_1+m_2} > x_{m_1+m_2}$, and m_3 'inclusive' censoring $X_{m_1+m_2+1} \geq x_{m_1+m_2+1}, \ldots, X_{m_1+m_2+m_3} \geq x_{m_1+m_2+m_3}$, where $m_1 + m_2 + m_3 = n$, the sample size. The former type is the customary way of defining censoring and is the only type considered in Kaplan and Meier (1958) and Susarla and Van Ryzin (1976). In addition, Ferguson and Phadia assumed the censoring points as given constants and not random variables as assumed by Susarla and Van Ryzin (1976). (Lemma 7.1 in Lo 1993a, shows that the distinction is immaterial as long as the distributions of censored variables are independent of F.) Under this sampling scheme, they derived the posterior mean.

Let $u_1 < u_2 < \ldots < u_k$ be the distinct ordered values among $x_1, \ldots, x_n$; $\delta_1, \ldots, \delta_k$ are number of uncensored observations at $u_1, \ldots, u_k$, respectively; $\lambda_1, \ldots, \lambda_k$ denote the number of 'exclusive' censoring at $u_1, \ldots, u_k$ respectively; $\mu_1, \ldots, \mu_k$ denote the number of 'inclusive' censoring at $u_1, \ldots, u_k$ respectively, so that $\sum_1^k \delta_i = m_1$, $\sum_1^k \lambda_i = m_2$, $\sum_1^k \mu_i = m_3$, $h_j = \sum_{i=j+1}^k (\delta_i + \lambda_i + \mu_i)$ denote the number of x_i greater than u_j; and $j(t)$ denotes the number of u_i less than or equal to t. If $F \in \mathcal{D}(\alpha)$, then the posterior expectation of the survival function $S(t)$ is

$$\mathcal{E}\big(S(t)|data\big) = \frac{\alpha(t,\infty) + h_{j(t)}}{\alpha(R) + n} \times \prod_{i=1}^{j(t)} \frac{(\alpha[u_i,\infty) + h_{i-1})(\alpha(u_i,\infty) + h_i + \lambda_i)}{(\alpha(u_i,\infty) + h_i)(\alpha[u_i,\infty) + h_i + \lambda_i + \delta_i)}. \tag{3.12}$$

The Susarla-Van Ryzin formula is really the above formula with all μ_i's equal to zero, since in that case $h_{i-1} = h_i + \lambda_i + \delta_i$ for $i = 1, \ldots, k$.

3.2.4 Estimation Based on Progressive Censoring

There is a broad class of experiments in which the Z_j's are observed sequentially, and cost and/or time considerations often entail termination of experimentation before all Z_j's have been observed. For example, a study may be curtailed at the k $(= k(n))$-th smallest order statistics $Z_{(k)}$, $1 \leq k \leq n$, and then in effect, the statistician has at his disposal only the data

$$\big\{\big(\delta_i^*, Z_{(i)}\big), 1 \leq i \leq k, Z_{(r)} > Z_{(k)}, r = k+1, \ldots, n\big\}, \tag{3.13}$$

where $\delta_i^* = 0$ or 1 according as $Z_{(i)}$ is a true survival time or censoring time.

Statistical procedure based on this type of data is referred to as progressively censoring scheme and is treated by Tiwari et al. (1988). Assuming that the first l, $1 \le l \le k$, $Z_{(i)}$'s are uncensored observations and proceeding as in (Sect. 3.2.1) gives

$$\begin{aligned}\widehat{S}_k(u) = \mathcal{E}\big\{S(u) \mid (0, Z_{l+1}), \ldots, (0, Z_k), 1 \le l \le k,\\ Z_{(r)} > Z_{(k)}, r = k+1, \ldots, n\big\},\end{aligned} \tag{3.14}$$

where the expectation is taken with respect to $\mathcal{D}(\alpha + \sum_{i=1}^{l} \delta_{Z_{(i)}})$. From Blum and Susarla (1977) one may observe that the conditional distribution of $S(u) \mid (0, Z_{l+1}), \ldots, (0, Z_k)$, $1 \le l \le k$, $Z_{(r)} > Z_{(k)}$, $r = k+1, \ldots, n$, is a mixture of Dirichlet processes (see also, Tiwari et al. 1988). Hence, from Blum and Susarla (1977) or Gardiner and Susarla (1981, 1983) the Bayes estimator $\widehat{S}_k$ in (3.14) becomes

$$\begin{aligned}\widehat{S}_k(u) = {} & \frac{\alpha(u,\infty) + N^+(u) + (n-k)I[Z_{(k)} > u]}{\alpha(R^+) + n}\\ & \times \prod_{j=1}^{k}\left(\frac{\alpha(Z_{(j)},\infty) + N^+(Z_{(j)}) + (n-k) + 1}{\alpha(Z_{(j)},\infty) + N^+(Z_{(j)}) + (n-k)}\right)^{I[\delta_j^*=0, Z_{(j)} \le u]}\\ & \times \left(\frac{\alpha(Z_{(k)},\infty) + (n-k)}{\alpha(Z_{(k)},\infty)}\right)^{[Z_{(k)} \le u]}.\end{aligned} \tag{3.15}$$

Note that by setting $k(n) = n$ in (3.15) yields the Susarla and Van Ryzin estimator (3.5) with $\lambda_j = 1$ for all j. Also, if there are no observed censoring times in the data (3.13), (3.15) reduces to

$$\begin{aligned}\widehat{S}(u) = {} & \left\{\frac{\alpha(u,\infty) + N^+(u) + (n-k)I[Z_{(k)} > u]}{\alpha(R^+) + n}\right\}\\ & \times \left(\frac{\alpha(Z_{(k)},\infty) + (n-k)}{\alpha(Z_{(k)},\infty)}\right)^{I[Z_{(k)} \le u]},\end{aligned} \tag{3.16}$$

which in turn may be viewed as a generalization of Ferguson's (1973) estimator $(\alpha(u,\infty) + N^+(u))/(\alpha(R^+) + n)$, when censoring is absent and $k(n) = n$. For $u \le Z_{(k)}$, (3.15) yields the Kaplan and Meier (1958) Product-Limit estimator, which itself reduces to the empirical survival function in the absence of censoring.

3.2.5 Estimation Based on Record-Breaking Observations

In certain industrial experiments one observes only the successive minima and the number of trials required to obtain the next minima. The objective is to estimate the survival function based on such data. Tiwari and Zalkikar (1991b) considered this problem and obtained the Bayes estimator for the survival function using the Dirichlet process prior.

Let $X_1, \ldots, X_n$ be iid random variables from a continuous distribution function F defined on $R^+ = (0, \infty)$ and let $S(t)$ be the corresponding survival function. The data is observed sequentially and can be represented as $Y_1, K_1, Y_2, K_2, \ldots,$ where Y_i's are successive minima and K_i's are the number of trials required to obtain a subsequent minimum. In harmony with the survival data, this data can be reformulated as follows. Let $Z_1 = X_1$ and $Z_i = \min\{Z_{i-1}, X_i\}$ and $\delta_i = I[X_i < Z_{i-1}]$, for $i = 2, 3, \ldots$. Clearly the pair Z_i and δ_i are neither independent nor have the same identical distribution. Let λ_i denote the multiplicities of Z_i (corresponding to $\delta_i = 0$). Then, using an approach similar to the one used in Susarla and Van Ryzin (1976), Tiwari and Zalkikar obtain the Bayes estimator of the survival function S with respect to $\mathcal{D}(\alpha)$ and under the weighted squared error loss as

$$\widehat{S}_\alpha(u) = \frac{\alpha(u, \infty) + N^+(u)}{\alpha(R^+) + n} \times \prod_{j=1}^{n} \left(\frac{\alpha[Z_j, \infty) + N^+(Z_j) + \lambda_j}{\alpha[Z_j, \infty) + N^+(Z_j)} \right)^{I[\delta_j = 0, Z_j \leq u]/\lambda_j}. \tag{3.17}$$

By taking the limit $\alpha(R^+) \to 0$, it is shown that the above estimator reduces to the nonparametric maximum likelihood estimate of S obtained by Samaniego and Whitaker (1988). Weak convergence of the above estimator is also established. Finally, considering the usual empirical Bayes set up of Sect. 2.2.4, they derive an empirical Bayes estimator and show that it is asymptotically optimal. Details can be found in their paper.

3.2.6 Estimation Based on Random Left Truncation

In most of the applications, we encounter censoring on the right. However, in Tiwari and Zalkikar (1993) the authors consider left truncation and derive the Bayes estimator for the survival function. Under the random left truncation model, it is assumed that we have independent random variables $X_1, \ldots, X_n$ and $T_1, \ldots, T_n$ from continuous distribution functions F and G, respectively, and we observe the pairs (X_i, T_i), $i = 1, 2, \ldots, n$ only if $X_i \geq T_i$, for all i, otherwise nothing is observed. Since G is continuous, T_i's are distinct. Regardless of F being continuous, they assume $F \in D(\alpha)$, and obtain the Bayes estimator of the survival function S as follows:

$$\widehat{S}(u) = \frac{\alpha(u, \infty) + n(S_n(u) - \overline{G}_n(u))}{\alpha\big(R^+\big)} \times \prod_{i:T_i<u}^{n} \frac{\alpha(T_i, \infty) + n(S_n(T_i^-) - \overline{G}_n(T_i^-))}{\alpha(T_i, \infty) + n(S_n(T_i^-) - \overline{G}_n(T_i))}, \tag{3.18}$$

where $S_n(u) = (1/n)\sum_{j=1}^n [X_j > u]$, $\overline{G}_n(u) = (1/n)\sum_{j=1}^n [T_j > u]$, and $\overline{G}(u^-) = 1 - G(u^-)$, $G(u^-)$ being the left sided limit at u. As $\alpha(R^+) \to 0$, they show that the limiting Bayes estimator is a rescaled PL estimator above the smallest truncating observation $T_{(1)}$. Below $T_{(1)}$, the sample does not provide any information and hence the Bayes estimator reduces to the limit of the prior guess. They also discuss the weak convergence of the above estimator and give a numerical example.

3.2.7 Estimation Based on Proportional Hazard Models

Ghorai (1989) derives the Bayes estimator of the survival function assuming the proportional hazard model. Let S_X and S_Y denote the survival functions of uncensored variable X and censored variable Y, respectively. Then, under the proportional hazard model, it is assumed that $S_Y = S_X^{\beta}$ for some $\beta > 0$. β is known as the *censoring parameter*. Since $\mathcal{E}(\delta_i) = P(X \leq Y) = (1+\beta)^{-1} = \theta$, say, θ is the expected proportion of uncensored observations. Since X's and Y's are assumed to be independent and $Z = \min(X, Y)$, $S_Z(t) = P(Z > t) = (S_X(t))^{1+\beta}$ or $S_X(t) = (S_Z(t))^{\theta}$. He assumes a priori $S_Z(t) \in \mathcal{D}(\alpha)$ and $\theta \sim Be(a, b)$, a beta distribution with parameters a and b, and that $S_Z(t)$ and θ are independent. Then the posterior distributions are $S_Z(t)|(\mathbf{Z}, \boldsymbol{\delta}) \in \mathcal{D}(\alpha + \sum_{i=1}^n \delta_{Z_i})$ and $\theta|(\mathbf{Z}, \boldsymbol{\delta}) \sim Be(a + N_u, b + n - N_u) = Be(a^*, b^*)$, say, where $N_u = \sum_{i=1}^n \delta_i$. Since $S_Z(t)$ and θ are independent a priori, they can be seen to be so a posterior as well. Now the Bayes estimator of $S_X(t)$ under L_1 loss function is

$$\begin{aligned}\widehat{S}_X(t) &= E\big[\big(S_Z(t)\big)^{\theta}|(\mathbf{Z}, \boldsymbol{\delta})\big]\\ &= \mathcal{E}_{Be(a^*,b^*)}\big[\mathcal{E}_{\mathcal{D}(\alpha+\sum_{i=1}^n \delta_{Z_i})}\big(S_Z^{\theta}(t)|\theta\big)\big]. \qquad (3.19)\end{aligned}$$

Appealing to the moments of the Dirichlet process, the quantity inside the square brackets is

$$\begin{aligned}&\mathcal{E}_{\mathcal{D}(\alpha+\sum_{i=1}^n \delta_{Z_i})}\big(S_Z^{\theta}(t)|\theta\big)\\ &\quad= \frac{\Gamma(\alpha(R^+)+n)}{\Gamma(\alpha(R^+)+n+\theta)}\frac{\Gamma(\alpha(t,\infty)+N^+(Z_1)+\theta)}{\Gamma(\alpha(t,\infty)+N^+(Z_1))}\cdots, \qquad (3.20)\end{aligned}$$

where, as before, $N^+(u) = \sum_{i=1}^n I[Z_i > u]$. However, explicit evaluation of this expression is difficult and Ghorai provides some approximations by expanding the ratios of gamma functions. Final evaluation of the estimator $\widehat{S}_X$ is then proceeded by taking the expectation of this quantity with respect to $Be(a^*, b^*)$. He proves some asymptotic properties of this estimator, in particular, almost sure consistency and weak convergence to a Gaussian process.

3.2.8 Modal Estimation

There is difficulty in defining the mode of an infinite-dimensional distribution. Like Ramsey (1972), Ferguson and Phadia (1979) avoid this difficulty by restricting the attention to finite-dimensional subsets of the variables. With $t_1, \ldots, t_k$ as arbitrary points, they define the modal estimate of $F(t)$ as the modal value of $F(t)$ in the joint distribution of $(F(t), F(t_1), \ldots, F(t_k))$, where $t_1, \ldots, t_k$ are arbitrary points that contain all exclusive censoring points. So assume that $t_1, \ldots, t_k$ are arbitrary points containing t and all exclusive censoring points. Further assume that they are arranged in an increasing order and that $F \in \mathcal{D}(\alpha)$. α is assumed to give positive mass to every open interval. Thus the vector $(p_1, p_2, \ldots, p_{k+1}) = (F(t_1), F(t_2) - F(t_1), \ldots, 1 - F(t_k))$ has a Dirichlet distribution with parameters $\beta_i = \alpha(t_i) - \alpha(t_{i-1})$ for $i = 1, \ldots, k+1$ with $\alpha(t_0) = 0$, $\alpha(t_{k+1}) = \alpha(R)$, and $\beta_i > 0$, for $i = 1, \ldots, k+1$. They take the density of the vector $(p_1, p_2, \ldots, p_k)$ with respect to the measure $d\nu = \prod_{i=1}^{k} dp_i / \prod_{i=1}^{k+1} p_i$ where $p_{k+1} = 1 - \sum_{i=1}^{k} p_i$, over the simplex

$$S_k = \left\{ (p_1, p_2, \ldots, p_k) : p_i \geq 0 \text{ for } i = 1, \ldots, k, \text{ and } \sum_{i=1}^{k} p_i \leq 1 \right\}. \tag{3.21}$$

The prior density of $(p_1, p_2, \ldots, p_k)$ over S_k with respect to $d\nu$ is proportional to $\prod_{i=1}^{k+1} p_i^{\beta_i}$. With this formulation they establish the following. Let α be such that it gives positive mass to every open interval. Then the posterior modal (with respect to ν) estimate of S, given the data is

$$\widehat{S}(t) = \frac{\alpha(t,\infty) + h_{j(t)}}{\alpha(R) + n} \prod_{i=1}^{j(t)} \frac{(\alpha(u_i,\infty) + h_i + \lambda_i)}{(\alpha(u_i,\infty) + h_i)}, \tag{3.22}$$

where u_i's are distinct observations in the sample, $j(t)$ denotes the number of u_i less than or equal to t and h_j denote the number of observations greater than u_j.

The above formula reveals that the estimate depends only on the censoring points among $t_1, \ldots, t_k$, and is thus independent of the choice of $t_1, \ldots, t_k$ provided all exclusive censoring points are included.

Remark 3.1 It is clear from the above results in Sect. 3.2 that the Bayes estimator of S with respect to a Dirichlet process prior under various sampling schemes turn out to be a version of Susarla-Van Ryzin estimator; and when the prior information tends to nil via $\alpha(R) \to 0$, they reduce to the nonparametric MLE, namely, the PL estimator. This may be construed as another attractive feature of the Dirichlet process.

It is possible to derive the above results by replacing the Dirichlet process with other priors. Some such applications are presented in the next section.

3.3 Estimation of a SF Based on Other Priors

In this section, Bayesian estimators of a survival function with respect to other priors, such as, processes neutral to the right, beta, gamma, beta-Stacy, etc. are presented. They have similar form as for the case of the Dirichlet process prior, and in many cases the estimators are a version of the Susarla-Van Ryzin estimator. In case of the processes neutral to the right, a posterior moment generating function is given. From these, the estimators for the case of uncensored data can easily be recovered. Also, an alternate approach of placing a prior via subsurvival functions is presented.

3.3.1 Estimation Based on an Alternate Approach

A different approach is adopted by Tsai (1986). He considers the joint distribution of (Z, δ) and assigns a Dirichlet process prior with parameter α^* (to be described below) to the pair. He obtains the Bayes estimators of subsurvival functions under the weighted squared error loss function and then combines the two estimators to produce an estimator for the survival function, which is then shown to be Bayes under a slightly different loss function. This approach does not require independence between the X_i's and Y_i's. In fact he does not even define censoring variables Y_i's. Instead he assumes the data available to be of the form (Z_i, δ_i) where $\delta_i = 1$ if Z_i, $= X_i$ and $\delta_i = 0$ if $Z_i < X_i$ for $i = 1, \dots, n$, the pairs (Z_i, δ_i) are independent, and $X_1, \dots, X_n \overset{iid}{\sim} S$. Furthermore, it is clear that (Z_i, δ_i)'s need not be identically distributed. The independence-like assumption makes the marginal distribution of X identifiable, and the estimate of S consistent. Since the marginal distribution is not Dirichlet under this assumption, the resulting estimator is distinct from that of Susarla and Van Ryzin. Here are some details.

Tsai places a Dirichlet process prior with parameter α^*, on $(\mathcal{R}^*, \mathcal{B}^*)$, where $\mathcal{R}^* = R^+ \times \{0, 1\}$ and $\mathcal{B}^* = \mathcal{B} \times \{\phi, \{0\}, \{1\}, \{0, 1\}\}$, $\mathcal{B}$ is a Borel field on R^+ and α^* is a non-null finite measure on $(\mathcal{R}^*, \mathcal{B}^*)$. Then, based on a random sample of size n, the Bayes estimators $\widehat{S}_u$ and $\widehat{S}_c$ of subsurvival functions $S_u(t) = P(Z > t, \delta = 1)$ and $S_c(t) = P(Z > t, \delta = 0)$, respectively, are derived under the loss function $L(S., \widehat{S}.) = \int_0^\infty (S. - \widehat{S}.)^2 dW$:

$$\widehat{S}_u(t) = \frac{\alpha^*((t, \infty), \{1\}) + \sum_{i=1}^n I[Z_i > t, \delta_i = 1]}{\alpha^*(\mathcal{R}^*) + n} \tag{3.23}$$

and

$$\widehat{S}_c(t) = \frac{\alpha^*((t, \infty), \{0\}) + \sum_{i=1}^n I[Z_i > t, \delta_i = 0]}{\alpha^*(\mathcal{R}^*) + n}. \tag{3.24}$$

To unify discrete and continuous cases of S, he follows pp. 7–9 in Kalbfleisch and Prentice (1980) and defines

$$\Lambda(t) = -\int_0^{t^+} \frac{dS(u)}{S(u^-)} \tag{3.25}$$

and

$$\gamma(\Lambda)(t) = \lim_{k\to\infty} \prod_{i=1}^{k} \left\{1 - \left[\Lambda(u_i) - \Lambda(u_{i-1})\right]\right\}, \tag{3.26}$$

where $0 = u_0 < u_1 < \ldots < u_k = t$, the integral and differential operators are Riemann-Stieltjes operators, and the limit $k \to \infty$ is taken as $\Delta u_k = u_k - u_{k-1} \to 0$. From the above,

$$S(t) = \gamma(\Lambda)(t) = \exp\left\{\oint_0^t \frac{dS(u)}{S(u^-)}\right\} \prod_{s\le t}\left(1 - \frac{\Delta S(s)}{S(s^-)}\right), \tag{3.27}$$

where the integral is over the intervals of points less than t for which S is continuous, and $\Delta S(s) = S(s^-) - S(s^+)$.

Thus a self-consistent (Efron 1967) estimator $\widehat{S}$ of S is obtained as

$$\begin{aligned}\widehat{S}(t) &= \gamma\left(-\int_0^{t^+} \frac{d\widehat{S}_u(u)}{\widehat{S}_u(u^-) + \widehat{S}_c(u^-)}\right)(t) \\ &= \gamma\left(-\int_0^{t^+} \frac{d(\alpha^*((u,\infty),\{1\}) + \sum_{i=1}^n I[Z_i > u, \delta_i = 1])}{\alpha^*([u,\infty),\{0,1\}) + \sum_{i=1}^n I[Z_i \ge u]}\right)(t). \end{aligned} \tag{3.28}$$

Then it is shown that $\widehat{S}$ is the Bayes estimator of S under the loss function

$$L(S,\widehat{S}) = \int_0^{\infty} \left[\gamma^{-1}(S)(t) - \gamma^{-1}(\widehat{S})(t)\right]^2 dW(t), \tag{3.29}$$

where γ^{-1} denotes the inverse operator of γ. He proves the estimator to be strongly consistent and derives the weak convergence results.

When $\alpha^*((t,\infty),\{0\}) = 0$ and $\alpha^*((t,\infty),\{1\}) = \alpha(t,\infty)$, then $\widehat{S}$ reduces to a version of the Susarla-Van Ryzin estimator under certain conditions.

Salinas-Torres et al. (2002) generalize this approach in the context of k competing risks. Suppose the risk set is $\{1,\ldots,k\}$ and let Δ be a subset of $\{1,\ldots,k\}$. Then they derive a Bayes estimator for the marginal survival function S_Δ by the above approach and show that the resulting estimator is also consistent and weakly convergent. It is discussed later (Sect. 3.5.3) in this chapter.

3.3.2 Estimation Based on Neutral to the Right Processes

In all of the applications discussed in the previous section, the Dirichlet process prior was used. In this section, we describe the results when a neutral to the right process is used as prior in the estimation of a survival function. This prior being conjugate with respect to the right censored data, the posterior distribution given the data is also neutral to the right. This result of Doksum (1974) was extended to the case of two types of censoring in Ferguson and Phadia (1979).

Since the prior distribution of F may give positive probability to the event that F has a jump at a fixed point, it is useful in such problems to generalize earlier treatments to allow two types of censoring: “inclusive” and “exclusive”. In this case, the description of the posterior distribution of F turns out to be simpler than that in the case of uncensored data. In fact, the posterior distribution is the same as in Doksum’s except that the jump at the point x does not have to be treated differently. The increment at x is treated as if it were to the left of x in the case of exclusive censoring and to the right of x for inclusive censoring.

Thus the nonparametric Bayesian estimation problem based on right censored data can be conveniently carried out by using processes neutral to the right as prior processes. As a particular case, Susarla and Van Ryzin’s result is derived. A slight deviation from earlier treatment is considered here in that the censoring variables y_i’s are assumed to be fixed constants rather than random variables. However, as noted in Sect. 3.2.3, Lemma 7.1 in Lo (1993a, 1993b) has shown that the results hold even in the case when y_i’s are assumed to be random with distributions G_i, as long as F and G_i’s are independent.

Complete description of the posterior distribution of F for application to the censored data in which two types of censoring is considered, is presented for a sample of size one.

Theorem 3.2 (Ferguson and Phadia) *Let F be a random distribution function neutral to the right, X be a sample of size one from F, and let x be a real number.*

(a) *The posterior distribution of F given $X > x$ is neutral to the right; the posterior distribution of an increment y to the right of x is the same as the prior distribution; the posterior distribution of an increment to the left of or including x is found by multiplying the prior density by e^{-y} and normalizing.*
(b) *The posterior distribution of F given $X \geq x$ is neutral to the right; the posterior distribution of an increment y to the right of or including x is the same as the prior distribution; the posterior distribution of an increment to the left of x is found by multiplying the prior density by e^{-y} and normalizing.*

For the general case, as mentioned in Chap. 1, it is easy to work with the moment generating function (MGF), $M_t(\theta) = \mathcal{E}e^{-\theta Y_t}$, where Y_t is a process with nonnegative independent increments.

The vectors $\mathbf{u}$, $\boldsymbol{\delta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ as defined in Sect. 3.2.3 will be referred to as the *data*. We will use the same notation as before. We also use $M_t^-(\theta)$ to denote the MGF of Y_t^-, $M_t^-(\theta) = \lim_{s\to t} M_s(\theta)$, for $s < t$. $G_u(s)$ denotes the prior distribution of the jump in Y_t at u, and $H_u(s)$ its posterior distribution, given $X = u$ for a single observation. Then the following result was obtained.

Theorem 3.3 (Ferguson and Phadia) *Let F be a random distribution function neutral to the right, and let $X_1, \ldots, X_n$, be a sample of size n from F, yielding data* $\mathbf{u}$, $\boldsymbol{\delta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$. *Then the posterior distribution of F given the data is neutral to the right, and Y_t has posterior MGF*

$$M_t(\theta \mid data) = \frac{M_t(\theta + h_j(t))}{M_t(h_j(t))} \times \prod_{i=1}^{j(t)} \left[\frac{M_{u_i}^-(\theta + h_{i-1})}{M_{u_i}^-(h_{i-1})} \cdot \frac{C_{u_i}(\theta + h_i + \lambda_i, \delta_i)}{C_{u_i}(h_i + \lambda, \delta_i)} \cdot \frac{M_{u_i}(h_i)}{M_{u_i}(\theta + h_i)} \right], \tag{3.30}$$

where, if u is a prior fixed point of discontinuity of Y_t,

$$C_u(\alpha, \beta) = \int_0^\infty e^{-\alpha z}\left(1 - e^{-z}\right)^\beta dG_u(z) \tag{3.31}$$

while, if u is not a prior fixed point of discontinuity of Y_t,

$$C_u(\alpha, \beta) = \begin{cases} \int_0^\infty e^{-\alpha z}(1 - e^{-z})^\beta dH_u(z) & \text{if } \beta \geq 1 \\ 1 & \text{if } \beta = 0. \end{cases} \tag{3.32}$$

Now it is easy to evaluate posterior moments of F. For example the posterior expectation of the survival function S is obtained by plugging $\theta = 1$ in the above expression, $\mathcal{E}(S(t)|data) = M_t(1|data)$. However, the difficulty is encountered in finding the posterior distribution $H_u(s)$ of the jump at the point of discontinuity. Nevertheless, it is shown that in the case of homogeneous processes this is easy to do. This was illustrated in two specific cases (Ferguson and Phadia 1979).

3.3.3 Estimation Based on a Simple Homogeneous Process

In the first case, let Y_t be a simple homogeneous process with the MGF

$$M_t(\theta) = \exp\left\{\gamma(t) \int_0^\infty \left(e^{-\theta z} - 1\right)e^{-\tau z}\left(1 - e^{-z}\right)^{-1} dz\right\}, \tag{3.33}$$

where γ is assumed to be continuous. Then, with the above data scheme and notations, it is shown that

$$\begin{aligned}\mathcal{E}\big(S(t)|data\big) &= e^{-\gamma(t)/(h_{j(t)}+\tau)} \\ &\quad \times \prod_{i=1}^{j(t)}\left[\exp\left\{\gamma(u_i)\frac{h_{i-1}-h_i}{(h_{i-1}+\tau)(h_i+\tau)}\right\}\right]\cdot\frac{(h_i+\lambda_i+\tau)}{(h_i+\lambda_i+\delta_i+\tau)},\end{aligned} \tag{3.34}$$

which is the Bayes estimator under the weighted squared error loss function.

If we have the knowledge of $S_0(t)$, the prior guess of the survival function, then $\gamma(t) = -\tau\log S_0(t)$. Substituting this in the above formula, we get

$$\begin{aligned}\mathcal{E}\big(S(t)|data\big) &= S_0(t)^{\tau/(h_{j(t)}+\tau)} \\ &\quad \times \prod_{i=1}^{j(t)}\left[S_0(t)^{-\tau(h_{i-1}-h_i)/((h_{i-1}+\tau)(h_i+\tau))}\right]\cdot\frac{(h_i+\lambda_i+\tau)}{(h_i+\lambda_i+\delta_i+\tau)}.\end{aligned} \tag{3.35}$$

Further if $S_0(t) > 0$ for all t, we have as $\tau \to 0$,

$$\mathcal{E}\big(S(t)|data\big) \to \begin{cases} \prod_{i=1}^{j(t)} \frac{h_i+\lambda_i}{h_i+\lambda_i+\delta_i} & \text{for } t < u_k \\ \frac{S_0(t)}{S_0(u_k)}\prod_{i=1}^{k} \frac{h_i+\lambda_i}{h_i+\lambda_i+\delta_i} & \text{for } t \geq u_k, \end{cases} \tag{3.36}$$

where $(h_k+\lambda_k)/(h_k+\lambda_k+\delta_k)$ is to be treated as 1 if it is $0/0$. This is a maximum likelihood estimator. If there are no censored observations, $\lambda_i = 0$, $h_i+\delta_i = h_{i-1}$ and the estimator reduces to the sample distribution function.

3.3.4 Estimation Based on Gamma Process

A second case is when the independent increments of the process Y_t have gamma distributions with parameters $\gamma(t)$ (assumed to be continuous) and τ. In this case the MGF takes a simple form

$$\begin{aligned}M_t(\theta) &= \left(\frac{\tau}{\tau+\theta}\right)^{\gamma(t)} \\ &= \exp\left\{\gamma(t)\int_0^\infty \big(e^{-\theta z}-1\big)e^{-\tau z}z^{-1}dN(z)\right\}.\end{aligned} \tag{3.37}$$

The posterior mean of the survival function given the above scheme of data turns out to be

$$\begin{aligned}\mathcal{E}\big(S(t)|data\big) &= M_t(1|data)\\ &= \left(\frac{h_j(t)+\tau}{h_j(t)+\tau+1}\right)^{\gamma(t)} \prod_{i=1}^{j(t)}\left[\left(\frac{(h_{i-1}(t)+\tau)(h_i(t)+\tau+1)}{(h_{i-1}(t)+\tau+1)(h_i(t)+\tau)}\right)^{\gamma(u_i)}\right.\\ &\quad \left.\times \frac{\phi_G(h_i+\lambda_i+\tau+1,\delta_i)}{\phi_G(h_i+\lambda_i+\tau,\delta_i)}\right], \end{aligned} \tag{3.38}$$

where

$$\phi_G(\alpha,\beta) = \sum_{i=0}^{\beta-1}\binom{\beta-1}{i}(-1)^i \log\left(\frac{\alpha+i+1}{\alpha+i}\right). \tag{3.39}$$

Note that $\mathcal{E}(S(t)) = M_t(1) = (\frac{\tau}{\tau+1})^{\gamma(t)}$. Thus, if we have a prior guess at $S(t)$, say, $S_0(t)$, we can choose $\gamma(t)$ such that $(\tau/(\tau+1))^{\gamma(t)} = S_0(t)$, for all t and for a fixed τ. For further observations and the effect of τ on the behavior of this estimate, see Ferguson-Phadia paper.

In the case of gamma process prior, Ghorai (1981) derived an empirical Bayes estimator of the survival function. Here $\gamma(\cdot)$ plays a role similar to $\alpha(\cdot)$ in the Dirichlet process. He assumes τ to be known, estimates $\gamma(\cdot)$ from the previous n-stages, and proceeds to determine the empirical Bayes estimator of S at the $(n+1)$-th stage on the line of Susarla and Van Ryzin (1978a) and Phadia (1980). For simplicity he also considers the case of sample size one. His estimator of S at the $(n+1)$-th stage turns out to be the above estimator in which γ is replaced by its estimator $\widehat{\gamma}$ and can be simplified as

$$\widehat{S}_{n+1}(u) = \begin{cases} (\frac{\tau+1}{\tau+2})^{\widehat{\gamma}(u)} & \text{if } u < Z_{n+1}\\ (\frac{\tau}{\tau+1})^{\widehat{\gamma}(u)}(\frac{(\tau+1)^2}{\tau(\tau+2)})^{\widehat{\gamma}(Z_{n+1})} & \text{if } u \geq Z_{n+1}, \delta_{n+1}=0\\ \kappa(\frac{\tau}{\tau+1})^{\widehat{\gamma}(u)}(\frac{(\tau+1)^2}{\tau(\tau+2)})^{\widehat{\gamma}(Z_{n+1})} & \text{if } u \geq Z_{n+1}, \delta_{n+1}=1, \end{cases} \tag{3.40}$$

where $\kappa = \ln(\frac{\tau+2}{\tau+1})/\ln(\frac{\tau+1}{\tau})$ and $\widehat{\gamma}(u)$ is such that

$$\left(\frac{\tau}{\tau+1}\right)^{\widehat{\gamma}(u)} = \frac{1+N^+(u)}{1+n}\prod_{j=1}^{n}\left(\frac{N^+(Z_j)+2}{N^+(Z_j)+1}\right)^{[\delta_j=0,Z_j\leq u]}_{.}. \tag{3.41}$$

Ghorai showed that the sequence of empirical Bayes estimators is asymptotically optimal with rate of convergence $O(n^{-1})$.

3.3.5 *Estimation Based on Beta Process*

As noted in Sect. 1.8, the cumulative hazard function $H(t) = \int_{[0,t]} dF(s)/F[s,\infty)$ is related to the distribution function via the correspondence $F(t) = 1 - \prod_{[0,t]}\{1 -$

$dH(s)\}$, where $\prod$ is the product integral. Suppose H has a beta process prior with parameters $c(\cdot)$ and H_0 as defined in Sect. 1.8. Suppose we have, as before $X_1, \ldots, X_n \overset{iid}{\sim} F, y_1, \ldots, y_n$ as censoring times and we observe $Z_i = \min\{X_i, y_i\}$ and $\delta_i = I[X_i \leq y_i]$. Then, using the posterior distribution of H given the data $(\mathbf{Z}, \boldsymbol{\delta})$ and applying the relevant formulas, Hjort (1990) derives the following Bayes estimator of $S(t)$ under L_1 loss.

$$\widehat{S}(t) = \mathcal{E}\big(S(t)|data\big) = \prod_{[0,t]} \left\{1 - \frac{c(s)dH_0(s) + dN(s)}{c(s) + M(s)}\right\}, \tag{3.42}$$

where $N(\cdot)$ is the counting process for uncensored observations, $dN(t) = N\{t\}$, the number of uncensored observations at t, and $M(t) = \sum_i^n I[Z_i \geq t]$, the number of observations surviving at t. As $c(\cdot) \to 0$, $\widehat{S}(t)$ tends to the PL estimator. It should be noted that here the prior is placed on the cumulative hazard function and not on the survival function itself. On the other hand, if F has a beta-Stacy prior with parameters $c(\cdot)$ and G, Walker and Muliere (1997a) obtain the same estimator as shown below.

3.3.6 Estimation Based on Beta-Stacy Process

Assume a beta-Stacy prior (see Sect. 1.9) with parameters $c(\cdot)$ and G for F. Then given a random sample from F, with possible right censored observations, the Bayes estimate of $S(t)$ with L_1 loss function, is given by (Walker and Muliere 1997a)

$$\widehat{S}(t) = \mathcal{E}\big(S(t)|data\big) = \prod_{[0,t]} \left\{1 - \frac{c(s)dG(s) + dN(s)}{c(s)G[s, \infty) + M(s)}\right\}, \tag{3.43}$$

where $N(\cdot)$ and $M(t)$ are as defined in the previous section. The PL estimator is obtained if $c(\cdot) \to 0$.

3.3.7 Estimation Based on Polya Tree Priors

Muliere and Walker (1997) present the estimation of a survival function using the posterior predictive distribution of a future observation. Assume that Π and $\mathcal{A}$, as described in Sect. 1.11 are given and $F \in \mathcal{PT}(\Pi, \mathcal{A})$. Let $\theta_1, \theta_2, \ldots | F \overset{iid}{\sim} F$. The posterior predictive distribution based on exact observations was given earlier as

$$\mathcal{P}[\theta_{n+1} \in B_{\underline{\epsilon}_m} | data] = \frac{\alpha_{\epsilon_1} + n_{\epsilon_1}}{\alpha_0 + \alpha_1 + n} \cdots \frac{\alpha_{\underline{\epsilon}_m} + n_{\underline{\epsilon}_m}}{\alpha_{\underline{\epsilon}_{m-1}0} + \alpha_{\underline{\epsilon}_{m-1}1} + n_{\underline{\epsilon}_{m-1}}}, \tag{3.44}$$

where n_ϵ is the number of observations in B_ϵ and $\underline{\epsilon}_k = \epsilon_1 \ldots \epsilon_k$. Note that if we let $\alpha_{\epsilon 0} + \alpha_{\epsilon 1} = \alpha_\epsilon$ for all ϵ, it reduces to $(\alpha(B_{\underline{\epsilon}_m}) + n_{\underline{\epsilon}_m})/(\alpha(R^+) + n)$ as one would obtain for the Dirichlet process. For the censored data, they show that the posterior predictive distribution of a future observation is given by

$$\mathcal{P}[\theta_{n+1} \in B_{\underline{\epsilon}_m} | data] = \frac{\alpha_{\epsilon_1} + n_{\epsilon_1}}{\alpha_0 + \alpha_1 + n} \cdots \frac{\alpha_{\underline{\epsilon}_m} + n_{\underline{\epsilon}_m}}{\alpha_{\underline{\epsilon}_{m-1}0} + \alpha_{\underline{\epsilon}_{m-1}1} + n_{\underline{\epsilon}_{m-1}} - \lambda_{\underline{\epsilon}_{m-1}}}, \quad (3.45)$$

where λ_ϵ is the number of observations that are censored in B_ϵ. If we take α measure such that $\alpha_\epsilon = \alpha(B_\epsilon)$ for all ϵ, then this estimator is essentially the same as Susarla-Van Ryzin estimator obtained using the Dirichlet process, since for such α_ϵ's, the Polya tree distribution reduces to the Dirichlet process as was noted in Ferguson (1974).

For the Polya tree Π, their construction uses the distinct censoring points, say, $t_1 < t_2 < \ldots < t_k$, and splits the right hand side interval $[t_i, \infty)$ into two intervals $[t_i, t_{i+1})$ and $[t_{i+1}, \infty)$, for $i = 1, 2, \ldots, k$. The construction of intervals to the left of these partitions remain arbitrary. Obviously the point of contention is the unsavory fact of using data in constructing the prior. They suggest some partial remedy to overcome this undesirable situation. Neath (2003) also uses Polya tree distributions for statistical modeling of censored data.

3.3.8 Estimation Based on an Extended Gamma Prior

In contrast to placing a prior on the space of survival functions, if the hazard rate $r(t)$ is assumed to be distributed a priori as an extended gamma process with parameters α and β, $\Gamma(\alpha(\cdot), \beta(\cdot))$ (see Sect. 1.7), and $S(x) = \exp\{-\int_{[0,x)} r(t)dt\}$, then Dykstra and Laud (1981) have shown that $\widehat{S}_{\alpha\beta}(t)$ given below is the Bayes estimator of $S(t) = P(X_{n+1} \geq t | X_1 = x_1, \ldots, X_n = x_n)$, the conditional survival function of a future observation given n current observations, under the usual L_1 loss function.

$$\widehat{S}_{\alpha\beta}(t) = \exp\left\{-\int_{[0,\infty)} \log\bigl(1 + \beta^*(s)(t-s)^+\bigr)d\alpha(s)\right\} \cdot \phi\bigl(\beta^{**}\bigr)/\phi\bigl(\beta^*\bigr), \quad (3.46)$$

where

$$\phi(\beta) = \int_{[0,x_n)} \cdots \int_{[0,x_1)} \prod_{i=1}^{n} \beta(z_i) \prod_{i=1}^{n} d\left[\alpha + \sum_{j=i+1}^{n} I_{(x_j,\infty)}\right](z_i), \quad (3.47)$$

with $\beta^{**}(s) = \beta^*(s)/[1 + \beta^*(s)(t-s)^+]$ and $\beta^*(t) = \beta(t)/[1 + \beta(t) \cdot \sum_{i=1}^{m}(x_i - t)^+]$.

3.3.9 Estimation Assuming Increasing Failure Rate

If the hazard rate (also known as failure rate) $r(t)$ is known to be increasing (nondecreasing), a different approach is proposed in Padgett and Wei (1981). In this case they propose a constant jump process prior on the space of all increasing failure rates. The process consists of constant jumps of size c at times $T_i > 0$, $i = 1, 2, \ldots,$ where T_i are arrival times of a Poisson process $\{N(t) : t > 0\}$ with intensity parameter ν. With respect to this prior and under the usual respective loss functions, the Bayesian estimates of the survival function, $S(x) = \exp\{-\int_{[0,x)} r(t)dt\}$, failure rate function, and the density function, based on right censored observations, are obtained by them. Although the derivation is straight forward, the expressions do not turn out to be simple. See their paper for details.

3.4 Linear Bayes Estimation of a SF

In Bayesian estimation of a survival curve S, Zehnwirth (1985) takes a different approach. In assuming the Dirichlet process or processes neutral to the right as priors, it is tacitly assumed that the hazard contributions of non-overlapping intervals are independent. Zehnwirth argues that this may not be the case in practice. In his paper he obtains a Bayesian estimator of S by estimating the hazard contributions between successive censoring points by linear Bayes rule. In doing so, he gains tractability and simplicity but his estimator turns out to be only an approximate Bayes estimator of the survival curve. He takes the loss function as point-wise squared error at distinct censoring points which is different from the usual weighted squared error loss function. It may be reasonable if the censoring points are assumed to be fixed, as is the case in some clinical studies where the trials are monitored at regular intervals.

Again as in the right censored data model, let $Z_{(1)}, Z_{(2)}, \ldots, Z_{(m)}$ be the distinct ordered censored observations among a sample of n observations. Let $N(u) = \sum_{j=1}^{n}[Z_j \geq u]$, $N^+(u) = \sum_{j=1}^{n}[Z_j > u]$ and λ_j stand for the number of censored observations at $Z_{(j)}$, $\sum_{j=1}^{m} \lambda_j = n$. Further denote $N(j) = N(Z_{(j)})$ for $j = 1, 2, \ldots, m$. Let $p(b|a) = P[X_i \geq b \mid X_i \geq a] = S(b^-)/S(a^-)$ if $P[X_i \geq a] > 0$, otherwise $p(b|a) = 0$. $p(b|a)$ represents the conditional probability of surviving up to time b given that the object has survived up to time a.

For any $u \in R^+$, let $Z_{(l)} \leq u \leq Z_{(l+1)}$, $l = 0, 1, \ldots, m$ with $Z_{(0)} = 0$, $Z_{(m+1)} = \infty$, $N(0) = n$ and $\lambda_0 = 0$. Consider the partition $[0, Z_{(1)}), [Z_{(2)}, Z_{(3)}), \ldots, [Z_{(l)}, u)$ of $[0, u)$. Then $S(u)$ can be written as

$$S(u) = p(u|Z_{(l)}) \prod_{j=0}^{l-1} p(Z_{(j+1)}|Z_{(j)}). \tag{3.48}$$

Zehnwirth now estimates each $p(b|a)$ by a linear Bayes rule (i.e. linear rule that best approximates the Bayes rule) by minimizing the risk

$$\begin{aligned} R(S, \widehat{S}) = \mathcal{E}\left[\left(a_{u,l} + b_{u,l}\frac{N^+(u)}{N(l) - \lambda_l} - p(u|Z_{(l)})\right)^2\right] \\ + \sum_{j=0}^{l-1} \mathcal{E}\left[\left(a_j + b_j\frac{N(j+1)}{N(j) - \lambda_j} - p(Z_{(j+1)}|Z_{(j)})\right)^2\right], \end{aligned} \tag{3.49}$$

over all $a_{u,l}$, $b_{u,l}$, a_j and b_j for $j = 0, 1, \ldots, l-1$. The linear Bayes estimator for $S(u)$ thus obtained for $Z_{(l)} \leq u \leq Z_{(l+1)}$, $l = 0, 1, \ldots, m$, is

$$\widehat{S}(u) = \left\{ \frac{N^+(u) + f(u|Z_{(l)}) g(u|Z_{(l)},)}{N(l) - \lambda_l + f(u|Z_{(l)})} \right\} \times \prod_{j=0}^{l-1} \left(\frac{N(j+1) + f(Z_{(j+1)}|Z_{(j)}) g(Z_{(j+1)}|Z_{(j)})}{N(j) - \lambda_j + f(Z_{(j+1)}|Z_{(j)})} \right), \tag{3.50}$$

where

$$f(b|a) = \frac{\mathcal{E}[p(b|a)(1 - p(b|a)]}{\mathrm{Var}[p(b|a)]} \quad \text{and} \quad g(b|a) = \mathcal{E}\big[p(b|a)\big]. \tag{3.51}$$

Here $f(b|a)$ may be interpreted as the number of individuals at risk at a and $f(b|a)g(b|a)$ as the number of survivors up to b among them.

So far no assumption is made regarding a prior for F. If F is assumed to be a neutral to the right process, then given data the posterior distribution of F is also neutral to the right. In this case $f(b|a)$ and $g(b|a)$ may be evaluated as follows. Note that for any two disjoint intervals $[0, a)$ and $[a, b)$, the survival probability satisfy $p(b|0) = p(a|0)p(b|a)$. By the independence property of neutral to the right processes, this expression can be written as $\mathcal{E}[S^r(b^-)/S^r(a^-)] = \mathcal{E}[S^r(b^-)]/\mathcal{E}[S^r(a^-)]$ for any r, a positive integer. This yields

$$g(b|a) = S_1\big(b^-\big)/S_1\big(a^-\big) \quad \text{and}$$
$$f(b|a) = \frac{S_1(b^-)/S_1(a^-) - S_2(b^-)/S_2(a^-)}{S_2(b^-)/S_2(a^-) - S_1^2(b^-)/S_1^2(a^-)}, \tag{3.52}$$

where S_1 and S_2 are the first and second moments of S, $S_1(u) = \mathcal{E}(S(u))$ and $S_2(u) = \mathcal{E}(S^2(u))$.

Now substituting these quantities, the linear Bayes estimator for $S(u)$ is given by

$$\widehat{S}(u) = \left\{ \frac{N^+(u) + f(u|Z_{(l)}) S_1(u^-)/S_1(z^-_{(l)})}{N(l) - \lambda_l + f(u|Z_{(l)})} \right\} \times \prod_{j=0}^{l-1} \left(\frac{N(j+1) + f(z_{(j+1)}|z_{(j)}) S_1(z^-_{(j+1)})/S_1(z^-_{(j)})}{N(j) - \lambda_j + f(z_{(j+1)}|z_{(j)})} \right). \tag{3.53}$$

On the other hand, if $F \sim \mathcal{D}(\alpha)$, then $S(u^-)$ has a $Be(\alpha[u, \infty), \alpha[0, u))$ distribution and therefore,

$$S_1\big(u^-\big) = \frac{\alpha[u, \infty)}{\alpha(R^+)} \quad \text{and} \quad S_2\big(u^-\big) = \frac{\alpha[u, \infty)(\alpha[u, \infty) + 1)}{\alpha(R^+)(\alpha(R^+) + 1)}. \tag{3.54}$$

This implies $f(b|a) = \alpha[a, \infty)$. Substituting this in the above expression yields the Bayes estimator of Susarla and Van Ryzin (1976). Other neutral to the right processes such as gamma or simple homogeneous processes may be used to evaluate

S_1 and S_2 yielding different linear estimates of S. In fact, besides the above independence assumption, all we need is the first two moments of S to compute this estimator.

3.5 Other Estimation Problems

In this section, we describe Bayesian solutions to some other estimation problems that have appeared in the literature.

3.5.1 Estimation of $P(Z > X + Y)$

Let X, Y and Z be independent and identically distributed as F, which is defined on R^+. Consider the problem of estimation of probability $\Delta(F)$

$$\Delta(F) = P(Z > X + Y) = \int_0^\infty \int_0^\infty S(x+y)dS(x)dS(y). \tag{3.55}$$

Assume $F \in \mathcal{D}(\alpha)$ and the squared error loss function, L_2. Based on a random sample of right censored data $(\mathbf{Z}, \boldsymbol{\delta})$ of size n, Zalkikar et al. (1986) derived the Bayes estimator of Δ as

$$\widehat{\Delta}(S) = \frac{(M+n)^2}{(M+n)^{(3)}}\left[-\int_0^\infty \widehat{S}_\alpha(2y)d\widehat{S}_\alpha(y) + (M+n)\Delta(\widehat{S}_\alpha)\right], \tag{3.56}$$

where $a^{(k)} = a(a+1)\ldots(a+k-1)$ and $\widehat{S}_\alpha$ is the Bayes estimator of S with respect to the Dirichlet process prior. When $M \to 0$, $\widehat{S}_\alpha \to \widehat{S}_{PL}$, the PL estimator of the survival function, therefore the estimator reduces to

$$\widehat{\Delta}(S) = \frac{n^2}{n^{(3)}}\left[-\int_0^\infty \widehat{S}_{PL}(2y)d\widehat{S}_{PL}(y) + n\Delta(\widehat{S}_{PL})\right]. \tag{3.57}$$

The empirical Bayes estimator is also derived by them using the procedure discussed in Sect. 2.2.4.

3.5.2 Estimation of $P(X \leq Y)$

The Bayesian estimator of $\Delta = P(X \leq Y) = \int FdG$ based on two independent samples, $X_1, \ldots, X_n$ from F and $Y_1, \ldots, Y_n$ from G (need not be of the same size) under the squared error loss was derived by Ferguson (1973). He assumed $F \in \mathcal{D}(\alpha_1)$ and independently, $G \in \mathcal{D}(\alpha_2)$. Based on samples $\mathbf{X}$ and $\mathbf{Y}$, he obtained the estimator as $\widehat{\Delta} = \int \widehat{F}_{\alpha_1} d\widehat{G}_{\alpha_2}$, where $\widehat{F}_{\alpha_1}$ and $\widehat{G}_{\alpha_2}$ are Bayes estimators of F and G, respectively. Its treatment from the empirical Bayes point of view was considered by Hollander and Korwar (1976), as indicated earlier.

It's extension to the case of right censored data was carried out in Phadia and Susarla (1979) as follows. Assume that we have $U_1, \ldots, U_n \overset{iid}{\sim} H_1$, and $V_1, \ldots, V_n \overset{iid}{\sim} H_2$, the censoring variables. All random variables are assumed to be mutually independent. We observe the pairs (S_i, T_i), $i = 1, 2, \ldots, n$, where $S_i = \min(X_i, U_i)$ and $T_i = \min(Y_i, V_i)$, and pairs (δ_i, η_i), where $\delta_i = I[X_i \leq U_i]$, and $\eta_i = I[Y_i \leq V_i]$ $i = 1, 2, \ldots, n$. Based on the data $\{\mathbf{S}, \boldsymbol{\delta}, \mathbf{T}, \boldsymbol{\eta}\}$ we want to estimate Δ. In the context of censored data, it is easy to handle if F and G are considered as *right sided* distribution functions. That is $F(t) = P(X > t)$ and $G(t) = P(Y > t)$. Then $\Delta = P(X \leq Y) = \int (1 - F) d(1 - G) = -\int G dF$. Therefore the Bayes estimate of Δ under the squared error loss is given by $\widehat{\Delta} = \int \mathcal{E}(G) d\mathcal{E}(F) = -\int \widehat{G}_{\alpha_2} d\widehat{F}_{\alpha_1}$ where the conditional expectation is taken with respect to the posterior distributions, and $\widehat{F}_{\alpha_1}$ and $\widehat{G}_{\alpha_2}$ are the Bayes estimators of F and G derived earlier.

Again when $\alpha(\cdot)$ and $\alpha(R)$ are unknown, the empirical Bayes methods can be used. This was done by the authors. For simplicity, they took the sample of size one at each stage and proposed the following estimator at the $(n + 1)$-th stage (based on n previous stages, each of sample size one): $\widehat{\Delta}_{n+1} = -\int_{-\infty}^{M_n} \widehat{G}_{\widehat{\alpha_2}} d\widehat{F}_{\widehat{\alpha_1}}$ where $\widehat{\alpha_1}$ and $\widehat{\alpha_2}$ are given by

$$\widehat{\alpha_1}(u) = \frac{N_1^+(u)}{n} \prod_{i=1}^{n} \left[\frac{N_1^+(S_i) + 2}{N_1^+(S_i) + 1} \right]^{[\delta_i = 0, S_i \leq u]} \quad \text{and}$$
$$\widehat{\alpha_2}(u) = \frac{N_2^+(u)}{n} \prod_{i=1}^{n} \left[\frac{N_2^+(T_i) + 2}{N_2^+(T_i) + 1} \right]^{[\eta_i = 0, T_i \leq u]}, \tag{3.58}$$

where $N_1^+(u) = \sum_{i=1}^{n} I[S_i > u]$, $N_2^+(u) = \sum_{i=1}^{n} I[T_i > u]$, and $\{M_n\}$ is a suitable sequence converging to ∞ as $n \to \infty$. Unlike in the uncensored case, the integral in the estimator here had to be restricted to the interval $(-\infty, M_n)$ to overcome divergence of some integrals arising in the bounds. They also discuss the choice of the sequence $\{M_n\}$. Note also that $\widehat{\alpha_1}$ and $\widehat{\alpha_2}$ do not depend on H_1 and H_2. Finally, the asymptotic optimality of the estimator is established and explicit expression for the rate of convergence of $O(n^{-1})$ is derived.

3.5.3 Estimation in Competing Risk Models

Consider the situation in which there are two competing causes of death labelled 1 and 2. With each cause of death i, $i = 1, 2$, associate a random variable T_i representing the time of death if i were only the cause of death. Then in practice, one observes only the $\min(T_1, T_2)$ and the cause of death 1 or 2. Based on this type of data, one needs to estimate the survival function $S(x, y) = P(X > x, Y > y)$ corresponding to a random probability P defined on $(R_+^2, \mathcal{B}_+^2)$, and $\mathcal{B}_+^2$ is the Borel field defined on R_+^2. Note that unlike in the right censored data case discussed in the previous sections, X and Y are not assumed to be independent.

Phadia and Susarla (1983) considered this problem and obtained the Bayes estimator of $S(x, y)$ with respect to a bivariate Dirichlet process prior and under the weighted squared error loss function.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from the unknown distribution function $F(x, y)$ defined on $R_+^2 = \{(0, \infty) \times (0, \infty)\}$. The data consists of (Z_i, δ_i), where $Z_i = \min(X_i, Y_i)$ and $\delta_i = I[X_i \le Y_i]$, $i = 1, 2, \ldots, n$. Let P be a Dirichlet process $\mathcal{D}(\alpha)$ defined on R_+^2 with parameter α, where α is a nonnegative finite measure on $(R_+^2, \mathcal{B}_+^2)$. The loss function used is $\int_{R_+^2} (S - \widehat{S})^2 dW$, where W is a known weight function on R_+^2. Suppose among the data there are k distinct z_i's and without loss of generality assume them to be ordered so that $0 < z_1 < \ldots < z_k < \infty$. Further let λ_i and μ_i be the number of censored and uncensored observations at z_i i.e. $\lambda_i = \#\{j \mid \min(X_j, Y_j) = z_i \text{ and } X_j > Y_j\}$, $\mu_i = \#\{j \mid \min(X_j, Y_j) = z_i \text{ and } X_j \le Y_j\}$. Then the Bayes estimator of $S(s, t)$ is obtained as

$$\begin{aligned}\widehat{S}(s,t) &= \mathcal{E}\big[S(s,t) \mid (Z, \delta)\big] \\ &= \frac{1}{\alpha(R_+^2) + n}\left\{\alpha\big((s,\infty)\times(t,\infty)\big) + N^+\big(\max(s,t)\big) + \sum_r \theta_r\right\}, \quad (3.59)\end{aligned}$$

where the summation is taken over all r such that $\min(s, t) < z_r < \max(s, t)$,

$$N^+(u) = \sum_{\{i : z_i > u\}} (\lambda_i + \mu_i) = \#\text{ of observations} > u, \quad (3.60)$$

$$\theta_r = \begin{cases} \lambda_r \alpha_s'(Z_r) & \text{if } s > t \\ \mu_r \alpha_t'(Z_r) & \text{if } s < t \\ 0 & \text{if } s = t, \end{cases} \quad (3.61)$$

and

$$\begin{aligned}\alpha_s'(Z_r) &= \lim_{\epsilon \searrow 0} \frac{\alpha(\{X > s, Z_r - \epsilon < Y \le Z_r\})}{\alpha(\{X > Y, Z_r - \epsilon < Y \le Z_r\})} \\ \alpha_t'(Z_r) &= \lim_{\epsilon \searrow 0} \frac{\alpha(\{Y > t, Z_r - \epsilon < X \le Z_r\})}{\alpha(\{X \le Y, Z_r - \epsilon < X \le Z_r\})},\end{aligned} \quad (3.62)$$

whenever the limits exist.

It should be noted that the Bayes estimator is a proper survival function. $\sum_r \theta_r$ in the numerator of the Bayes formula represents a quantity which may be considered as a sum of 'conditional' probabilities each weighed by the number of ties at the point of conditioning. If we take $s = t$, then it reduces to a 2-dimensional analogue of the Bayes estimator obtained by Ferguson (1973). If we set $t = 0$, we get the Bayes estimator of the marginal $S(s, 0)$.

This result can be extended to the case of competing risks models where there are three or more competing (dependent) causes of failure and we observe only the life time of the component and the cause of failure.

As an example, the authors compute the Bayes estimator by taking α measure to be continuous with density,

$$\alpha'(x,y) = \begin{cases} \beta(\beta+\gamma)e^{(\beta+\gamma)y} & \text{if } 0 \le x < y \\ \gamma(\beta+\gamma)e^{(\beta+\gamma)x} & \text{if } 0 \le y < x, \end{cases} \tag{3.63}$$

which is a special case of Freund's bivariate exponential distribution (Johnson and Kotz 1970), for $\beta, \gamma > 0$. Then by straight forward substitution and simplification, the Bayes estimator for $s > t$ is obtained as

$$\widehat{S}(s,t) = \frac{1}{n+1}\left\{e^{(\beta+\gamma)s}\left[1+\gamma(s-t)\right] + N^{+}(s) + \sum_{t<z_r<s} \lambda_r e^{(\beta+\gamma)(z_r - s)}\right\}. \tag{3.64}$$

A similar expression can be obtained for $s < t$. For $t = s$, $\widehat{S}(s,\infty) = (n+1)^{-1}\{\alpha((s,\infty)\times(s,\infty)) + N^{+}(s)\}$.

In the above treatment the posterior distribution of the joint survival function was not derived. Neath and Samaniego (1996) shaped this problem in the general framework of a multiple decrement model, and in the bivariate case obtained the posterior distribution of S given the data. However, their approach is different and employs the special feature of the identified minima in updating the parameter α. They introduce a new random variable ξ having a distribution

$$\mathcal{P}\{\xi \in B\} = \mathcal{P}\{U \in B | Z\}, \tag{3.65}$$

for any set $B \in \mathcal{B}_{+}^{2}$, and where U represents a complete observation (X,Y), and Z, the identified minimum of the pair. That is, ξ represents the unobserved complete realization from P. They prove that if $P \in \mathcal{D}(\alpha)$, then the posterior distribution of P given Z is a mixture of Dirichlet processes with random parameter measure $\alpha + \delta_{\xi}$. They extend this result to the sample of size n, for the case when the sample is drawn from a continuous distribution and α is a continuous measure. Finally the limiting posterior distribution is derived showing that as the sample size grows to infinity, the posterior distribution becomes degenerate. In a subsequent paper (Neath and Samaniego 1997) they handle the case of sample being drawn from a discrete distribution and α being a discrete measure.

In the above papers, the authors do not assume X and Y to be independent. In contrast, if the components are assumed to be independent, Salinas-Torres et al. (2002) propose a different approach in estimating the survival function based on a subset of failure causes by using Peterson's (1977) formula of expressing the survival function as a function of subsurvival functions. It is essentially the extension of Tsai's (1986) approach, where he considers two competing risks and uses Peterson's formula in deriving a self-consistent estimator and shows that it is Bayes with respect to a certain loss function (see Sect. 3.3.1). Salinas-Torres et al. approach is presented here briefly.

Consider the competing risks model with k competing causes of system's failure. Let X_j denote the failure time of the j-th component and its (marginal) survival function be denoted by $S_j(t) = P(X_j > t)$, $j = 1, \ldots, k$. Let $Z = \min(X_1, \ldots, X_k)$ and $S_j^*(t) = P(Z > t, \delta = j)$, be the subsurvival function of the j-th component, $j = 1, \ldots, k$. Then if $\delta = j$, $Z = X_j$ and $S(t) = P(Z > t) = \sum_{j=1}^{k} S_j^*(t)$. Let Δ be a non-empty subset of $\{1, 2, \ldots, k\}$ and denote its complement by Δ^c. Corresponding to Δ, let $S_\Delta^*(t) = P(Z > t, \delta \in \Delta)$ and $S_\Delta(t) = P(\min_{j\in\Delta} X_j > t)$ be the subsurvival and survival functions, respectively. Peterson's formula for $S_\Delta(t)$, as a function of subsurvival functions is

$$S_\Delta(t) = \varphi\big(S_\Delta^*(\cdot), S_{\Delta^c}^*(\cdot); t\big), \quad \text{for } t \le \min(t_{S_\Delta}, t_{S_{\Delta^c}}), \tag{3.66}$$

where

$$\varphi\big(F(\cdot), G(\cdot); t\big) = \exp\left\{ \oint_0^t \frac{dF(s)}{F(s) + G(s)} \right\} \prod_t \frac{F(s_+) + G(s_+)}{F(s_-) + G(s_-)}, \tag{3.67}$$

$t_{S_\Delta} = \sup\{t : S_\Delta(t) > 0\}$, and $\oint_0^t$ is the integral over the union of intervals of points less than t for which $F(\cdot)$ is continuous. $\prod_t$ indicates the product over the set $\{s \le t : s \text{ is a jump point of } F\}$. Let the loss function be

$$L\big(\mathbf{S}^*, \widehat{\mathbf{S}}^*\big) = \int_0^\infty \|\mathbf{S}^* - \widehat{\mathbf{S}}^*\|^2 dW(t), \tag{3.68}$$

where $\|\cdot\|$ stands for the usual norm, $\mathbf{S}^* = (S_1^*, \ldots, S_k^*)$, and $\widehat{\mathbf{S}}^* = (\widehat{S}_1^*, \ldots, \widehat{S}_k^*)$ is an estimator of $\mathbf{S}^*$. Suppose we have a sample of size n and let $Z_1, \ldots, Z_n$ be the minima of observations. Further assume that among them, $Z_{(1)} < \ldots < Z_{(m)}$ are m distinct ordered minima. Subsurvival functions $S_\Delta^*(t)$ are estimated by its natural estimator, $\widehat{S}_\Delta^*(t) = (1/n)\sum_{i=1}^{n} I[Z_i > t, \delta \in \Delta]$ and $\widehat{S}(t) = \widehat{S}_\Delta^*(t) + \widehat{S}_{\Delta^c}^*(t)$. Combining the various estimates, Salinas-Torres et al. establish the following result. Suppose the vector function $(\alpha_1(s, \infty), \ldots, \alpha_k(s, \infty))$ in s is continuous on $(0, t)$, $t > 0$ and $S_\Delta(t)$ and $S_{\Delta^c}(t)$ have no common points of discontinuities, then, for $t \le Z_{(m)}$,

$$\begin{aligned} \widehat{S}_\Delta(t) &= \varphi\big(\widehat{S}_\Delta^*, \widehat{S}_{\Delta^c}^*; t\big) \\ &= \widehat{S}(t)\pi_k(t) \exp\left\{ \frac{-1}{\alpha(R) + n} \right\} \sum_{j\in\Delta^c} \int_0^t \frac{d\alpha_j(s, \infty)}{\widehat{S}(s)} \end{aligned} \tag{3.69}$$

is the Bayes estimator of $S_\Delta(t)$ under the above loss function, where $n_j = \sum_{i=1}^{n} I[Z_i \ge Z_{(j)}]$, $d_j = \sum_{i=1}^{n} I[Z_i = Z_{(j)}, \delta_i = 1]$, $j = 1, \ldots, m$ and

$$\pi_k(t) = \prod_{i:Z_{(i)}\le t} \frac{\sum_{j=1}^{k} \alpha_j(Z_{(i)}, \infty) + n_i - d_i}{\sum_{j=1}^{k} \alpha_j(Z_{(i)}, \infty) + n_i}. \tag{3.70}$$

They also establish strong consistency and weak convergence of the estimator. When $k = 2$, $\Delta = \{1\}$ and $\Delta^c = \{2\}$, it reduces to the usual right censored data model. In this case, $\alpha_1(t, \infty) + \alpha_2(t, \infty) = \alpha(t, \infty)$ for each t, and the product $\widehat{S}(t)\pi_k(t)$ is analogous to Tsai's (1986) Eq. (3.2). It is also similar to the Susarla-Van Ryzin estimator derived in Sect. 3.1. Likewise, if $\alpha_j(s, \infty) \to 0$, for $j = 1, 2$, the estimator reduces to the PL estimator.

3.5.4 Estimation of Cumulative Hazard Rates

So far we have been dealing with the estimation of a survival function. In this section we describe the nonparametric Bayesian estimation of cumulative hazard rates obtained in Hjort (1990) by assuming a beta process prior of Sect. 1.8. Let $X \sim F$ taking values in the discrete time scale $\{0, b, 2b, \ldots\}$ (without loss of generality we take $b = 1$). For $j = 0, 1, 2 \ldots$, let $f(j) = P\{X = j\}$, $F(j) = P\{X \leq j\} = \sum_{i=0}^{j} f(i)$, $h(j) = P\{X = j | X \geq j\} = f(j)/(1 - F(j^-))$, and cumulative hazard rate $H(j) = \sum_{i=0}^{j} h(i)$. Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ be a random sample subjected to right censorship, and thus the data consists of $Z_i = \min\{X_i, y_i\}$ and $\delta_i = I[X_i \leq y_i]$, y_i being censoring time for the i-th individual, $i = 1, \ldots, n$. Let N be the counting process of uncensored observations and M the number-at-risk process given by

$$\begin{aligned} N(j) &= \sum_{i=1}^{n} I[Z_i \leq j \text{ and } \delta_i = 1]; \quad \text{and} \\ M(j) &= \sum_{i=1}^{n} I[Z_i \geq j], \quad j \geq 0. \end{aligned} \tag{3.71}$$

Treating H as a stochastic process with independent summands and having a beta process $\mathcal{B}e\{c(\cdot), H_0(\cdot)\}$ prior, $h(j)$ is distributed as $Be\{c(j)h_0(j), c(j)(1 - h_0(j))\}$. Then as noted before, $\mathcal{E}(h(j)) = h_0(j) = dH_0(j)$ is the prior guess of $h(j)$ and $\mathcal{E}(H(j)) = H_0(j)$, and $\text{Var}(h(j)) = h_0(j)(1 - h_0(j))/[c(j) + 1]$ as the prior 'uncertainty'. The posterior distribution of H (discrete time version of property 4 of Sect. 1.8) is

$$H|data \sim \mathcal{B}e\left\{c + M, \sum \frac{cdH_0 + dN}{c + M}\right\}, \tag{3.72}$$

and the nonparametric Bayesian estimator of H under the weighted quadratic loss is given by

$$\widehat{H}(j) = \mathcal{E}(H|data) = \sum_{i=0}^{n} \frac{c(i)h_0(i) + dN(i)}{c(i) + M(i)}. \tag{3.73}$$

Similarly in the time-continuous case, it is shown (not to minimize the efforts required) that the analysis leads to the estimator,

$$\widehat{H}(t) = \int_0^t \frac{c(s)dH_0(s) + dN(s)}{c(s) + M(s)}.$$

It is worth noting that as $c(\cdot) \to 0$, these estimators reduce to the usual nonparametric Nelson-Aalen estimator and as $c(\cdot) \to \infty$, it simply reduces to the prior guess H_0—properties observed earlier for the distribution functions.

3.5.5 *Estimation of Hazard Rates*

Assume that a hazard rate $r(t)$ has an extended gamma process prior (see Sect. 1.7) $\Gamma(\alpha(\cdot), \beta(\cdot))$. Given a sample of n observations, the posterior distribution of $r(t)$ was given in expression (1.80) for censored and in expression (1.81) for exact observations.

The posterior mean of $r(t)$ based on m observations (censored and exact) is

$$\widehat{r}(t) = \frac{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \int_{[0,t)} \prod_{i=0}^m \beta^*(z_i) \prod_{i=0}^m d[\alpha + \sum_{j=i+1}^m I_{(x_j,\infty)}](z_i)}{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \prod_{i=1}^m \beta^*(z_i) \prod_{i=1}^m d[\alpha + \sum_{j=i+1}^m I_{(x_j,\infty)}](z_i)}. \tag{3.74}$$

Clearly, $\widehat{r}(t)$ is the Bayes estimator under the loss function $\int_{[0,\infty)} (r(t) - \widehat{r}(t))^2 dW(t)$ subject to the condition

$$\int_{[0,\infty)} \int_{[0,t)} \beta^2(s) d\alpha(s) dW(t) < \infty. \tag{3.75}$$

For the computational purposes, they show that the multi-dimensional integrals in $\widehat{r}(t)$ can be reduced to a type that involves only one-dimensional integrals.

The Bayesian estimator of the survival function derived under the assumption $r(t) \backsim \Gamma(\alpha(\cdot), \beta(\cdot))$ is given in Sect. 3.3.8.

3.5.6 *Markov Chain Application*

Hjort (1990) extends the results of the last subsection to the case of nonhomogeneous Markov Chain and obtains Bayesian estimators of transition probabilities and cumulative hazard rates. Let $X = \{X(r) : r = 0, 1, 2, \ldots\}$ be a Markov chain with state space $\{1, \ldots, k\}$ and transition probabilities from i to j,

$$p_{ij}(r, s) = P\{X(s) = j | X(r) = i\}, \quad 0 \le r \le s, i, j = 1, \ldots, k. \tag{3.76}$$

The treatment in the last subsection corresponds to the state $\{1, 2\}$ and the possible transitions being only from state 1 to 2. Also, $h(j)$ corresponds to one-step probabilities $h_{ij}(s) = p_{ij}(s-1, s)$ and the cumulative hazard rate from i to j is $H_{ij}(s) = \sum_{r=1}^{s} h_{ij}(r)$, $s \geq 1$. The data now available is of the form X observed up to and including time t, and let $X_t = \{X(r) : r = 0, 1, 2, \ldots, t\}$ collected on n individuals moving around in the state space independently of each other, each with transition probability p_{ij}. Let the data be represented as follows. $X_{t(l)}^{(l)} = \{X^{(l)}(r) : r = 0, 1, \ldots, t(l)\}, l = 1, 2, \ldots, n,$

$$dN_{ij}(r) = \sum_{l=1}^{n} I\big[X^{(l)}(r-1) = i, X^{(l)}(r) = j\big]; \quad \text{and} \tag{3.77}$$

$$M_i(r) = \sum_{l=1}^{n} I\big[X^{(l)}(r-1) = i, r \leq t(l)\big], \quad r \geq 1. \tag{3.78}$$

Here, $M_i(r)$ is the number of individuals at risk in state i just before time r and are subject to transition probability $h_{ij}(r)$ to one of $k-1$ states, $j \neq i$, or may remain in state i with probability $h_{ii}(r) = 1 - \sum_{j \neq i, j=1}^{k} h_{ij}(r)$. $M_i(r)$ does not include those that had $X^{(l)}(r-1) = i$, but were censored before r. The increments $dN_{ij}(r)$ add up to counting processes N_{ij}, and $N_{ij}(s)$ counts the number of transitions i to j observed in the time interval $(0, s]$.

Assuming a prior distribution for the $k(k-1)$ cumulative hazard rates H_{ij} which specifies that its summands are independent and that k rows of $h(r)$ are independently distributed according to a Dirichlet distribution with parameters $c_i(r)h_{0i1}(r), \ldots, c_i(r)h_{0ik}(r)$ for the i-th row. Thus, $\mathcal{E}(h_{ij}(r)) = h_{0ij}(r)$ and therefore, $H_{0ij}(s) = \sum_{r=1}^{s} h_{0ij}(r)$, is the prior guess at H_{ij}. Then the nonparametric Bayesian estimator for $h_{ij}(r)$ with respect to a quadratic loss is the posterior mean

$$\widehat{h}_{ij}(r) = \mathcal{E}\big(h_{ij}(r)|data\big) = \frac{c_i(r)h_{0i1}(r) + dN_{ij}(r)}{c_i(r) + M_i(r)}, \tag{3.79}$$

and for $H_{ij}(s)$ is

$$\widehat{H}_{ij}(s) = \sum_{r=1}^{s} \frac{c_i(r)h_{0i1}(r) + dN_{ij}(r)}{c_i(r) + M_i(r)}, \quad s \geq 1. \tag{3.80}$$

Similarly, the Bayes estimator of waiting time distribution F_i for state i, defined as $F_i(s) = P\{X \text{ leaves } i \text{ before time } s\}$ is obtained as

$$\widehat{F}_i(s) = 1 - \prod_{r=1}^{s} \left[1 - \frac{c_i(r)dH_{0,i\cdot}(r) + dN_{i\cdot}(r)}{c_i(r) + M_i(r)}\right]. \tag{3.81}$$

Similar analysis in the time-continuous case led to the estimator

$$\widehat{H}_{ij}(t) = \int_0^t \frac{c_{ij}(s)dH_{0,ij}(s) + dN_{ij}(s)}{c_{ij}(s) + M_i(s)}. \tag{3.82}$$

The estimate of waiting time distribution $G_i([s,t]) = P\{X(u) \equiv i, u \in [s,t] | X(s) = i\} = \prod_{[s,t]}\{1 - dH_{i\cdot}(u)\}$ for $s \le t$, is given by

$$\widehat{G}_i\big([s,t]\big) = \prod_{[s,t]}\left[1 - \frac{c_i dH_{0,i\cdot} + dN_{i\cdot}}{c_i + M_i}\right], \tag{3.83}$$

where $N_{i\cdot} = \sum_{j \ne i} N_{ij}$, $H_{0,i\cdot} = \sum_{j \ne i} H_{0,ij}$.

3.5.7 Estimation for a Shock Model

A problem of Bayesian analysis of shock models and wear processes using the Dirichlet process prior was studied in Lo (1981). Suppose a device is subject to shocks occurring randomly according to a Poisson process $N = \{N(t); t \in R\}$ with intensity parameter λ. The i-th shock inflicts a random amount X_i, $i = 1, 2 \ldots$ of damage on the device. The X_i's are assumed to be iid F defined on R^+. The process is observed until a fixed time T. Thus we have for the data, $N(T)$ the number of shocks occurring during the time interval $[0, T]$, and $X_1, \ldots, X_{N(T)}$, the amounts of damages. The task is to estimate the survival probability $S(t)$ that the device survives beyond time $t \in [0, T]$. Lo considers the nonparametric Bayesian estimation approach to this problem and derives the Bayes estimator for $S(t)$. He assumes the pair (λ, F) to be independent random variables and places a gamma $G(\nu, \theta)$ and a Dirichlet process $\mathcal{D}(\alpha)$ priors on λ and F, respectively. He derives the posterior distribution of (λ, F) given the process $N(t)$ up to time T, denoted by $\underline{N}_T$ and $\mathbf{X}_T = (X_1, \ldots, X_{N(T)})$, and shows that it has again the same structure as the prior but with parameters updated. Also, λ and F turn out to be independent as one would expect. Symbolically,

$$(\lambda, F)|\underline{N}_T, \mathbf{X}_T \sim G\big(\nu + N(T), \theta + T\big) \times \mathcal{D}\left(\alpha + \sum_{i=1}^{N(T)} \delta_{X_i}\right). \tag{3.84}$$

Now for any f, a real valued integrable function, the conditional expectation of $f(\lambda, F)$, given the data, i.e. $\mathcal{E}[f(\lambda, F)|\underline{N}_T, \mathbf{X}_T]$ can be computed. In particular, he deals with the Bayesian estimator of survival probability $S(t) = \sum_{k=0}^{\infty} \overline{P}_k e^{-\lambda t}(\lambda t)^k / k!$, where $\overline{P}_k$ is the probability that the device survives k shocks during the time interval $[0, t]$, known as the capacity or threshold of the device.

Then the Bayes estimator of S under the loss function $L(S, \widehat{S}) = \int_R (S(t) - \widehat{S}(t))^2 dW(t)$ is obtained as

$$\widehat{S}(t) = \mathcal{E}\big[S(t)|\underline{N}_T, \mathbf{X}_T\big] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathcal{E}[\overline{P}_k|\underline{N}_T, \mathbf{X}_T]\mathcal{E}\big[e^{-\lambda t}\lambda^k|\underline{N}_T, \mathbf{X}_T\big]. \tag{3.85}$$

Using the fact that $\overline{P}_k$ depends on F only, and the fact that λ and F are independent under the posterior distribution, the estimator reduces to

$$\widehat{S}(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{\Gamma(\nu + N(T) + k)}{\Gamma(\nu + N(T))} \times \left(\frac{1}{\theta + T + t}\right)^k \left(\frac{\theta + T}{\theta + T + t}\right)^{\nu + N(T)} \cdot \mathcal{E}[\overline{P}_k | \underline{N}_T, \mathbf{X}_T]. \tag{3.86}$$

He evaluates this expression in closed form for three particular cases of $\overline{P}_k$. They are (1) $\overline{P}_k = P\{X_1 + \ldots + X_k \leq y | N(t) = k\}$, the sum of the damages does not exceed the capacity or threshold; (2) $\overline{P}_k = \prod_{i=1}^k F(y_i)$, the threshold values changes after each shock; and (3) $\overline{P}_k = [F(y)]^k$, the fixed threshold model. Details can be found in his paper.

In another paper Lo (1982) treats the problem of estimation of the intensity parameter γ of a nonhomogeneous Poisson point process based on a random sample from the process. Assuming a weighted gamma distribution $G(\alpha, \beta)$ as prior for γ, and given a random sample $N_1, \ldots, N_n$ of n functions from this process, he shows that the posterior distribution of γ is again a gamma distribution $G(\alpha + \sum_{j=1}^n N_j, \beta/(n\beta + 1))$.

Kim (1999) considers a more general model called the *multiplicative intensity model*. A stochastic process $N(t)$ defined on the time interval $[0, T]$ is called a counting process if the sample paths are right continuous step functions with $N(0) = 0$ and having a finite number of jumps, each of size one. It is called a multiplicative intensity model if the cumulative intensity process has certain form. Kim considers nonparametric inference procedures for such a model. Lo's result may be considered as a special case of Kim's treatment. Further Kim adopts a semimartingale approach instead of the Lévy measure approach for the independent increment processes.

3.5.8 Estimation for a Age-Dependent Branching Process

An interesting application of the Dirichlet process is given in Johnson et al. (1979) for the Bayesian estimation of the distributions of offsprings and life-lengths in the context of a Bellman-Harris branching process. It is based on the family tree up to time T. Start with a population of one ancestor. Using their notation, with every offspring x, associate one nonnegative random variable t_x, the life-length, and a point process n_x, the reproduction of x. Assume the pairs (n_x, t_x) as iid with probability distribution $P \times G$, where $n_x \sim P$, known as offspring distribution, and $t_x \sim G$, the life-length distribution of the individual in the process. P is taken as a discrete distribution $\{p_j\}_{j=0}^{\infty}$, $p_j \geq 0$ for all j and $\sum_{j=0}^{\infty} p_j = 1$, and G a distribution function on $(0, \infty)$. The authors derive Bayesian estimators of P and G and give explicit expressions as shown below.

Assume P and G to be independent having Dirichlet processes $\mathcal{D}(\alpha_1)$ and $\mathcal{D}(\alpha_2)$, with finite non-null measures α_1 and α_2 as priors, respectively. The support of α_1 is restricted to $\mathbb{N} = \{0, 1, 2, \dots\}$ and the loss function assumed is,

$$L\big((P, G), (\widehat{P}, \widehat{G})\big) = a_1 \sum_{j=0}^{\infty} W_1(j)(p_j - \widehat{p}_j)^2 + a_2 \int_0^{\infty} \big(G(t) - \widehat{G}(t)\big)^2 dW_2(t), \tag{3.87}$$

where $a_1, a_2 \geq 0$ and W_1 and W_2 are known weight functions on $\mathbb{N}$ and $(0, \infty)$, respectively. For data we have $N_l(T) = \#$ of splits of size l in $[0, T]$, $l = 0, 1, 2, \dots$, $D_{t_1}, \dots, D_{t_n}$ age at death of n individuals who died in $[0, T]$, and $S_{t_1}, \dots, S_{t_m}$ survival times of m individuals who survived time T. Based on this data, Bayes estimators of P and G, assuming them to be independent of each other, are obtained as follows.

1. The Bayes estimator of P under the above loss function is given by $\widehat{P} = \{\widehat{p}_j\}$, where
$$\widehat{p}_j = \frac{\alpha_1(j) + N_l(T) I[j = l]}{\alpha_1(\mathbb{N}) + \sum_{l=0}^{\infty} N_l(T)}. \tag{3.88}$$
2. The conditional distribution of $P | N_l(T)$, $l \in \mathbb{N}$ is $\mathcal{D}(\alpha_1(\cdot) + \sum_{l=0}^{\infty} N_l(T) I[l \in (\cdot)])$.
3. Assuming α_2 to be nonatomic, the Bayes estimator of G under the above loss function is given by $\widehat{G}$, where
$$1 - \widehat{G}(t) = \frac{\alpha_2(t, \infty) + N^+(t)}{\alpha_2\big(R^+\big) + m + n} \prod_{j=1}^{k} \left(\frac{\alpha_2(S^*_{t_j}, \infty) + N(S^*_{t_j})}{\alpha_2(S^*_{t_j}, \infty) + N(S^*_{t_j}) - \lambda^*_j} \right)^{I[S_{t_j} \leq t]}, \tag{3.89}$$
where $N^+(t) = \#$ of deaths and survival times greater than t, $N(t) = \#$ of deaths and survival times greater than or equal to t, $S^*_{t_1}, \dots, S^*_{t_k}$ are the k distinct observations among $S_{t_1}, \dots, S_{t_m}$, and λ^*_j are their multiplicities.

Remark 3.2 As $\alpha_1(\mathbb{N}) \to 0$, $\widehat{p}_j \to$ MLE of p_j. The Bayes estimator under squared error loss of the mean M of the offspring distribution P is given by

$$\widehat{M} = \frac{\sum_{l=0}^{\infty} l \alpha_1(\{l\}) + \sum_{l=0}^{\infty} l N_l(T)}{\alpha_1(\mathbb{N}) + \sum_{l=0}^{\infty} N_l(T)}. \tag{3.90}$$

Remark 3.3 The estimator $1 - \widehat{G}(t)$ looks similar to the Susarla-Van Ryzin estimator of the survival function, but two vital differences were noted by the authors. In that estimator, the total sample size n which includes censored and uncensored observations, is known and fixed constant ahead of the sampling. Here n and m are random variables. Second, in that treatment, censoring times were taken to be independent of the survival times. Here the censoring random variables associated with $S_{t_1}, \ldots, S_{t_m}$ are not independent of the random life-times $D_{t_1}, \ldots, D_{t_n}$.

They also treat the case when P and G are not independent, and the prior for the pair (P, G) is taken to be the Dirichlet process defined on the product space $\mathbb{N} \times R^+$, and point out that the estimator $\widehat{p}_j$ in this case not only depends on the splits n_x but also on the life-lengths and survival times.

3.6 Hypothesis Testing $H_0 : F \leq G$

Earlier in Sect. 2.9.1, hypothesis testing relative to the null hypothesis $H_0 : F \leq F_0$ against the alternative $H_1 : F \nleq F_0$ was considered from a decision theoretic point of view. In this section we consider its two-sample analog when the data is right censored. The case of uncensored data can easily be handled as a special case.

In the non-Bayesian context, Gehan (1965) had obtained procedures to test the hypothesis $H_0 : F = G$ against one and two sided alternatives for the censored data. His test statistic was a natural extension of Wilcoxon-Mann-Whitney statistic. Efron (1967) produced an alternative test statistic as an improvement over Gehan's. Phadia and Susarla (1979) used a decision theoretic approach for this testing problem. The statistic that emerged is quite different from those of Gehan (1965) and Efron (1967).

Using the same notations as earlier in the two-sample problem (see Sect. 3.5.2), we want to test the hypothesis $H_0 : F \leq G$ against $H_1 : F \nleq G$ based on the data $\{\mathbf{S}, \boldsymbol{\delta}, \mathbf{T}, \boldsymbol{\eta}\}$. F and G are considered as *right sided* distribution functions. The loss function used here is an appropriate modification of that used in one sample case (see Sect. 2.9).

$$\begin{aligned} L\big((F, G), a_0\big) &= \int (F - G)^+ dW \quad \text{and} \\ L\big((F, G), a_1\big) &= \int (F - G)^- dW, \end{aligned} \tag{3.91}$$

where $L((F, G), a_i)$ indicates the loss when action a_i (deciding in favor of H_i) is taken for $i = 0, 1$, W is a weight function, $\mathrm{a}^+ = \max\{a, 0\}$ and $\mathrm{a}^- = -\min\{a, 0\}$ for any $a \in R$, as before. Assume F and G to have Dirichlet process priors with parameters α_1 and α_2, respectively. Let $\xi_n = \mathcal{P}\{\text{taking action } a_0 \mid \textit{data}\}$. Then the Bayes rule with respect to these priors is given by the test statistic

$$\xi_n = I\big[\psi(\alpha_1, \alpha_2) \leq 0\big], \tag{3.92}$$

where

$$\begin{aligned} \psi(\alpha_1, \alpha_2) &= \mathcal{E}\big[L\big((F, G), a_0\big) - L\big((F, G), a_1\big) \mid \textit{data}\big] \\ &= \int \big(\widehat{F}_{\alpha_1}(u) - \widehat{G}_{\alpha_2}(u)\big) dW(u), \end{aligned} \tag{3.93}$$

and $\widehat{F}_{\alpha_1}$ and $\widehat{G}_{\alpha_2}$ are the Bayes estimators of F and G derived earlier. The minimum Bayes risk against the Dirichlet process priors is

$$\begin{aligned} r_n^*(\alpha_1, \alpha_2) &= \inf_{\xi_n}(\alpha_1, \alpha_2, \xi_n) \\ &= \mathcal{E}\big[I\big[\psi(\alpha_1, \alpha_2) \le 0\big]\psi(\alpha_1, \alpha_2)\big] + \mathcal{E}\big[L\big((F, G), a_1\big)\big], \end{aligned} \tag{3.94}$$

which can easily be evaluated. When α_1 and α_2 are unknown, the empirical Bayes method of earlier sections was recommended. The test statistics ξ_n is replaced at the $(n+1)$-stage by $\widehat{\xi}_{n+1}$with ψ replaced by $\widehat{\psi}_{n+1}$ given by $\widehat{\psi}_{n+1}(\cdot) = \int(\widehat{F}_{\widehat{\alpha}_1}(u) - \widehat{G}_{\widehat{\alpha}_2}(u))dW(u)$, where as before, $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ given in (3.58) may be used. It is shown that the above test statistic is asymptotically optimal with the rate of convergence $O(n^{1/2})$.

Damien and Walker (2002) present a different approach for comparing two treatments in which the Bayes Factor is used.

3.7 Estimation in Presence of Covariates

Most of the analysis presented for the censored data can be extended to incorporate regressor variables. To accommodate covariates in the analysis of right censored survival data, a common practice in the non-Bayesian context is to use the Cox's model. Kalbfleisch (1978) and Wild and Kalbfleisch (1981) initiated this approach in the Bayesian framework

Suppose we have positive survival times X_i's distributed according to a distribution function F and is associated with a set of covariates $\mathbf{W}_i^T = (W_{i1}, \ldots, W_{ik})$, a transpose of the column vector. The Cox's model, also known as the proportional hazard model, is expressed as $S(t; \mathbf{W}) = S_0(t)^{\exp(\boldsymbol{\beta}\mathbf{W})}$, where $\boldsymbol{\beta}$ is a row vector of regression coefficients and S_0 is the baseline survival function, or in terms of hazard rates, as $\lambda(t|\mathbf{w}) = \lambda_0(t)\exp(\boldsymbol{\beta}\mathbf{W})$. The main interest in covariate data analysis centers around the estimation and hypothesis testing of $\boldsymbol{\beta}$, and in such cases, S_0 may be regarded as a nuisance parameter. On the other hand, one may be interested in the estimation of S_0 itself.

As stated in Chap. 1, Kalbfleisch considered these problems from a Bayesian approach. Writing $S_0(t) = e^{-H(t)}$, it is immediately clear that $H(t)$ may be viewed as a nondecreasing process with independent nonnegative increments. Thus with $H(0) = 0$, and as $t \to \infty$, $H(t) \to \infty$, the theory of the nondecreasing processes with independent increments can be used. For the task of covariate analysis, Kalbfleisch treats $H(t)$ as a nuisance parameter having a certain prior distribution and carried out the estimation of $\boldsymbol{\beta}$ by determining the marginal distribution of observations as a function of $\boldsymbol{\beta}$ having $H(t)$ eliminated. For the prior it was convenient to choose specifically a gamma process with parameters $cH_0(t)$ and c, so that $\mathcal{E}(H(t)) = H_0(t)$ and $\text{Var}(H(t)) = H_0(t)/c$. The parameter H_0 serves as a prior guess at H, and c the precision parameter. To avoid difficulties of prior fixed points of discontinuities, he assumed H_0 to be absolutely continuous. The gamma process is easy to handle in deriving the posterior distribution of $H(t)$ given the observations, and thus one can easily compute $\mathcal{E}(e^{-\Lambda(t)}|\text{observations})$ for a fixed $\boldsymbol{\beta}$.

Following the paper of Ferguson and Phadia (1979), Wild and Kalbfleisch (1981) placed the above approach in a more general setting by using the processes neutral to the right as priors for $H(t)$ instead of the gamma process. By proposing a simple adjustment in the derivations given by Ferguson and Phadia, they were able to extend Ferguson and Phadia's treatment to the regression analysis problem. If we let $Y_t = -\log S(t)$ and $Y_{0t} = -\log S_0(t)$, then $Y_t = Y_{0t}e^{\boldsymbol{\beta}\mathbf{W}}$. Since $e^{\boldsymbol{\beta}\mathbf{W}}$ is treated as non-random, it is easy to see that if $F_0 = 1 - S_0$ is neutral to the right, so is $F = 1 - S$. Now the theorems of Ferguson and Phadia are applicable with the following variations. The posterior density of an increment in Y_{0t} is obtained by multiplying the prior density, say, $dG(y)$ with $\exp(-ye^{\boldsymbol{\beta}\mathbf{W}})$ and normalizing; likewise the distribution of jump at x should be adjusted. Similarly, if the given observation is censored, the posterior distribution of an increment changes only to the left of x and is found by multiplying the prior density by $\exp(-ye^{\boldsymbol{\beta}\mathbf{W}})$ (instead of $\exp(-y)$) and renormalizing.

The full description of the posterior distribution for the sample size $n = 1$ can be reformulated as follows:

Theorem 3.4 (Wild and Kalbfleisch) *Let F_0 be a random distribution neutral to the right. Then given an observation $X = x$ or $X > x$ from F, the posterior distribution of F_0 is also neutral to the right.*

For $X = x$,

(i) *the posterior distribution of an increment in Y_{0t} to the right of x is the same as the prior distribution.*

(ii) *An increment $Y_{0t} - Y_{0s}$ for $s < t$ left of x with a prior density $dG(y)$ has the posterior density proportional to* $\exp(-ye^{\boldsymbol{\beta}\mathbf{W}})dG(y)$.

(iii) *There is a jump discontinuity $J = Y_{0x} - Y_{0x}^{-}$ at x, in the posterior whether there was one in the prior or not. If the prior density of J is assumed to be $dG_x(s)$, then it has the posterior density proportional to* $(1 - \exp(-se^{\boldsymbol{\beta}\mathbf{W}}))dG_x(s)$.

For the case $X > x$,

(i) *the posterior distribution of an increment in Y_{0t} to the right of x is the same as the prior distribution.*

(ii) *The posterior distribution of an increment $Y_{0t} - Y_{0s}$ for $s < t \le x$ has the description same as in* (iii) *above.*

For the general case of sample size $n > 1$, it is convenient to derive a formula for the posterior moment generating function (MGF). It is essentially the same as in Ferguson and Phadia except that now with the adjustments mentioned above, instead of counting the number of observations at a point or beyond a point, we count the exponential scores $e^{\boldsymbol{\beta}\mathbf{W}_i}$.

Let $u_1, \ldots, u_k$ be the distinct values among $x_1, \ldots, x_n$ ordered so that $u_1 < u_2 < \ldots < u_k$. Let $C(i)$ and $D(i)$, respectively, be the sets of labels of censored and uncensored observations at u_i, $R(i)$ be the set of labels of all observations that are greater than u_i, $G_{u_i}(s)$ be the prior distribution of a jump, J in Y_{0t} at u_i, $H_{u_i}(s)$ be the posterior distribution of J given that a failure occurred at u_i, $M_t(\theta) = \mathcal{E}(\exp(-Y_{0t}\theta))$ denote the MGF of Y_{0t}, and $M_t^-(\theta)$ denote the MGF of Y_t^-, $M_t^-(\theta) = \lim_{s\to t} M_s(\theta)$, $s < t$. Then we have

Theorem 3.5 (Wild and Kalbfleisch) *Let F_0 be a random distribution function neutral to the right, and let $X_1, \ldots, X_n$, be a sample of independent observations such that X_i is distributed according to $F_i = 1 - (1 - F_0(t))^{\exp(\boldsymbol{\beta}\mathbf{W}_i)}$, where $\mathbf{W}_i$ is the associated vector of covariates for X_i. Then the posterior distribution of F_0 given the data is neutral to the right, and Y_{0t} has posterior MGF*

$$M_t(\theta \mid data) = \frac{M_t(\theta + h_{j(t)})}{M_t(h_{j(t)})} \times \prod_{i=1}^{j(t)} \left[\frac{M_{u_i}^-(\theta + h_{i-1})}{M_{u_i}^-(h_{i-1})} \cdot \frac{C_{u_i}(\theta + h_i + \lambda_i, d_i)}{C_{u_i}(h_i + \lambda_i, d_i)} \cdot \frac{M_{u_i}(h_i)}{M_{u_i}(\theta + h_i)} \right], \tag{3.95}$$

where $h_i = \sum_{l\in R(i)} \exp(\boldsymbol{\beta}\mathbf{W}_l)$, $\lambda_i = \sum_{l\in C(i)} \exp(\boldsymbol{\beta}\mathbf{W}_l)$ and d_i is the number of observations in $D(i)$.

If u_i is a prior fixed point of discontinuity of Y_{0t}, then

$$C_{u_i}(\alpha, d_i) = \int_0^\infty e^{-\alpha z} \prod_{l\in D(i)} \left(1 - \exp\left(-ze^{\boldsymbol{\beta}\mathbf{W}_l}\right)\right) dG_{u_i}(z), \tag{3.96}$$

where the product is taken over d_i observations in $D(i)$; while, if u_i is not a prior fixed point of discontinuity of Y_{0t}, then

$$C_{u_i}(\alpha, d_i) = \begin{cases} 1 & \text{if } d_i = 0 \\ \int_0^\infty e^{-\alpha z} dH_{u_i}(z) & \text{if } d_i = 1 \\ \int_0^\infty e^{-\alpha z} \prod\{1 - \exp(-ze^{\boldsymbol{\beta}\mathbf{W}_l})\} dH_u(z) & \text{if } d_i > 1, \end{cases} \tag{3.97}$$

where the product is taken over d_{i-1} observations. (Note that one observation is needed to generate a fixed point of discontinuity at u_i (see Ferguson and Phadia 1979).)

In application, difficulties are encountered in evaluating the posterior distribution H_u of a jump at u where a single observation fell. However, as noted earlier, for certain specific processes neutral to the right it is relatively simple. The gamma process prior is one such process and Wild and Kalbfleisch evaluate the Bayes estimator of F_0 in this particular case. For the gamma process prior, the independent increments of the process Y_{0t} have gamma distributions with shape parameter $\nu(t)$ and intensity parameter τ. Since for this homogeneous process there are no prior fixed points of discontinuities, we need to consider only the second part of the above formula.

For this case, an application of Theorem 5 of Ferguson and Phadia for the gamma process, yields

$$\frac{C_{u_i}(\alpha+1,d_i)}{C_{u_i}(\alpha,d_i)}=\begin{cases}1 & \text{if } d_i=0\\ \log\frac{\alpha+1+\exp(\boldsymbol{\beta}\mathbf{W})}{\alpha+1}\Big/\log\frac{\alpha+\exp(\boldsymbol{\beta}\mathbf{W})}{\alpha} & \text{if } d_i=1\\ \frac{\int_0^\infty e^{-(\alpha+1)z}\prod_{l\in D(i)}\{1-\exp(-ze^{\boldsymbol{\beta}\mathbf{W}_l})\}z^{-1}dz}{\int_0^\infty e^{-\alpha z}\prod_{l\in D(i)}\{1-\exp(-ze^{\boldsymbol{\beta}\mathbf{W}_l})\}z^{-1}dz} & \text{if } d_i>0.\end{cases} \tag{3.98}$$

Further noting that the MGF of the gamma process is $M_t(\theta)=(\tau/(\tau+\theta))^{\nu(t)}$, and putting the various quantities together in the formula for the posterior MGF, they obtain an expression for the Bayes estimator similar to (3.15) in Ferguson-Phadia paper.

$$\begin{aligned}E\big(1-F(t)|data\big)&=M_t(1|data)\\&=\left(\frac{h_{j(t)}+\tau}{h_{j(t)}+\tau+1}\right)^{\gamma(t)}\\&\quad\times\prod_{i=1}^{j(t)}\left[\left(\frac{(h_{i-1}+\tau)(h_i+\tau+1)}{(h_{i-1}+\tau+1)(h_i+\tau)}\right)^{\gamma(u_i)}\right.\\&\quad\left.\times\frac{C_{u_i}(h_i+\lambda_i+\tau+1,d_i)}{C_{u_i}(h_i+\lambda_i+\tau,d_i)}\right].\end{aligned}\tag{3.99}$$

The difference is in the quantities

$$h_i=\sum_{l\in R(i)}\exp(\boldsymbol{\beta}\mathbf{W}_l),\qquad \lambda_i=\sum_{l\in C(i)}\exp(\boldsymbol{\beta}\mathbf{W}_l)\tag{3.100}$$

and the ratio of C_{u_i}'s (as defined above) is used in place of the ratios of φ_G's.

Burridge (1981) extends the above analysis to group data. Clayton (1991) develops computational procedures for the results. His model assumes the increments in the cumulative hazard to be nonnegative and independent in disjoint intervals and uses the gamma process to model the baseline cumulative hazard function. This approach has the disadvantage of highly discrete and independent hazards in disjoint intervals. Sinha (1998) presents an analysis of using a correlated prior process for the baseline hazard.

In his paper on Beta process, Hjort (1990) extends the covariate analysis by recasting the Cox model in terms of hazard functions as $1-dH_i(s)=\{1-dH(s)\}^{\exp(\boldsymbol{\beta}\mathbf{w}_i)}$, where $dH_i(s)$ is the hazard function of the i-th individual. By assuming $\boldsymbol{\beta}$ known and $H\in\mathcal{B}e\{c(\cdot),H_0(s)\}$, he shows that the posterior distribution of H given the data is a process with independent increments and is distributed again like a beta process between jumps. But at jumps the distribution is somewhat complicated. Using this approach, he derives the Bayes estimator for H under the weighted squared error loss. His work parallels the work of Wild and Kalbfleisch (1981) but the difference is that these authors found it necessary to assume the covariates to be constant in time, whereas in his derivation they can be time-dependent.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, *6*, 701–726.

Ammann, L. P. (1984). Bayesian nonparametric inference for quantal response data. *Annals of Statistics*, *12*, 636–645.

Ammann, L. P. (1985). Conditional Laplace transforms for Bayesian nonparametric inference in reliability theory. *Stochastic Processes and Their Applications*, *20*, 197–212.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, *2*, 1152–1174.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions*. New York: Wiley.

Basu, D., & Tiwari, R. C. (1982). A note on the Dirichlet process. In G. Kallianpur, P. R. Krishnaiah & J. K. Ghosh (Eds.), *Statistics and probability: essays in honor of C. R. Rao* (pp. 89–103).

Berry, D. A., & Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet process. *Annals of Statistics*, *7*, 558–568.

Bhattacharya, P. K. (1981). Posterior distribution of a Dirichlet process from quantal response data. *Annals of Statistics*, *9*, 803–811.

Binder, D. A. (1982). Nonparametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society. Series B. Methodological*, *44*, 388–393.

Blackwell, D. (1973). Discreteness of Ferguson selections. *Annals of Statistics*, *1*, 356–358.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *Annals of Statistics*, *1*, 353–355.

Blum, J., & Susarla, V. (1977). On the posterior distribution of a Dirichlet process given randomly right censored observations. *Stochastic Processes and Their Applications*, *5*, 207–211.

Breth, M. (1978). Bayesian confidence bands for a distribution function. *Annals of Statistics*, *6*, 649–657.

Breth, M. (1979). Nonparametric Bayesian interval estimation. *Biometrika*, *66*, 641–644.

Bulla, P., Muliere, P., & Walker, S. (2007). Bayesian nonparametric estimation of a bivariate survival function. *Statistica Sinica*, *17*, 427–444.

Bulla, P., Muliere, P., & Walker, S. (2009). A Bayesian nonparametric estimator of a multivariate survival function. *Journal of Statistical Planning and Inference*, *139*, 3639–3648.

Burridge, M. (1981). Empirical Bayes analysis of survival data. *Journal of the Royal Statistical Society. Series B. Methodological*, *43*, 65–75.

Campbell, G., & Hollander, M. (1978). Rank order estimation with the Dirichlet prior. *Annals of Statistics*, *6*(1), 142–153.

Christensen, R., Hanson, T., & Jara, A. (2008). Parametric nonparametric statistics: an introduction to mixtures of finite Polya trees. *The American Statistician*, *62*, 296–306.

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

Chung, Y., & Dunson, D. B. (2011). The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*, *63*, 59–80.

Cifarelli, D. M., & Regazzini, E. (1979a). Considerazioni generali sull'impostazione bayesiana di problemi non parametrici, Part I. *Rivista di matematica per le scienze economiche e sociali*, *2*, 39–52.

Cifarelli, D. M., & Regazzini, E. (1979b). Considerazioni generali sull'impostazione bayesiana di problemi non parametrici, Part II. *Rivista di matematica per le scienze economiche e sociali*, *2*, 95–111.

Clayton, M. K. (1985). A Bayesian nonparametric sequential test for the mean of a population. *Annals of Statistics*, *13*, 1129–1139.

Clayton, M. K. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrika*, *47*, 467–485.

Clayton, M. K., & Berry, D. (1985). Bayesian nonparametric bandits. *Annals of Statistics*, *13*, 1523–1534.

Connor, R. J., & Mosimann, J. E. (1969). Concept of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, *64*, 194–206.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Annals of Statistics*, *16*, 1475–1489.

Dalal, S. R. (1979a). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Processes and Their Applications*, *9*, 99–107.

Dalal, S. R. (1979b). Nonparametric and robust Bayes estimation of location. In *Optimizing methods in statistics* (pp. 141–166). New York: Academic Press.

Dalal, S. R., & Hall, G. J. (1980). On approximating parametric Bayes models by nonparametric Bayes models. *Annals of Statistics*, *8*, 664–672.

Dalal, S. R., & Phadia, E. G. (1983). Nonparametric Bayes inference for concordance in bivariate distributions. *Communications in Statistics. Theory and Methods*, *12*(8), 947–963.

Damien, P., & Walker, S. (2002). A Bayesian nonparametric comparison of two treatments. *Scandinavian Journal of Statistics*, *29*, 51–56.

Damien, P., Laud, P. W., & Smith, A. F. M. (1995). Approximate random variate generation form infinitely divisible distributions with applications to Bayesian inference. *Journal of the Royal Statistical Society. Series B. Methodological*, *57*, 547–563.

Damien, P., Laud, P. W., & Smith, A. F. M. (1996). Implementation of Bayesian non-parametric inference based on beta processes. *Scandinavian Journal of Statistics*, *23*, 27–36.

Dey, D., Müller, P., & Sinha, D. (Eds.) (1998). *Lecture notes in statistics*. *Practical nonparametric and semiparametric Bayesian statistics*. New York: Springer.

Dey, J., Erickson, R. V., & Ramamoorthi, R. V. (2003). Some aspects of neutral to right priors. *International Statistical Review*, *71*(2), 383–401.

Doksum, K. A. (1972). Decision theory for some nonparametric models. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, vol. 1: theory of statistics* (pp. 331–343).

Doksum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability*, *2*, 183–201.

Doss, H. (1984). Bayesian estimation in the symmetric location problem. *Zeitschrift Für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *68*, 127–147.

Doss, H. (1985a). Bayesian nonparametric estimation of the median; part I: computation of the estimates. *Annals of Statistics*, *13*, 1432–1444.

Doss, H. (1985b). Bayesian nonparametric estimation of the median; part II: asymptotic properties of the estimates. *Annals of Statistics*, *13*, 1445–1464.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics*, *22*, 1763–1786.

Drăghici, L., & Ramamoorthi, R. V. (2000). A note on the absolute continuity and singularity of Polya tree priors and posteriors. *Scandinavian Journal of Statistics*, *27*, 299–303.

Duan, J. A., Guindani, M., & Gelfand, A. E. (2007). Generalized spatial Dirichlet process model. *Biometrika*, *94*, 809–825.

Dubins, L. E., & Freedman, D. A. (1966). Random distribution functions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 2: contributions to probability theory, part 1* (pp. 183–214).

Dunson, D. B., & Park, J. H. (2008). Kernel stick-breaking processes. *Biometrika*, *95*, 307–323.

Dykstra, R. L., & Laud, P. (1981). A Bayesian nonparametric approach to reliability. *Annals of Statistics*, *9*, 356–367.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 4: biology and problems of health.*

Engen, S. (1975). A note on the geometric series as a species frequency model. *Biometrika*, *62*, 697–699.

Engen, S. (1978). *Stochastic abundance models with emphasis on biological communities and species diversity*. London: Chapman & Hall.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, *89*, 268–277.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*, 577–588.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, *3*, 87–112.

Fabius, J. (1964). Asymptotic behavior of Bayes estimates. *Annals of Mathematical Statistics*, *35*, 846–856.

Fabius, J. (1973). Neutrality and Dirichlet distributions. In *Transactions of the 6th Prague conference on information theory, statistical decision functions and random processes* (pp. 175–181).

Feller, W. (1966). *An introduction to probability theory and its applications* (Vol. II). New York: Wiley

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, *2*, 615–629.

Ferguson, T. S. (1982). Sequential estimation with Dirichlet process priors. In S. Gupta & J. Berger (Eds.), *Statistical decision theory and related topics III* (Vol. 1, pp. 385–401).

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In H. Rizvi & J. S. Rustagi (Eds.), *Recent advances in statistics* (pp. 287–302). New York: Academic Press.

Ferguson, T. S., & Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Annals of Statistics*, *7*, 163–186.

Ferguson, T. S., Phadia, E. G., & Tiwari, R. C. (1992). Bayesian nonparametric inference. In M. Ghosh & P. K. Pathak (Eds.), *IMS lecture notes—monograph series: Vol. 17. Current issues in statistical inference: essays in honor of D. Basu* (pp. 127–150).

Freedman, D. A. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Annals of Mathematical Statistics*, *34*, 1386–1403.

Gardiner, J. C., & Susarla, V. (1981). A nonparametric estimator of the survival function under progressive censoring. In J. Crowley & R. A. Johnson (Eds.), *IMS lecture notes—monograph series: Vol. 2. Survival analysis* (pp. 26–40).

Gardiner, J. C., & Susarla, V. (1983). Weak convergence of a Bayesian nonparametric estimator of the survival function under progressive censoring. *Statistics & Decisions*, *1*, 257–263.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, *52*, 203–223.

Gelfand, A. E., Kottas, A., & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, *100*, 1021–1035.

Ghahramani, Z., Griffiths, T. L., & Sollich, P. (2007). Bayesian nonparametric latent feature models (with discussion and rejoinder). In J. M. Bernardo et al. (Eds.), *Bayesian statistics* (Vol. 8).

Oxford: Oxford University Press.

Ghorai, J. K. (1981). Empirical Bayes estimation of a distribution function with a gamma process prior. *Communications in Statistics. Theory and Methods*, *10*(12), 1239–1248.

Ghorai, J. K. (1989). Nonparametric Bayesian estimation of a survival function under the proportional hazard model. *Communications in Statistics. Theory and Methods*, *18*(5), 1831–1842.

Ghorai, J. K., & Susarla, V. (1982). Empirical Bayes estimation of probability density function with Dirichlet process prior. In W. Grossmann et al. (Eds.), *Probability and statistical inference*. Dordrecht: Reidel.

Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Springer series in statistics*. *Bayesian nonparametric*. New York: Springer.

Ghosh, J. K., Hjort, N. L., Messan, C., & Ramamoorthi, R. V. (2006). Bayesian bivariate survival estimation. *Journal of Statistical Planning and Inference*, *136*, 2297–2308.

Ghosh, M. (1985). Nonparametric empirical Bayes estimation of certain functionals. *Communications in Statistics. Theory and Methods*, *14*(9), 2081–2094.

Ghosh, P., Lahari, M., & Tiwari, R. C. (1989). Nonparametric empirical Bayes estimation of the distribution and the mean. *Communications in Statistics. Theory and Methods*, *18*(1), 121–146.

Griffin, J. E., & Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, *101*, 179–194.

Griffiths, R. C. (1980). Unpublished notes.

Griffiths, T. L., & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in neural information processing systems* (Vol. 18). Cambridge: MIT.

Griffiths, T. L., & Ghahramani, Z. (2011). The Indian buffet process: an introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224.

Gross, A. J., & Clark, V. A. (1975). *Survival distributions. Reliability applications in biomedical sciences*. New York: Wiley.

Hall, G. J. Jr. (1976). Sequential search with random overlook probabilities. *Annals of Statistics*, *4*, 807–816.

Hall, G. J. Jr. (1977). Strongly optimal policies in sequential search with random overlook probabilities. *Annals of Statistics*, *5*, 124–135.

Hannum, R. C., & Hollander, M. (1983a). Robustness of Ferguson's Bayes estimator of a distribution function. *Annals of Statistics*, *11*, 632–639.

Hannum, R. C., & Hollander, M. (1983b). Correction: Robustness of Ferguson's Bayes estimator of a distribution function. *Annals of Statistics*, *11*, 1267.

Hannum, R. C., Hollander, M., & Langberg, N. A. (1981). Distributional results for random functionals of a Dirichlet process. *Annals of Probability*, *9*, 665–670.

Hanson, T. E. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, *101*, 1548–1565.

Hanson, T. E. (2007). Polya trees and their use in reliability and survival analysis. In *Encyclopedia of statistics in quality and reliability* (pp. 1385–1390). New York: Wiley.

Hanson, T. E., & Johnson, W. O. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, *97*, 1020–1033.

Hanson, T. E., Branscum, A., & Gardner, I. (2008). Multivariate mixtures of Polya trees for modelling ROC data. *Statistical Modelling*, *8*, 81–96.

Hjort, N. L. (1985). *Bayesian nonparametric bootstrap confidence intervals*. (Technical Report No. 240). Stanford, CA: Stanford University, Department of Statistics.

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

Hjort, N. L., Homes, C., Müller, P., & Walker, S. G. (2010). *Cambridge series in statistical and probabilistic mathematics*. *Bayesian nonparametrics*. Cambridge: Cambridge University Press.

Hollander, M., & Korwar, R. M. (1976). Nonparametric empirical Bayes estimation of the probability that $X \leq Y$. *Communications in Statistics. Theory and Methods*, *5*(14), 1369–1383.

Hollander, M., & Korwar, R. M. (1982). Nonparametric Bayesian estimation of the horizontal distance between two populations. In *Nonparametric statistical inference I*. New York: North-Holland.

Ibrahim, J. L., Chen, M., & Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer.

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

Ishwaran, H., & Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, *87*, 371–390.

Ishwaran, H., & Zarepour, M. (2002). Exact and approximate sum representation for the Dirichlet process. *Canadian Journal of Statistics*, *30*(2), 269–283.

James, L. F. (2006). Poisson calculus for spatial neutral to the right processes. *Annals of Statistics*, *34*, 416–440.

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics—continuous multivariate distributions*. New York: Wiley.

Johnson, N. L., Kotz, S., & Balkrishnan, N. (1997). Multivariate Ewens distribution. In *Discrete multivariate distributions* (pp. 232–246). New York: Wiley.

Johnson, R. A., Susarla, V., & Van Ryzin, J. (1979). Bayesian non-parametric estimation for age-dependent branching processes. *Stochastic Processes and Their Applications*, *9*, 307–318.

Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival data. *Journal of the Royal Statistical Society. Series B. Methodological*, *40*, 214–221.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457–481.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Annals of Statistics*, *27*, 562–588.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, *21*, 59–78.

Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B. Methodological*, *75*, 1–22.

Kingman, J. F. C. (1993). *Poisson processes*. Oxford: Clarendon.

Korwar, R. M., & Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Probability*, *1*, 705–711.

Korwar, R. M., & Hollander, M. (1976). Empirical Bayes estimation of a distribution function. *Annals of Statistics*, *4*, 581–588.

Kraft, C. H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability*, *1*, 385–388.

Kraft, C. H., & van Eeden, C. (1964). Bayesian bioassay. *Annals of Mathematical Statistics*, *35*, 886–890.

Kuo, L. (1986a). A note on Bayes empirical Bayes estimation by means of Dirichlet processes. *Statistics & Probability Letters*, *4*, 145–150.

Kuo, L. (1986b). Computations of mixtures of Dirichlet processes. *SIAM Journal on Scientific and Statistical Computing*, *7*, 60–71.

Kuo, L. (1988). Linear Bayes estimators of the potency curve in bioassay. *Biometrika*, *75*, 91–96.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Annals of Statistics*, *20*, 1222–1235.

Lavine, M. (1994). More aspects of Polya trees for statistical modelling. *Annals of Statistics*, *22*, 1161–1176.

Lijoi, A., & Prünster, I. (2010). Models beyond the Dirichlet process. In N. L. Hjort et al. (Eds.), *Cambridge series in statistical and probabilistic mathematics. Bayesian nonparametrics* (pp. 80–136).

Lo, A. Y. (1981). Bayesian nonparametric statistical inference for shock models and wear processes. *Scandinavian Journal of Statistics*, *8*, 237–242.

Lo, A. Y. (1982). Bayesian nonparametric statistical inference for Poisson point processes. *Zeitschrift Für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *59*, 55–66.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates; I. Density estimates. *Annals of Statistics*, *12*, 351–357.

Lo, A. Y. (1986). Bayesian statistical inference for sampling a finite population. *Annals of Statistics*, *14*, 1226–1233.

Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap. *Annals of Statistics*, *15*(1), 360–375.

Lo, A. Y. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, *16*, 1684–1695.

Lo, A. Y. (1991). A characterization of the Dirichlet process. *Statistics & Probability Letters*, *12*, 185–187.

Lo, A. Y. (1993a). A Bayesian bootstrap for censored data. *Annals of Statistics*, *21*, 100–123.

Lo, A. Y. (1993b). A Bayesian method for weighted sampling. *Annals of Statistics*, *21*, 2138–2148.

MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In D. Dey, P. Müller, & D. Sinha (Eds.), *Practical nonparametric and semiparametric Bayesian statistics* (pp. 23–44).

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*. Alexandria: Am. Statist. Assoc.

MacEachern, S. N. (2000). *Dependent Dirichlet processes*. Unpublished manuscript, The Ohio State University, Department of Statistics.

Mauldin, R. D., Sudderth, W. D., & Williams, S. C. (1992). Polya trees and random distributions. *Annals of Statistics*, *20*, 1203–1221.

McCloskey, J. W. (1965). *A model for the distribution of individuals by species in an environment*. Unpublished Ph.D. thesis, Michigan State University.

Muliere, P., & Walker, S. (1997). A Bayesian non-parametric approach to survival analysis using Polya trees. *Scandinavian Journal of Statistics*, *24*, 331–340.

Neath, A. A. (2003). Polya tree distributions for statistical modeling of censored data. *Journal of Applied Mathematics & Decision Sciences*, *7*(3), 175–186.

Neath, A. A., & Bodden, K. (1997). Bayesian nonparametric confidence bounds for a distribution function. *Journal of Statistical Computation and Simulation*, *59*, 147–160.

Neath, A. A., & Samaniego, F. J. (1996). On Bayesian estimation of the multiple decrement function in the competing risks problem. *Statistics & Probability Letters*, *31*, 75–83.

Neath, A. A., & Samaniego, F. J. (1997). On Bayesian estimation of the multiple decrement function in the competing risks problem, II. *Statistics & Probability Letters*, *35*, 345–354.

Ongaro, A., & Cattaneo, C. (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics & Probability Letters*, *67*, 33–45.

Padgett, W. J., & Wei, L. J. (1981). A Bayesian nonparametric estimator of survival probability assuming increasing failure rate. *Communications in Statistics. Theory and Methods*, *10*(1), 49–63.

Patil, G. P., & Taillie, C. (1977). Diversity as a concept and its implications for random communities. *Bulletin of the International Statistical Institute*, *47*, 497–515.

Perman, M., Pitman, J., & Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, *92*, 21–39.

Peterson, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association*, *72*, 854–858.

Petrone, S. (1999). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, *26*, 373–393.

Phadia, E. G. (1971). *Minimax estimation of a cumulative distribution function*. (Technical Report 71-1). Columbus, OH: The Ohio State University, Division of Statistics.

Phadia, E. G. (1973). Minimax estimation of a cumulative distribution function. *Annals of Statistics*, *1*, 1149–1157.

Phadia, E. G. (1974). Best invariant confidence bands for a continuous cumulative distribution function. *Australian Journal of Statistics*, *16*(3), 148–152.

Phadia, E. G. (1980). A note on empirical Bayes estimation of a distribution function based on censored data. *Annals of Statistics*, *8*(1), 226–229.

Phadia, E. G. (2007). On bivariate tailfree processes. In *Proceedings of the 56th session of the International Statistical Institute*, Lisbon, Portugal. Electronic version.

Phadia, E. G., & Susarla, V. (1983). Nonparametric Bayesian estimation of a survival curve with dependent censoring mechanism. *Annals of the Institute of Statistical Mathematics*, *35*, 389–400.

Phadia, E. G., & Susarla, V. (1979). An empirical Bayes approach to two-sample problems with censored data. *Communications in Statistics. Theory and Methods*, *8*(13), 1327–1351.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, *102*, 145–158.

Pitman, J. (1996a). Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability*, *28*, 525–539.

Pitman, J. (1996b). Some developments of the Blackwell-MacQueen urn scheme. In T. S. Ferguson, L. S. Shapley & J. B. MacQueen (Eds.), *Statistics, probability and game theory. Papers in honor or David Blackwell* (pp. 245–267). Hayward: IMS.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, *25*, 855–900.

Pruiit, R. C. (1988). *An inconsistent Bayes estimate in bivariate survival curve analysis*. Preprint, University of Minnesota.

Ramsey, F. L. (1972). A Bayesian approach to bioassay. *Biometrics*, *28*, 841–858.

Randles, R. H., & Wolf, D. A. (1979). *Introduction to the theory of nonparametric statistics*. New York: Wiley.

Reich, B. J., & Fuentes, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Annals of Applied Statistics*, *1*, 240–264.

Rodriguez, A., Dunson, D. B., & Gelfand, A. E. (2010). Latent stick-breaking processes. *Journal of the American Statistical Association*, *105*, 647–659.

Salinas-Torres, V. H., Pereira, C. A. B., & Tiwari, R. C. (2002). Bayesian nonparametric estimation in a series system or a competing-risks model. *Journal of Nonparametric Statistics*, *14*, 449–458.

Samaniego, F. J., & Whitaker, L. R. (1988). On estimating population characteristics from record-breaking observations. II. Nonparametric results. *Naval Research Logistics*, *35*, 221–236.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statistica Sinica*, *2*, 639–650.

Sethuraman, J., & Tiwari, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In S. Gupta & J. Berger (Eds.), *Statistical decision theory and related topics III* (Vol. 1, pp. 305–315).

Sinha, D. (1997). Time-discrete beta-process model for interval-censored survival data. *Canadian Journal of Statistics*, *25*, 445–456.

Sinha, D. (1998). Posterior likelihood methods for multivariate survival data. *Biometrics*, *54*, 1463–1474.

Steck, G. P. (1971). Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distributions. *Annals of Mathematical Statistics*, *42*, 1–11.

Susarla, V., & Phadia, E. G. (1976). Empirical Bayes testing of a distribution function with Dirichlet process priors. *Communications in Statistics. Theory and Methods*, *5*(5), 455–469.

Susarla, V., & Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, *71*, 897–902.

Susarla, V., & Van Ryzin, J. (1978a). Empirical Bayes estimation of a distribution (survival) function from right-censored observations. *Annals of Statistics*, *6*, 740–754.

Susarla, V., & Van Ryzin, J. (1978b). Large sample theory for a Bayesian nonparametric survival curve estimator based on censored samples. *Annals of Statistics*, *6*, 755–768.

Susarla, V., & Van Ryzin, J. (1980). Addendum to "Large sample theory for a Bayesian nonparametric survival curve estimator based on censored samples". *Annals of Statistics*, *8*, 693.

Teh, Y. W., & Gorur, D. (2010). Indian buffet processes with power-law behavior. In *Advances in neural information processing systems* (Vol. 22).

Teh, Y. W., & Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In N. L. Hjort et al. (Eds.), *Cambridge series in statistical and probabilistic mathematics*.

Bayesian nonparametrics.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004). Hierarchical Dirichlet processes. In *Advances in neural information processing systems* (Vol. 17). Cambridge: MIT Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*, 1566–1581.

Teh, Y. W., Gorur, D., & Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In M. Meila & X. Shen (Eds.), *Proceedings of the international conference on artificial intelligence and statistics* (Vol. 11, pp. 556–563). Brookline: Microtome Publishing.

Thibaux, R., & Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In M. Meila & X. Shen (Eds.), *Proceedings of the international conference on artificial intelligence and statistics* (Vol. 11, pp. 564–571). Brookline: Microtome Publishing.

Tiwari, R. C. (1981). *A mathematical study of the Dirichlet process*. Ph.D. dissertation, Florida State University, Department of Statistics.

Tiwari, R. C. (1988). Convergence of the Dirichlet invariant measures and the limits of Bayes estimates. *Communications in Statistics. Theory and Methods*, *17*(2), 375–393.

Tiwari, R. C., & Lahiri, P. (1989). On robust Bayes and empirical Bayes estimation of means and variances from stratified samples. *Communications in Statistics. Theory and Methods*, *18*(3), 913–926.

Tiwari, R. C., & Zalkikar, J. N. (1985). Empirical Bayes estimation of functionals of unknown probability measures. *Communications in Statistics. Theory and Methods*, *14*, 2963–2996.

Tiwari, R. C., & Zalkikar, J. N. (1991a). Empirical Bayes estimate of certain estimable parameters of degree two. *Calcutta Statistical Association Bulletin*, *34*, 179–188.

Tiwari, R. C., & Zalkikar, J. N. (1991b). Bayesian inference of survival curve from record-breaking observations: estimation and asymptotic results. *Naval Research Logistics*, *38*, 599–609.

Tiwari, R. C., & Zalkikar, J. N. (1993). Nonparametric Bayesian estimation of survival function under random left truncation. *Journal of Statistical Planning and Inference*, *35*, 31–45.

Tiwari, R. C., Jammalamadaka, S. R., & Zalkikar, J. N. (1988). Bayes and empirical Bayes estimation of survival function under progressive censoring. *Communications in Statistics. Theory and Methods*, *17*(10), 3591–3606.

Tsai, W. Y. (1986). Estimation of survival curves from dependent censorship models via a generalized self-consistent property with nonparametric Bayesian estimation application. *Annals of Statistics*, *14*, 238–249.

Walker, S. G., & Damien, P. (1998). A full Bayesian nonparametric analysis involving a neutral to the right process. *Scandinavian Journal of Statistics*, *25*, 669–680.

Walker, S. G., & Mallick, B. K. (1997a). A note on the scale parameter of the Dirichlet process. *Canadian Journal of Statistics*, *25*, 473–479.

Walker, S. G., & Mallick, B. K. (1997b). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *59*, 845–860.

Walker, S. G., & Mallick, B. K. (1999). Semiparametric accelerated life time models. *Biometrics*, *55*, 477–483.

Walker, S. G., & Muliere, P. (1997a). Beta-Stacy processes and a generalization of the Polya-urn scheme. *Annals of Statistics*, *25*(4), 1762–1780.

Walker, S. G., & Muliere, P. (1997b). A characterization of Polya tree distributions. *Statistics & Probability Letters*, *31*, 163–168.

Walker, S. G., & Muliere, P. (1999). A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate. *Annals of Statistics*, *27*(2), 589–599.

Walker, S. G., & Muliere, P. (2003). A bivariate Dirichlet process. *Statistics & Probability Letters*, *64*, 1–7.

Walker, S. G., Damien, P., Laud, P., & Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *61*, 485–527.

West, M. (1990). *Bayesian kernel density estimation*. (Discussion Paper 90-A02). Duke University, Institute of Statistics and Decision Sciences.

West, M. (1992). Modelling with mixtures (with discussion and rejoinder). In J. M. Bernardo et al. (Eds.), *Bayesian statistics* (Vol. 4). Oxford: Oxford University Press.

Wild, C. J., & Kalbfleisch, J. D. (1981). A note on a paper by Ferguson and Phadia. *Annals of Statistics*, *9*, 1061–1065.

Wolpart, R. L., & Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, *85*, 251–267.

Yamato, H. (1975). A Bayesian estimation of a measure of the difference between two continuous distributions. *Reports of the Faculty of Science, Kagoshima University. Mathematics, Physics and Chemistry*, *8*, 29–38.

Yamato, H. (1977a). Relations between limiting Bayes estimates and the U-statistics for estimable parameters of degree 2 and 3. *Communications in Statistics. Theory and Methods*, *6*, 55–66.

Yamato, H. (1977b). Relations between limiting Bayes estimates and the U-statistics for estimable parameters. *Journal of the Japan Statistical Society*, *7*, 57–66.

Yamato, H. (1984). Characteristic functions of means of distributions chosen from a Dirichlet process. *Annals of Probability*, *12*, 262–267.

Yamato, H. (1986). Bayes estimates of estimable parameters with a Dirichlet invariant process. *Communications in Statistics. Theory and Methods*, *15*(8), 2383–2390.

Yamato, H. (1987). Nonparametric Bayes estimates of estimable parameters with a Dirichlet invariant process and invariant U-statistics. *Communications in Statistics. Theory and Methods*, *16*(2), 525–543.

Yang, M., Hanson, T., & Christensen, R. (2008). Nonparametric Bayesian estimation of a bivariate density with interval censored data. *Computational Statistics & Data Analysis*, *52*(12), 5202–5214.

Zacks, S. (1971). *The theory of statistical inference*. New York: Wiley.

Zalkikar, J. N., Tiwari, R. C., & Jammalamadaka, S. R. (1986). Bayes and empirical Bayes estimation of the probability that $Z > X + Y$. *Communications in Statistics. Theory and Methods*, *15*(10), 3079–3101.

Zehnwirth, B. (1981). A note on the asymptotic optimality of the empirical Bayes distribution function. *Annals of Statistics*, *9*, 221–224.

Zehnwirth, B. (1985). Nonparametric linear Bayes estimation of survival curves from incomplete observations. *Communications in Statistics. Theory and Methods*, *14*(8), 1769–1778.

Author Index

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

Subject Index

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