Springer Series in Statistics

Eswar G. Phadia

Prior Processes and Their Applications

Nonparametric Bayesian Estimation

Second Edition

Springer

Springer Series in Statistics

More information about this series at http://www.springer.com/series/692

Eswar G. Phadia

Prior Processes and Their Applications

Nonparametric Bayesian Estimation

Second Edition

Springer

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

ISSN 0172-7397 ISSN 2197-568X (electronic)
Springer Series in Statistics
ISBN 978-3-319-32788-4 ISBN 978-3-319-32789-1 (eBook)
DOI 10.1007/978-3-319-32789-1

Library of Congress Control Number: 2016940383

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To my

Daughter SONIA

and

Granddaughter ALEXIS

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. In contrast, Doksum's approach which was more general than the Dirichlet process, but restricted to the real line, did not receive the same kind of attention since the posterior distributions were not easily computable nor the parameters meaningfully interpretable. Ferguson's 1974 (*Annals of Statistics)* paper gave a simple formulation for the posterior distribution of the neutral to the right process, and its application to the right censored data was detailed in Ferguson and Phadia (1979). In fact, it was pointed out in this paper that the neutral to the right process is equally convenient to handle right censored data as is Dirichlet process for uncensored data and offers more flexibility. These papers revealed the advantage of using independent increment processes, and their concrete application in the reliability theory saw the development of gamma process (Kalbfleisch 1978), extended gamma process (Dykstra and Laud 1981), and beta process (Hjort 1990), as well as beta-Stacy process (Walker and Muliere 1997a,b). These processes lead to a class of neutral to the right type processes.

Thus it could rightly be said that, prior to 1974, the subject of nonparametric Bayesian inference did not exist. The above two papers laid the foundation of this branch of statistics. Following the publication of Ferguson's 1973 paper, there was a tremendous surge of activity in developing nonparametric Bayesian procedures to handle many inferential problems. During the decades of the 1970s and 1980s, hundreds of papers were published on this topic. 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. 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, e.g., the recent paper by Chung and Dunson (2011), 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 where the audience can get exposed to the theory and applications of this useful subject so that they can apply them in practice. This is the prime motivator for undertaking the present task.

The objective of this book is to present the material on the Dirichlet process, its properties, and its various applications, as well as other prior processes that have been discovered through the 1990s and their applications, in solving Bayesian inferential problems based on data that may possibly be right censored, sequential, or quantal response data. 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. They have roots in the Ferguson-Sethuraman countable 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 of their applications.

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 detail and algorithms are developed to simulate the posterior distributions so that the Bayesian analysis can proceed. These aspects are covered 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).

This material is an outgrowth of my lecture notes developed during the week-long lectures I gave at Zhongshan University in China in 2007 on this topic, followed by lectures at universities in India 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 as much as possible, so that the reader may find it easy to read the original publications.

The first part is devoted to introducing various prior processes, their formulation, and their properties. The Dirichlet process and its immediate generalizations are presented first. The neutral to the right processes and the processes with independent increments, which give rise to 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 to place greater weights, where it is deemed appropriate, by selecting suitable partitions in developing the prior. 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 their origin in the Ferguson-Sethuraman infinite sum representation and the manner in which the weights are constructed. They are collectively called here as *Ferguson-Sethuraman processes*.

The second part contains various inferential 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 part is devoted to presenting inferential procedures based on censored data. Heavy emphasis is given to the estimation of the survival function since it plays an important role in the survival data analysis. This is followed by other examples which include estimation procedures in certain stochastic process models.

Ferguson's seminal paper, and others that followed, has opened up a dormant area of nonparametric Bayesian inference. During the last four decades, a considerable attention has been given to this area, and great stride is made in solving many nonparametric problems and extending some usual approaches (see Müller and Quintana 2004). For example, in problems where the observations are subjected to random error, traditionally the errors are assumed to be distributed as normal with mean zero. Now it is possible to assume them to be having an unknown distribution whose prior is concentrated around the normal distribution or symmetric distributions with mean zero and carry out the analysis. Moreover, in many applications when the prior information tends to nil, the estimators reduce to the usual

maximum likelihood estimators—a desirable property. Obviously, it is impossible to include all these methods and applications in this manuscript. However, a long list of references is included for the reader to explore relevant areas of interest further.

Since this book discusses various prior processes, and their properties and inferential procedures in solving problems encountered in practice and limits deeper technical details, it is ideal to serve as a comprehensive introduction to the subject of nonparametric Bayesian inference. 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 accessible book. 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.

Preface to the Revision

Following the publication of the book in 2013, I have noticed that there has been continued and intensified interest in applying nonparametric Bayesian methods in the analysis of statistical data. Therefore, it is important that I should update the book to reflect the current interest. This is the main rationale for this revision. I have not only supplemented but expanded the earlier edition with additional material which would make the book "richer" in the content. As a consequence, I have reorganized the topics of the first part into cohesive but separate chapters. The second and third parts of the earlier edition (new Chaps. 6 and 7) remain unchanged as the applications mentioned there were obtained mostly in closed form and have limited applicability in the present environment of dealing with large and complex data. Highlights of the improvement in the revised edition are as follows.

The Dirichlet process and its variants are grouped together in Chap. 2. Starting in 2006, there has been growing interest in developing hierarchical and mixture models. Accordingly, a new section is added to describe them in more detail. Implementation of these models in carrying out full Bayesian analysis requires the knowledge of posterior distributions. Unfortunately, they are not usually in closed form but are often complicated and intractable—a major hurdle. This makes it necessary to generate them via simulation for which computational procedures such as Gibbs sampler, blocked Gibbs sampler, and slice and retrospective sampling are developed in the literature. These methods are described here and steps of relevant algorithms provided by the authors are included while discussing specific models.

A major development that occurred during the last decade was the exploitation of Sethuraman's representation of a Dirichlet process in modeling data that included covariates and spatial data, time series data, dependent groups data, etc., and gave rise to what is known as dependent (Dirichlet) processes. To reflect this development and continued interest, the material of the earlier edition has been expanded to include several new processes, thus forming a separate chapter—Chap. 3, under the heading of Ferguson-Sethuraman processes. This chapter not only includes

dependent processes but also one- and two-parameter Poisson-Dirichlet processes and a species sampling model.

As mentioned before, the basic processes developed earlier such as neutral to the right, gamma, beta, and beta-Stacy were essentially based on processes with independent increments and their associated Levy measures. Therefore, it made sense to present them cohesively under a single chapter, Chap. 4. The Chinese restaurant process, Indian buffet process, and stable and kernel beta processes also find place in this chapter. Since a random probability measure may be viewed as a completely random measure, which in turn can be constructed via the Poisson process with a specific Levy measure as its mean measure, some fundamental definitions and theorems related to them are also included for the sake of ready reference. The material of tailfree and Polya tree processes forms Chap. 5.

Throughout this revision, I have added additional explanations whenever warranted, including outlines of proofs of major theorems and derivations of basic processes such as the Dirichlet process, and beta and beta-Stacy processes and their variants, as well as of processes that are popular in other areas, all for better understanding the mechanism behind them. Also some further generalization of these processes have been included. In addition, scores of new references have been added to the list of references making it easy for interested readers to explore further. While this book focuses on the fundamentals of nonparametric Bayesian approach, a recently published book *Bayesian Nonparametric Data Analysis* by Mull et al. (2015, Springer), is a good source of Bayesian treatment in modeling and data analysis and could serve as a complement to the present volume.

I sincerely believe that this expanded version would better serve the readers interested in this area of statistics.

Acknowledgment

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 prepublication copy of Ferguson's seminal 1973 paper which led to my doctoral dissertation at 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 Zhongshan University in China at the behest of Qiqing Yu of Binghamton University. I thank him and the Zhongshan 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 since then. My sincere thanks go to anonymous reviewers whose comments and generous suggestions improved the manuscript. I must have exchanged scores of emails and had countless conversations with Dr. Eva Hiripi, Editor of Springer, during the last four years. Her patience, understanding, and helpful suggestions were instrumental in shaping the final products of the first and second editions, as well as her decision to publish it in Springer Statistics Series. My heartfelt thanks go to her. The production staff at Springer including Ulrike Stricker-Komba and Mahalakshmi Rajendran at SPi Technologies India Private Ltd including Edita Baronaite did a fantastic job in detecting missing references and producing the final product. They deserve my thanks.

Since my retirement from WPU, the Department of Statistics at Wharton School, University of Pennsylvania has been kind enough to host me as visiting scholar to pursue the revision of the first edition. I am very grateful to the faculty and staff of the department, especially Ed George and Mark Low, for extending their support, courtesy, and cooperation which enabled me to complete the revision successfully. I offer my sincere thanks to all of them. I also thank the two anonymous reviewers for their very complimentary reviews of the revised edition.

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 thank my wife and companion, Jyotsna my daughter, Sonia, for their support and my granddaughter, Alexis, who, at her tender age, provided me happiness and stimulus to keep working on the revision in spite of my retirement.

Wayne, NJ, USA Eswar G. Phadia
July 2016

Contents

Chapter 1
Prior Processes: An Overview

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 and discuss circumstances in which they are appropriate to use and their relative merits and shortcomings in solving inferential problems. In subsequent sections we provide more details on each of them and state their properties. To preserve the historical perspective, they are mostly organized in the order of their discovery and development. The last two chapters contain various applications based on censored and uncensored data.

In the Bayesian approach, the unknown distribution function from which the sample arises is itself considered as a parameter. Thus, we need to construct prior distributions on the space of all distribution functions, to be denoted by $\mathcal{F}(\chi)$, defined on a sample space χ, or on all probability measures, Π defined on certain probability space, $(\mathfrak{X}, \mathcal{A})$, where $\mathcal{A}$ is σ-field of subsets of $\mathfrak{X}$. To be more precise, let X be a random variable defined on some probability space $(\Omega, \sigma(\Omega), Q)$ taking values in $(\mathfrak{X}, \mathcal{A})$, and let $\mathcal{F}(\chi)$ denote the space of all distribution functions defined on the sample space $(\mathfrak{X}, \mathcal{A})$.

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 function is characterized by treating p as random. In this case, a prior on $\mathcal{F}(\chi)$ may then be specified by simply assigning a prior distribution to p on Π, say uniform, $U(0, 1)$ or a beta distribution, $\mathrm{Be}(a, b)$ with parameters $a > 0$, and $b > 0$. A prior distribution on $\mathcal{F}(\chi)$ or Π will be denoted by $\mathfrak{P}$ whenever needed.

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1_1

As a second example, consider the multinomial experiment with the sample space, $\chi = \{1, 2, \ldots, k\}$, In this case, $\mathcal{F}(\chi)$ is the space of all distribution functions corresponding to a $(k-1)$-dimensional probability simplex $S_k = \{(p_1, p_2, \ldots, p_k) : 0 \leq p_i \leq 1, \sum_{i=1}^{k} p_i = 1\}$ of probabilities. Then a prior distribution $\mathcal{P}$ can be specified on $\mathcal{F}(\chi)$ by defining a measure on S_k which yields the joint distribution of $(p_1, p_2, \ldots, p_k)$, say, the Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, where $\alpha_i \geq 0$ for $i = 1, 2, \ldots, k$.

These are examples of finite dimensional priors. Our concern now is to extend these formulations to infinite dimension. In such situations the prior is a stochastic process with parameter space as σ-algebra of subsets of the underlying space.

While the distribution function F 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, $P(a, b] = F(b) - F(a)$. However, many of the applications are given in terms of the distribution function or its functional. The advantage of considering P is then it is easy to talk about arbitrary space which may include R^k instead of R^1 alone. The Dirichlet process (DP) is defined in this way on an arbitrary space. Ferguson derives his results for a random probability measure which is a special case of random measures introduced by Kingman (1967). Random measures are generated by the Poisson process. They provide a tool to treat priors in a unified approach as shown in Hjort et al. (2010). However, such an approach does not provide any insight into how the priors originated to start with.

Defining a prior for an unknown F on $\mathcal{F}$ or for a P on Π gives 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 lead 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 non-existent. Earlier attempts in defining such priors on $\mathcal{F}$ can be traced to Dubins and Freedman (1966) 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 chapter and present them formally in subsequent chapters.

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. The first one is based on specifying 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 variables 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 measure, also known as the *stick-breaking* construction. These are described here informally without going into the underlying technicalities.

The first method introduced by Ferguson (1973) is to define a family of consistent finite dimensional distributions of probabilities of sets of a measurable partition of a set on an arbitrary space, and then appealing to the Kolmogorov's extension theorem. 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 $P \sim \mathcal{D}(\alpha)$ (In our presentation, as has been a common practice, 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 treated the parameter α itself as random, indexed by u, u having a certain distribution H and proposed the *mixture of Dirichlet processes*, i.e., $P \sim \int \mathcal{D}(\alpha_u)\, dH(u)$. The other by Dalal (1979a) who treated the measure α as invariant under a finite

group of transformations and proposed a *Dirichlet Invariant process* over a class of invariant distributions which included, symmetric distributions around a location ξ, or distributions having a median at 0.

The remarkable feature of the Dirichlet process (DP) is that it is defined on abstract spaces and serves as a "base" prior, and is the main source for various generalizations in many different directions. This makes it possible to generate new prior processes allowing not only a great deal of flexibility in modeling, but at the same time are tailored for different statistical problems (see Fig. 1.1). For example, by treating α itself as a random measure having certain prior distribution, say the DP, hierarchical models were proposed in Teh et al. (2006); 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; by writing $f(x) = \int K(x,u)dG(u)$ with a known kernel K, and taking $G \sim \mathcal{D}(\alpha)$, Lo (1984) was able to place priors on the space of density functions. Further examples based on countable representation of the DP are given ahead.

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 $-\infty < t_1 < t_2 < \cdots < t_m < \infty$. 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. This simple requirement provides a tremendous flexibility in generating priors. Since a distribution function can be reparametrized 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. Some examples are as follows.

Kalbfleisch (1978) defined a *gamma process* by assuming the increments to be distributed as the gamma distribution; Dykstra and Laud (1981) proposed an *extended gamma process*, by defining 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 to define a *beta process* which places a prior on the space of cumulative hazard functions, and via the above parametrization, on the CDFs as well; Thibaux and Jordan (2007) defined a *(Hierarchical) beta process* by modifying the Levy measure of the beta process; and Walker and Muliere (1997a) introduced the *beta-Stacy process* by assuming the increments to be distributed as beta-Stacy distribution. There are other related processes as well. They all belong to the family of Levy processes.

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 distribution. *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. Such priors are particularly appropriate when one wishes to model a random F (Walker et al. 1999) with fixed locations based on some prior guess of F, say, F_0.

As a fourth approach, Blackwell and MacQueen (1973) showed that a prior process can also be defined by constructing a sequence of exchangeable random variables via the Polya urn scheme and then appealing to a theorem of de Finetti. If $X_1, X_2, \ldots$ is a sequence of exchangeable random variables with a common distribution P, then for every n and sets $A_1, \ldots, A_n \in \mathcal{A}$,

$$\mathcal{P}(X_i \in A_i : i = 1, \ldots, n) = \int \prod_{i=1}^{n} P(A_i)\, Q(dP),$$

where Q is known as the de Finetti measure, and serves as a prior distribution of P. The prior processes (actually measures) discussed here are the different forms of Q. In particular, they showed that the Dirichlet process can also be defined in this way. This approach is especially suitable when one is interested in prediction problems, i.e., in deriving the predictive distribution of X_{n+1} given $X_1, \ldots, X_n$. However identification of Q is usually a problem.

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, $i = 1, 2, \ldots, k$. Draw a ball at random of color i and define the random variable X_1 so that $\mathcal{P}(X_1 = i) = \overline{\alpha}_i$, 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 $\mathcal{P}(X_2 = j \mid X_1 = i) = (\alpha_j + \delta_j)/(\sum_{i=1}^{k} \alpha_i + 1)$, where $\delta_j = 1$ if $j = i$, 0 otherwise. This is the 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. Blackwell and MacQueen generalize this scheme by taking a continuum of colors α. Then a theorem of de Finetti assures us 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. In this approach, besides exchangeability all that is needed essentially is the predictive probability rule to define a prior.

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 problematic. This approach was adopted by Mauldin et al. (1992) who used a generalized Polya urn scheme to generate sequences of exchangeable random variables and based upon them, defined a Polya tree process. Pitman (1996b) gives other examples.

It is interesting to note that the DP has representation under all of the above approaches and it is the only prior which can be obtained by any of the above approaches.

In addition to the above four methods, the countable mixture representation of a random probability measure has been found recently to be a useful tool in developing several new processes, some of which are variants of the DP suitable for handling specific applications. 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 representation,

$$P = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}, \tag{1.2.1}$$

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 $0 \leq p_i \leq 1$ and $\sum_{i=1}^{\infty} p_i = 1$. Ferguson's weights were constructed using normalized gamma variates. 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 giving a simple construction, the so-called *stick-breaking* construction, and the interest was renewed. His weights are constructed as follows:

$$p_1 = V_1 \text{ and } p_i = V_i \prod_{j=1}^{i-1} (1 - V_j), i = 1, 2, \ldots, \text{ and } V_i \overset{\text{iid}}{\sim} \text{Be}(1, \alpha(\mathfrak{X})). \tag{1.2.2}$$

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 were developed, which we call *Ferguson–Sethuraman* processes. Examples are:

If the infinite sum in (1.2.1) is truncated at a fixed or random $N < \infty$, it generates a class of *discrete distribution priors* studied by Ongaro and Cattaneo (2004);

by replacing the parameters $(1, \alpha(\mathfrak{X}))$ of the beta distribution with real numbers (a_i, b_i), $i = 1, 2, \ldots$ Ishwaran and James (2001) defined *stick-breaking priors;* by indexing ξ_i with a covariate $\mathbf{x} = (x_1, \ldots, x_k)$, denoted as $\xi_{i\mathbf{x}}$, MacEachern (1999) defined a class of *dependent DPs* which includes *spatial* and *time-varying* processes as well; by replacing the degenerate probability measure δ by a nondegenerate positive probability measure G, Dunson and Park (2008) introduced *kernel DPs*. The above stick-breaking construction as well as the prediction rule based on a generalized Polya urn scheme proposed by Blackwell and MacQueen (1973) has been found useful in the development of new processes, two of which are 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.

A further generalization is also possible. Recall that Ferguson (1973) defined the DP alternatively by taking a normalized gamma process. This suggests a natural generalization by defining a random distribution function via a normalized increasing additive process $Z(t)$ (or independent increment process) with $Z = \lim_{t\to\infty} Z(t) < \infty$. Regazzini et al. (2003) pursue this path. Note that Doksum (1974) used the reparametrization $F(t) = 1 - e^{-Y_t}$, with Y_t as an increasing additive process [Walker and Muliere (1997a), took Y_t to be a log-beta process].

A brief exposé of these major processes follows. Details are discussed in subsequent chapters organized by grouping together related processes.

A recently published chapter by Lijoi and Prünster (2010) provides a unified framework for several priors processes in terms of the concept of completely random measures studied by Kingman (1967), which is a generalization to abstract spaces of independent increment processes on the real line. They can be generated via the Poisson process Kingman (1993) by specifying the appropriate mean measure of the Poisson process. This will be further elaborated in Chap. 4. Lijoi and Prünster's formulation is elegant but essentially the same. 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 view 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.

1.3 Prior Processes

In this section we introduce major processes briefly.

Ferguson's Dirichlet process is an extension of the k-dimensional Dirichlet distribution to a process. It 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) define two types of conjugacy: 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 only the parameters are updated. Neutral to the right processes are 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 class 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 $P \sim \mathcal{D}(\alpha)$, 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., $P|\mathbf{X} \sim \mathcal{D}\left(\alpha + \sum_{i=1}^{n} \delta_{x_i}\right)$ (parametric conjugacy). Thus it is easy to compute the posterior distribution, by simply updating the parameter of the prior distribution. 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 function 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,b) 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 F_n. 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 by giving several illustrative examples. His initiative set the tone and created a surge in the activity. Numerous papers were published thereafter describing its utility in treating many of nonparametric problems from the Bayesian point of view. These applications include sequential estimation, empirical Bayes estimation, confidence bands, hypothesis testing, and survival data analysis, to name a few and presented in Chaps. 6 and 7. 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. Also in the predictive distribution of a future observation, the probabilities of selecting a new or duplicating a previously selected observation do not depend upon the number of distinct observations encountered thus far. However, to remedy this deficiency a two-parameter Poisson–Dirichlet process (Pitman and Yor 1997) is developed.

A major deficiency though is that its support is confined to discrete probability measures only. Nevertheless, 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. In fact, the Sethuraman representation has unleashed a flood of new processes

to model various types of data, as indicated later. Thus its 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 extensions, are proposed in the literature as mentioned above. They are outlined next.

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 denoted 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 $P \sim \mathcal{DGI}(\alpha)$, and $X_1, \ldots, X_n$ is a sample of size n from P, then the posterior distribution of P given $X_1, \ldots, X_n$ is $\mathcal{DGI}(\alpha + \sum_{i=1}^{n} \delta^g_{X_i})$, where $\delta^g_{X_i} = (1/k) \sum_{i=1}^{k} \delta_{gX_i}$. It is found to be useful in solving certain estimation problems regarding location.

In dealing with the estimation of dose–response curve or estimation based on the right censored data, if the Dirichlet process prior is assumed, it was found that the posterior distribution was not a Dirichlet process, but a mixture of Dirichlet processes. This lead 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 $P \sim \int \mathcal{D}(\alpha_u)\, dH(u)$. It has some attractive properties and is flexible enough to handle purely parametric or semiparametric models. This has lead to the development of mixture models. In fact, its applications in modeling high dimensional and complex data have exploded in recent years (Müller and Quintana 2004; 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, $P \sim \int_U D(\alpha_u) dH(u)$, then $P|\boldsymbol{\theta} \sim \int_U D(\alpha_u + \sum_{i=1}^{n} \delta_{\theta_i}) dH_{\boldsymbol{\theta}}(u)$, 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. 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 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, $-\infty < t_1 < t_2 < \cdots < t_m < \infty$, for $m = 1, 2, \ldots$, the successive 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 exist independent random variables $V_1, \ldots, V_m$ such that the distribution of the vector $(1 - F(t_1), 1 - F(t_2), \ldots, 1 - F(t_m))$ is same as the distribution of $(V_1, V_1V_2, \ldots, \prod_1^m V_i)$. Thus the prior can be described in terms of several quantities providing more flexibility. Furthermore the Dirichlet process defined on the real line is a neutral to the right process. Doksum proved the conjugacy property with respect to the data which may include right censored observations as well, 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 its own weaknesses. Its 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 the neutral to the right process. They alleviate the drawbacks, and at the same time, are more manageable, parameters are interpretable and they are conjugate with respect to the right censored data.

The neutral to the right process can also be viewed in terms of a process with independent nonnegative increments (Doksum 1974; Ferguson 1974) via the reparametrization $F(t) = 1 - e^{-Y_t}$, where Y_t is a process with independent nonnegative increments (also known as positive Levy process). The DP corresponds to one of these Y_t processes. Thus a prior on $\mathcal{F}$ can be placed by using such processes. This representation 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. 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 processes with independent nonnegative increments are extensively studied and they have been used successfully in developing priors with appropriate modification of the Lévy measure involved. They all belong to the family of Levy processes. 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 the right censored 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, $-\infty < t_1 < t_2 < \cdots < t_m < \infty$, let q_j denote the hazard contribution of the interval $[t_{j-1}, t_j)$, i.e., $q_j = \big(F(t_j) - F(t_{j-1})\big)/(1 - F(t_{j-1}))$. 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 $Y(t)$ 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 (1981) 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. Its 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 becomes 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 the F may not have a density, he defines a new general form of the cumulative hazard function H such that $F(t) = 1 - \Pi_0^t\{1 - dH(t)\}$, where Π 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 distribution. Since the beta distribution lacks the necessary convolution properties, he had to get around it by defining 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. When B is viewed as a measure of the beta process, it turns out to be the de Finetti measure of the Indian buffet process.

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. In fact 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 also emerges as a posterior distribution with respect to the right censored data when the prior is assumed to be a Dirichlet process. It has some additional pluses as well. Its parameters have reasonable interpretation; it is a neutral to the right process; and the posterior expectation of the survival function obtained in Susarla and Van Ryzin (1978a,b) turns out to be a special case.

The random probability measure associated with many of the above processes is completely random measures (Kingman 1967) on the real line. As the completely random measures can be constructed via the Poisson process Kingman (1993) with suitable mean measures, so are these processes. For example, the gamma process with parameter $c > 0$, and base measure G_0 is generated when the mean measure is given by $\nu(d\theta, dp) = cp^{-1}e^{-cp}dpG_0(d\theta)$, $p > 0$. Let $\{(\theta_i, p_i)\}$ denote the points obtained from the Poisson process with mean measure ν and define $\xi_i = p_i / \sum_{j=1}^{\infty} p_j$. Then $P = \sum_{i=1}^{\infty} \xi_i \delta_{\theta_i}$ is the Dirichlet process with parameters $\alpha = G_0(\Omega)$ and $F_0 = \alpha^{-1}G_0$. It should be noted that due to normalization, the Dirichlet process is not a completely random measure since for A_1, $A_2 \in \Omega$, $P(A_1)$ and $P(A_1)$ are not independent but negatively correlated. Beta process with parameter $c > 0$, and base measure B_0 is generated when the mean measure of Poisson process is given by $\nu(d\theta, dp) = cp^{-1}(1-p)^{c-1}dpB_0(d\theta)$, $0 < p < 1$.

The *tailfree* and *Polya tree* processes 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. 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 resulting expressions in closed form and the parameters involved are difficult to interpret adequately. 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 having a beta distribution. This way the expressions are manageable. He names the resulting process as a *Polya Tree process.* It is shown 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.

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*.

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

In addition to the core processes described above, the Ferguson–Sethuraman countable mixture representation of the DP, as alluded to in the previous section, proved to be an important tool in developing a large number of prior processes in order to address nonparametric Bayesian treatment of models involving different and complex types of data. Many of these are offshoots of the Dirichlet process but there are others as well. We describe them briefly here and more detailed treatment will be given in Chap. 3.

Recall that the Sethuraman's representation of the Dirichlet process with parameter α is $P = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}$ where δ_{ξ_i} denotes a discrete measure concentrated at ξ_i, ξ_i's are independent and identically distributed according to the distribution $\alpha/\alpha(\mathfrak{X})$; and p_i (known as random weights) are chosen independently of ξ_i as defined in (1.2.2). Based on this representation, possibilities for developing several new prior processes by varying the way the weights and locations are defined, seem natural.

By allowing the possibility of the infinite sum to be truncated at $N \leq \infty$, and assuming the distribution of V_i in the construction of p_i as $\text{Be}(a_i, b_i)$, $a_i \geq 0$, $b_i \geq 0$, instead of $\text{Be}(1, \alpha(\mathfrak{X}))$, Ishwaran and James (2001) define a family of priors called *stick-breaking priors*. Besides the DP, it includes several other priors as well. By truncating the sum to a positive integer N, which could be random as well, a class of *discrete prior processes* can be generated. Ongaro and Cattaneo (2004) follow this approach and point out that such processes lack conjugacy property. If $a_i = a$ and $b_i = b$, then we have a process known as *two-parameter beta process* (Ishwaran and Zarepour 2000). On the other hand, if $a_i = 1 - a$ and $b_i = b + ia$, with $0 \leq a < 1$ and $b > -a$, then it identifies the *2-parameter Poisson–Dirichlet* process described by Pitman and Yor (1997) (also known as *Pitman–Yor process*). The process itself is a two-parameter generalization of the *Poisson–Dirichlet* distribution 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 $b = \alpha(\chi)$, and when $b = 0$, we obtain a stable-law process.

The above generalization allowed truncation of the infinite sum and modification of the weights, but locations ξ_i's and the degenerate measure δ were untouched. By modifying them as well, several new processes have been proposed with their newer applications in the fields as diverse as machine learning, population genetics, and ecology. Therefore we enlarge the family, and since they all originate from the countable mixture representations of Ferguson and Sethuraman, they should rightly be called, as we do, *Ferguson–Sethuraman priors*, and reserve the phrase *stick-breaking* to indicate the process of construction of the weights. Thus, Ferguson–Sethuraman processes have a basic form of countable mixture of (mostly unit) masses placed at iid location (mostly) random variables such that the sum of the weights (need not be constructed as in (1.2.2) is equal to 1 and

$$P(\cdot) = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}(\cdot). \tag{1.3.1}$$

To accommodate covariates, the locations ξ_i and/or the weights (via V_i's) are modified to depend on vectors of covariates. 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 $P_{\mathbf{x}}(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_{j\mathbf{x}}}(\cdot)$. 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 $P_{\mathbf{x}}$ forms a class of random probability measures 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, which could also be made to depend on $\mathbf{x}$. When weights are constructed as in (1.2.2), this class of priors includes processes such as the *dependent Dirichlet* (MacEachern 1999), *ordered Dirichlet* (Griffin and Steel 2006), *kernel stick breaking* (Dunson and Park 2008), *and local Dirichlet processes* (Chung and Dunson 2011). 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. On the other hand, Rodriguez et al. (2008) replace each random location ξ in the above infinite sum representation with a random probability measure drawn from a Dirichlet process, thus creating a nested affect. They named the resulting process as *nested Dirichlet process*. Dunson and Park (2008) replaced the unit measure δ by a nondegenerate measure in developing their process called the *kernel based stick-breaking* process.

In defining the beta process mentioned earlier, Hjort's (1990) primary interest was in developing a prior distribution on the space of cumulative hazard functions and therefore the sample paths were defined on the real line. Thibaux and Jordan (2007) saw the need to define the sample paths of the beta process on more general spaces. So they modify the underlying Levy measure of the beta process and show that the resulting process serves as a prior distribution over a class of sparse binary matrices encountered in featural modeling. It is conjugate with respect to a Bernoulli process and is the de Finetti measure for the Indian buffet process. It

can also be constructed by the stick-breaking construction method. Teh and Gorur (2009) generalize the beta process by introducing a stability parameter thereby incorporating the power-law behavior and name the resulting process as *Stable-beta process*. Ren et al. (2011) define *kernel beta process* to model covariate data similar to the dependent DP.

Unfortunately, there are no explicit expressions in closed form for the posterior distributions in case of the above processes and one has to rely on simulations methods. For this purpose, simulation algorithms are developed and provided by the respective authors. The development on this line has seen a tremendous growth in modeling large and complex data in fields outside the mainstream statistics and scores of papers have been published in recent years. Two processes have especially caught the attention of practitioners in different fields, and a third to lesser extent.

The *Chinese restaurant process* (CRP) is a process of generating a sample from the Dirichlet process and is equivalent to the extended Polya urn scheme introduced by Blackwell–MacQueen (1973). It is related to the culinary interpretation where in a stream of N patrons enter a restaurant and are randomly seated on tables. It describes the marginal distribution of the Dirichlet process in terms of random partition of patrons determined by K tables they occupy. Samples from the Dirichlet process are probability measures and samples from the CRP are partitions. Teh et al. (2006) proposed a further generalization as *franchised CRP* which corresponds to a hierarchical Dirichlet process.

The *Indian Buffet process* (IBP) proposed by Griffiths and Ghahramani (2006) is essentially a process to define a prior distribution on the equivalence class of sparse binary matrices (entries of the matrix have binary responses, 0 or 1) consisting of a finite number of rows and an unlimited number of columns. Rows are interpreted as objects (patrons) and columns (tables) as potentially unlimited features. It has applications in dealing with featural models where the number of features may be unlimited and need not be known a priori. In contrast to the CRP, here the matrix may have entries of 1's in more than one column in each row. It is an iid mixture of Bernoulli processes with mixing measure the beta process. 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. Like the DP, the IBP also can be constructed by the stick-breaking construction. Further two and three parameter extensions of the process are proposed in the literature. A further generalization is proposed by Titsias (2008) where the binary matrices have been replaced by nonnegative integer valued matrices and a distribution called *infinite gamma-Poisson process* is defined as prior on the class of such equivalent matrices. In this model, the features are allowed to reoccur more than once.

Other processes, some of which are generalizations of the Dirichlet process and beta process are also mentioned briefly. They include, for example, finite dimensional DPs and multivariate DP. Various attempts to extend some of these processes fruitfully to the bivariate case have remained challenging and mostly unsuccessful so far. However in one case it is shown that it can be done. A bivariate

tailfree process (Phadia 2007; Paddock et al. 2003) is constructed on a unit square and presented in the last section.

Many of the above mentioned processes may be considered as special cases of a class of models proposed by Pitman (1996b) called *species sampling models*,

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot) + \left(1 - \sum_{j=1}^{\infty} p_j\right) Q(\cdot), \tag{1.3.2}$$

where Q is the probability measure corresponding to a continuous distribution G, $\xi_j \overset{\text{iid}}{\sim} G$, and weights p_j's are constrained by the condition, $\sum_{j=1}^{\infty} p_j \leq 1$. However, this model is not popular in statistics.

In summary, from the above description of various processes, it is clear that each process has its own merits as well as certain limitations. They have been developed, in some cases, to address specific needs. Nevertheless, Dirichlet process and its generalizations have emerged as most useful tools in carrying out nonparametric Bayesian analysis. 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, computational feasibility, and interpretation of parameters involved and the results obtained, play critical roles in the decision. Judging from the various applications that are published in the literature and presented in the applications chapters, it appears that the Dirichlet process and mixtures of Dirichlet processes have a substantial edge over the others not withstanding their limitations. However, this seems to be changing in the current trend towards dealing with *Big Data.*

Figure 1.1 depicts the connection among the various prior 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 focussed in the past on those applications where the expressions could be obtained in closed form and the obvious choice was the Dirichlet process. However in recent years a tremendous progress is made in developing computational methods, such as Gibbs sampler, importance sampling, slice and retrospective sampling, etc., for simulating the posterior distributions for implementation of Bayesian analysis which has made it possible to handle more complex models. And in view of the phenomenal increase in cheap computing power, the previous limitations have almost dissipated. In fact, the mixtures of Dirichlet processes 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 as indicated earlier, to implement full Bayesian analysis in treating covariate models, random effect models, hierarchical models, wavelet models, etc. The present trend has been to combine parametric

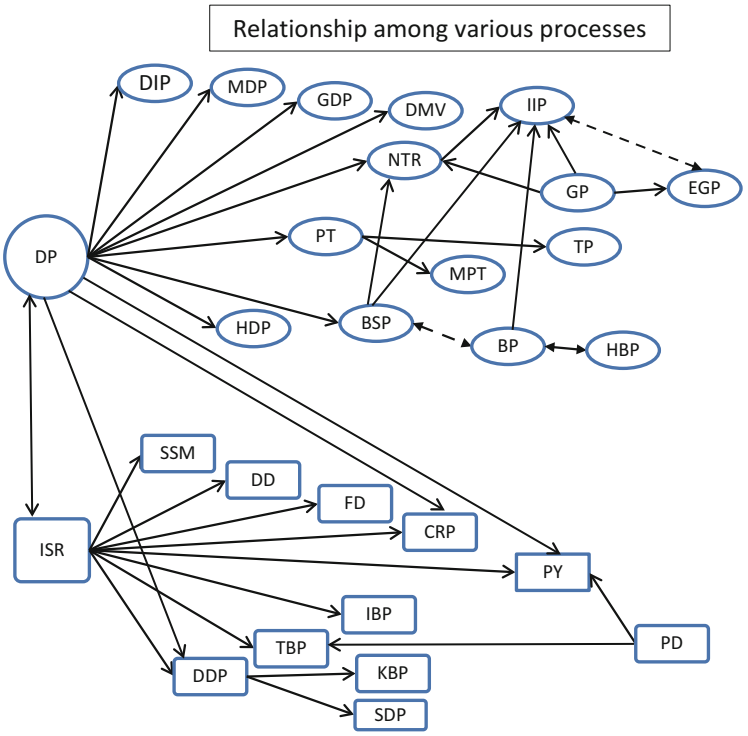

Fig. 1.1 An arrow A → B signifies either B generalizes A; or B originates from A; or A can be viewed as a particular case of B. Some relations need not be quite precise. A → B suggests B can be reached from A via a transformation. Processes in rectangles are identified as Ferguson-Sethuraman processes. *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, *IBP* Indian Buffet Process, *IIP* Independent Increments Process, *ISR* Infinite Sum Representation, *KBP* Kernel Based Process, *NTR* Neutral to the right Process, *PD* Poisson–Dirichlet Process, *PT* Polya Tree Process, *PY* Pitman–Yor (Two-parameter Poisson–Dirichlet Process), *SSM* Species sampling Model, *TBP* Two-parameter Beta Process, *TP* Tailfree Process

and semiparametric models in modeling such data. Books authored by Dey et al. (1998), Ibrahim et al. (2001), and Müller et al. (2015) contain numerous examples and applications, and are a good source to refer if one wants to explore further from the application angle. See also Favaro and Teh (2013) for an extended list of references.

Chapter 2
Dirichlet and Related Processes

2.1 Dirichlet Process

The Dirichlet process is an extension of the k-dimensional Dirichlet distribution to a stochastic process. It 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. Its success can be traced to its mathematical tractability, simple and attractive properties, and easy interpretation of its parameters. Dirichlet process (DP) and its offshoots, such as mixtures of Dirichlet processes (MDPs), hierarchical Dirichlet process (HDP), and dependent and spatial Dirichlet processes, are most important and widely used priors in modeling high dimensional and 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 $\alpha = MF_0$ for a random distribution function F (or a random probability measure [RPM] P) is a probability measure on the space of all distribution functions F (space of all RPMs) and is governed by two parameters: a baseline distribution function F_0 that defines the "location" or "center" or "prior guess" 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. It is concentrated on discrete probability distributions. Ferguson defines the Dirichlet process more broadly in terms of an RPM. We will follow his lead.

In this section we will present various features of the Dirichlet process: the original and alternative definitions; the close relationship between the parameter α and the random probability P; drawing a sample P; the posterior distribution of P given a random sample from it; procedures for generating samples from the same; numerous properties and characterization of the DP; and its various extensions.

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1_2

2.1.1 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 $[0, 1]$. That is, $\{P(A) : A \in \mathcal{A}\}$ is a stochastic process whose sample functions are probability measures on $(\mathfrak{X}, \mathcal{A})$. P being a probability is a measurable map from some probability space $(\Omega, \sigma(\Omega), \mu)$ to the space $(\Pi, \sigma(\Pi))$. Alternatively, it means that $P(\cdot, \cdot)$ is a measurable map from the product space $\Omega \times \mathcal{A}$ into $[0, 1]$ such that for every $\omega \in \Omega$, $P(\omega, \cdot)$ is a probability measure on $(\mathfrak{X}, \mathcal{A})$, and for every set $A \in \mathcal{A}$, $P(\cdot, A)$ is a measurable function (random variable) on $(\Omega, \sigma(\Omega))$ taking values in $[0, 1]$. In our treatment we will suppress reference to $\omega \in \Omega$ unless it is required for clarity. The distribution of P is a probability measure on $\left([0, 1]^{\mathcal{A}}, \sigma\left(\mathcal{B}^{\mathcal{A}}\right)\right)$ where $\sigma\left(\mathcal{B}^{\mathcal{A}}\right)$ denotes the σ-field generated by the field $\mathcal{B}^{\mathcal{A}}$ of Borel cylinder sets in $[0, 1]^{\mathcal{A}}$. 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 on the parameter space $(\Pi, \sigma(\Pi))$ 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 chapter on overview. Because of its simplicity and analytical tractability, the Dirichlet Process has been widely used despite its limitation that it gives positive probability to discrete distributions only. However it turns out to be an asset in certain areas of applications such as modeling grouped and covariate data and species sampling, as will be seen later in Chap. 3.

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

$$f(x_1, \ldots, x_{k-1}) = \frac{\Gamma(\gamma_1 + \cdots + \gamma_k)}{\Gamma(\gamma_1) \cdots \Gamma(\gamma_k)} \prod_{i=1}^{k-1} x_i^{\gamma_i - 1} \left(1 - \sum_{i=1}^{k-1} x_i\right)^{\gamma_k - 1}, \tag{2.1.1}$$

over the simplex: $S : \left\{(x_1, \ldots, x_{k-1}) : x_i \geq 0, i = 1, \ldots, k-1, \sum_{i=1}^{k-1} x_i \leq 1\right\}$ 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. Its tailfree property shows that the Dirichlet process is a tailfree process. Also, its neutral to the right property shows that the Dirichlet process is a neutral to the right process. These and many other properties of the Dirichlet distribution are discussed in great details by Basu and Tiwari (1982).

Before giving a formal definition, we must first fix the notion of a random probability measure. Since P is a stochastic process, it can be defined by specifying the joint distribution of the finite dimensional vector of random variables $(P(A_1), \ldots, P(A_m))$, for every positive integer m and every arbitrary sequence of measurable sets $A_1, \ldots, A_m$ belonging to $\mathcal{A}$, such that Kolmogorov consistency is satisfied. This would then imply that there exists a probability distribution $\boldsymbol{P}$ on $\left([0,1]^{\mathcal{A}}, \sigma\left(\mathcal{B}^{\mathcal{A}}\right)\right)$ yielding these finite dimensional distributions. Since such a sequence can be expressed in terms of mutually disjoint sets $B_1, \ldots, B_k$, with $\cup_{i=1}^{k} B_i = \mathfrak{X}$ (by taking intersections of the A_i and their complements), it is sufficient to define the joint distribution of $(P(B_1), \ldots, P(B_k))$ with $P(\varnothing) = 0$ which meets the consistency condition. Therefore it is sufficient to satisfy the following condition:

Condition C: *If* $\left(B'_1, \ldots, B'_{k'}\right)$ *and* $(B_1, \ldots, B_k)$ *are measurable partitions, and if* $\left(B'_1, \ldots, B'_{k'}\right)$ *is a refinement of* $(B_1, \ldots, B_k)$ *with* $B_1 = \cup_{j=1}^{r_1} B'_j, \ldots, B_k = \cup_{j=r_{k-1}+1}^{k'} B'_j$, *then the distribution of*

$$\left(\sum_{j=1}^{r_1} P\left(B'_j\right), \ldots, \sum_{j=r_{k-1}+1}^{k'} P\left(B'_j\right)\right) \tag{2.1.2}$$

as obtained from the joint distribution of $\left(P\left(B'_1\right), \ldots, P\left(B'_{k'}\right)\right)$, *is identical to the distribution of* $(P(B_1), \ldots, P(B_k))$.

This condition is shown to be sufficient to validate Kolmogorov consistency condition, and thus the existence of a measure on $\left([0,1]^{\mathcal{A}}, \sigma\left(\mathcal{B}^{\mathcal{A}}\right)\right)$ yielding the given finite dimensional distribution will be established. Now to define the DP which is a measure on $\left([0,1]^{\mathcal{A}}, \sigma\left(\mathcal{B}^{\mathcal{A}}\right)\right)$, all that is to be done is to specify the finite dimensional joint distribution. This is taken to be the Dirichlet distribution.

We say P is a *random probability measure* on $(\mathfrak{X}, \mathcal{A})$ (i.e., a measurable map from some probability space $(\Omega, \sigma(\Omega), Q)$ into $(\Pi, \sigma(\Pi))$), if the condition C is satisfied; 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. In this connection it is worth noting that Kingman (1967) defined a *completely random measure* (CRM) on an abstract measurable space $(\Psi, \sigma(\Psi))$ as a measure μ such that for any disjoint sets $A_1, A_2, \ldots \in \sigma(\Psi)$, random variables $\mu(A_1)$, $\mu(A_1), \ldots$ are mutually independent. More detailed reference is made later in Sect. 4.1.

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

Definition 2.1 (Ferguson) Let α be a non-null nonnegative 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 $(B_1, \ldots, B_k)$ of $\mathfrak{X}$, the distribution of the vector $(P(B_1), \ldots, P(B_k))$ is Dirichlet distribution, $D(\alpha(B_1), \ldots, \alpha(B_k))$.

In this definition, a parallel can be seen of the definition of a Poisson process (see Sect. 4.1) on abstract spaces (Kingman 1993) and a gamma process. For example, recall that a random measure μ on a measurable space $(\Psi, \sigma(\Psi))$ is a gamma process with parameter α if for any positive integer k and disjoint measurable sets $A_1, \ldots, A_k \in \sigma(\Psi)$, $\{\mu(A_i) : i = 1, .., k\}$ is a family of independent gamma random variables with mean $\alpha(A_i)$, $i = 1, \ldots, k$, respectively, and scale parameter unity.

By verifying the Kolmogorov consistency criterion, Ferguson 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 and with the property that the finite dimensional joint distribution of probabilities of sets $A_1, \ldots, A_k$ is a Dirichlet distribution. $\mathcal{D}(\alpha)$ may thus be considered as a prior distribution on the space Π of probability measures in the sense that each realization of the process yields a probability measure on $(\mathfrak{X}, \mathcal{A})$. Some immediate consequences of this definition are as follows:

(**a**) The Dirichlet process chooses a discrete RPM with probability one. This is true even when α is assumed to be continuous.
(**b**) The Dirichlet process is the only process 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.
(**c**) Let $P \sim \mathcal{D}(\alpha)$ 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 $(A, \mathcal{A} \cap A)$ 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, $D(\alpha(A_1), \ldots, \alpha(A_k))$.
(**d**) Let $\{\pi_m; m = 1, 2, \ldots\}$ be a nested tree of measurable partitions of $(R, \mathcal{B})$; that is $\pi_1, \pi_2, ..$ be a nested sequence of measurable partitions such that π_{m+1} is a refinement of π_m for each m and $\cup_0^\infty \pi_m$ generates $\mathcal{B}$. Then the Dirichlet process is tailfree with respect to every tree of partitions.
(**e**) Dirichlet process is neutral to the right with respect to every sequence of nested, measurable ordered partitions.

The parameter, α, can in fact be represented by two quantities. 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. The parameter M plays a significant role. Ferguson gave it the interpretation of prior sample size. However, some unsavory features have been pointed out in Walker et al. (1999). It controls the smoothness of F as well as the variability from F_0. The prior-to-posterior parameter update is $M \to M + n$ and $F_0 \to (MF_0 + nF_n)/(M + n)$ which is a linear combination of F_0 and F_n. When $M \to \infty$, F tends to the prior guess F_0 ignoring the sample information. On the other hand, if $M \to 0$, the prior provides no information. However, Sethuraman and Tiwari (1982) have shown that this interpretation is misleading. Actually in that case F degenerates to a single random point Y_0 selected according to F_0. Thus providing definite information that F is discrete. This fact about M 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 alternative representations of the definition of Dirichlet process have been proposed in the literature which are described next.

2.1.1.1 Alternative Representations of the Definition

Gamma Representation 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 normalized gamma variables [also true for any normalized random independent increments (Regazzini et al. 2003)]. Let $\mathcal{G}$ be a gamma process with intensity parameter $\gamma(\cdot) = MF_0(\cdot)$, i.e., $\gamma(A) \sim G(MF_0(A), 1)$, a gamma distribution with parameters $MF_0(A)$ and 1. Then $F(\cdot) = \gamma(\cdot)/\gamma(\mathfrak{X}) \sim \mathcal{D}(MF_0)$. It is defined in terms of a countable mixture $\sum_{j=1}^{\infty} P_j \delta_{\xi_j}$ of point masses at random points (of independent increments of gamma process) 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 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, $\mathcal{P}(J_1 \le x_1) = \exp\{N(x_1)\}$ for $x_1 > 0$, and for $j = 2, 3, \ldots$, $\mathcal{P}\left(J_j \le x_j | J_{j-1} = x_{j-1}, \ldots, J_1 = x_1\right) = \exp\left\{N\left(x_j\right) - N\left(x_{j-1}\right)\right\}$ 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 $Z_1 = \sum_{j=1}^{\infty} J_j$ converges with probability one and is a gamma variate with parameters $\alpha(\mathfrak{X})$ and 1, $G(\alpha(\mathfrak{X}), 1)$. Define $P_j = J_j / \sum_{i=1}^{\infty} J_i$, then $P_j \ge 0$ 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

Theorem 2.2 (Ferguson) *The RPM defined by*

$$P(A) = \sum_{j=1}^{\infty} P_j \delta_{\xi_j}(A),\ A \in \mathcal{A}, \tag{2.1.3}$$

where δ_x is the degenerate probability measure at x, is a Dirichlet process on $(\mathfrak{X}, \mathcal{A})$ with parameter α.

The key step of the proof involves in showing that for any k and any arbitrary partition $(B_1, \ldots, B_k)$ of $\mathfrak{X}$, the distribution of $(P(B_1), \ldots, P(B_k)) = \frac{1}{Z_1}\left(\sum_{j=1}^{\infty} J_j \delta_{\xi_j}(B_1), \ldots, \sum_{j=1}^{\infty} J_j \delta_{\xi_j}(B_1)\right)$ is $D(\alpha(B_1), \ldots, \alpha(B_k))$. Towards this end, he shows that for $i = 1, \ldots, k$, $\sum_{j=1}^{\infty} J_j \delta_{\xi_j}(B_i) \overset{\text{ind}}{\sim} G(\alpha(B_i), 1)$ by identifying them as independent increments of the gamma process [using a theorem from Ferguson and Klass (1972)], and that their sum is $Z_1 \sim G(\alpha(\mathfrak{X}), 1)$. This immediately yields the desired result using a property of the Dirichlet distribution.

This representation is counter intuitive. It shows that the DP is a countable mixture of point masses at randomly selected locations (selected iid $\overline{\alpha}$) and decreasing weights are the normalized gamma variates. Further it shows immediately that P is a discrete probability measure a.s., a fact proved by various authors. This has inspired the idea of defining a.s. discrete nonparametric priors by way of normalized completely random measures. Dirichlet process is a normalized completely random measure. In fact, without normalization one can deal with random measures themselves (as opposed to random *probability* measures) which also yield a.s. discrete random measure priors and are found to be useful in certain applications. This is so since the random discrete measures can be constructed using the Poisson processes with specific Levy measures as their intensity measures.

Levy Measure Representation Since the gamma process is a completely random measure (CRM) and CRM can be constructed by the Poisson process, Pitman (1996a) describes the above weights P_j's by ranking the points of a Poisson process. Let $J_1^* > J_2^* > \cdots$ be the points of a Poisson process on $(0, \infty)$ with mean measure $\nu(x) = \alpha(\mathfrak{X}) x^{-1} e^{-x} dx$. Then $P_j = J_j^* / \sum_{i=1}^{\infty} J_i^*$. P_j's are in descending order and the distribution of these weights is the Poisson–Dirichlet distribution, $PD(\alpha(\mathfrak{X}))$ developed by Kingman (1975) and discussed later in Sect. 3.4.1. Ishwaran and James (2001) describe them slightly differently. Let $\gamma_k = \varepsilon_1 + \cdots + \varepsilon_k$, where each ε_i is distributed as exponential distribution with parameter 1, that is, $\varepsilon_i \overset{\text{iid}}{\sim} \exp(1)$. 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 $x > 0$. 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. Then

$$P(\cdot) = \sum_{j=1}^{\infty} \frac{N^{-1}(\gamma_j)}{\sum_{j=1}^{\infty} N^{-1}(\gamma_j)} \delta_{\xi_j}(\cdot). \tag{2.1.4}$$

This form relies on the random weights constructed using infinitely divisible random variables.

As a further generalization, Regazzini et al. (2003) introduced a class of normalized random distribution functions with independent increments. The idea is to take the random probability distribution as a normalized increment process $F(t) = Z(t)/\overline{Z}$, where $\overline{Z} = \lim_{t \to \infty} Z(t) < \infty$ a.s. Ferguson took $Z(t)$ to be a gamma process. Nieto-Barajas et al. (2004) consider the case of normalized random distribution functions driven by an increasing additive process L, i.e., $Z(t) = \int K(t, x) dL(x)$ and provide conditions on K and L so that F is a random probability distribution function a.s.

Stick-Breaking Representation Ferguson's alternative definition was based on weights constructed using gamma variates. An extremely useful and popular constructive definition of the DP, known as *Stick-breaking representation,* is given by Sethuraman (1994) [see also Sethuraman and Tiwari (1982)] in which the weights are derived using a beta distribution with parameters 1 and $\alpha(\mathfrak{X})$, denoted

as Be $(1, \alpha(\mathfrak{X}))$. His representation of an RPM P having a Dirichlet prior $\mathcal{D}(\alpha)$ is

$$P(A) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(A),\ A \in \mathcal{A}, \tag{2.1.5}$$

where ξ_j's are as above,

$$p_1 = V_1,\ \text{and for } j \geq 2,\ p_j = V_j \Pi_{i=1}^{j-1}(1 - V_i) \text{ with } V_j \overset{\text{iid}}{\sim} \text{Be}\,(1, \alpha(\mathfrak{X}))\,, \tag{2.1.6}$$

and independent of $\xi_1, \xi_2, \ldots$. That is, the weights are generated by the so-called *stick-breaking (SB)* 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 jth 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 $p_j = V_j \prod_{l=1}^{j-1}(1 - V_l) = V_j\left(1 - q_j\right)$ where $q_j = \sum_{i=1}^{j} p_i$. This also shows that p_j's go to zero a.s. very fast. This process results in a sequence of weights $p_1, p_2, \ldots$ of Sethuraman representation. The weights will be denoted by SBW $(\alpha(\mathfrak{X}))$ and V 's are known as *stick-breaking ratios.* In contrast to Ferguson's weights, these weights need not be ordered. Ferguson's weights are equivalent to these weights rearranged in a decreasing order. That is, Ferguson weights are order statistics of these weights, $p_{(1)}, p_{(2)}, \ldots$. This construction provides a simple way to simulate F with $\mathcal{D}(\alpha)$ prior and also to compute the distribution of functionals of F such as $\int g dF = \sum p_i g(\xi_i)$ for a measurable function g on $\mathfrak{X}$. As will be seen later, the stick-breaking construction has been found useful in representing other priors as well.

The SB construction generates a random discrete distribution $p = (p_1, p_2, \ldots)$ and is of interest to many authors (see Pitman 1996a). The distribution of the vector p is known as *GEM* $(\alpha(\mathfrak{X}))$ distribution after Griffiths (1980), Engen (1978) and McCloskey (1965) who contributed to its development (Johnson et al. 1997, calls it as *generalized Engen–McCloskey* and credits Griffiths for its popularization). It is discussed later in Sect. 3.4 that if a sequence $\left\{p_i^*\right\}$ with $p_1^* \geq p_2^* \geq \cdots$ is defined by ranking the sequence $\{p_i\}$ with GEM$(\alpha(\mathfrak{X}))$ distribution, then the sequence $\left\{p_i^*\right\}$ has a Poisson–Dirichlet distribution with parameter $\alpha(\mathfrak{X})$.

Like the Ferguson's alternative definition, Sethuraman representation also reveals that the random probability P is a countable mixture of point masses at random locations $\xi_1, \xi_2, \ldots$, chosen according to the distribution $\overline{\alpha}(\cdot)$. However, the mixing probabilities $p_1, p_2, \ldots$ are different and 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 $(1, \alpha(\mathfrak{X}))$. 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 extensions/generalization of the Dirichlet process.

Theorem 2.3 (Sethuraman) *P as defined above in (2.1.5) is a Dirichlet process with parameter α.*

Its proof is interesting. We need to show first that a measure $\mathbf{P}$ exists, and second that its finite dimensional distributions are Dirichlet distributions.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots)$ be two sequences of iid random variables, independent of each other. Let θ_i be distributed as beta distribution, Be $(1, \alpha(\mathcal{X}))$, and Y_i distributed as $\overline{\alpha}(\cdot) = \alpha(\cdot)/\alpha(\mathcal{X}), i = 1, 2, \ldots$. Let $p_1 = \theta_1$ and $p_i = \theta_i \prod_{j=1}^{i-1} (1 - \theta_j)$. Now define a probability function P as

$$P_{\boldsymbol{\theta},\mathbf{Y}}(A) = P(A) = \sum_{i=1}^{\infty} p_i \delta_{Y_i}(A), \quad A \in \mathcal{A}.$$

Clearly by definition, P is random, discrete distribution with probability one, and is a stochastic process $\{P_A : A \in \mathcal{A}\}$. P is a measurable map from $(\mathfrak{X}, \mathcal{A})$ into $(\Pi, \sigma(\Pi))$.

We need to show that the distribution of P is a Dirichlet measure, $\mathcal{D}_\alpha$ defined on $(\Pi, \sigma(\Pi))$. Let $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \ldots)$ and $\mathbf{Y}^* = (Y_1^*, Y_2^*, \ldots)$ be two sequences such that $\theta_n^* = \theta_{n+1}$ and $Y_n^* = Y_{n+1}$ for $n = 1, 2, \ldots$. It is easily seen that P satisfies the equality

$$P_{\boldsymbol{\theta},\mathbf{Y}}(A) = \theta_1 \delta_{Y_1}(A) + (1 - \theta_1) P_{\boldsymbol{\theta}^*,\mathbf{Y}^*}(A), \quad A \in \mathcal{A}.$$

Note that $(\boldsymbol{\theta}^*, \mathbf{Y}^*)$ has the same distribution as $(\boldsymbol{\theta}, \mathbf{Y})$ and is independent of (θ_1, Y_1). Thus P satisfies the following distributional equation:

$$P \stackrel{\text{st}}{=} \theta_1 \delta_{Y_1} + (1 - \theta_1) P. \tag{2.1.7}$$

From the theory of distributional equation it follows that there exists a distribution $\mathbf{P}$ of P on $(\Pi, \sigma(\Pi))$ satisfying the above equation. His Lemma 3.3 [which is: *Suppose W, U be random variables with $W \in [-1, 1]$ and U taking values in linear space, V also a random variable in the same linear space as U and independent of (W, U), and satisfying distributional equation $V = U + WV$. Then, there is only one distribution for V*] shows that the distribution is unique. Thus the existence of its distribution is ascertained.

Now the task is to prove that the distribution is the Dirichlet measure $\mathcal{D}_\alpha$. A measure on $(\Pi, \sigma(\Pi))$ is Dirichlet if its finite dimensional distributions are Dirichlet distributions. Therefore consider the partition $(B_1, \ldots, B_k)$ of $\mathfrak{X}$ and the corresponding vector of random variables $\bar{P} = (P(B_1), \ldots, P(B_k))$. Let $\bar{D} = (\delta_{Y_1}(B_1), \ldots, \delta_{Y_1}(B_k))$ and $\mathbf{e}_j$ be a k-dimensional vector with 1 at the jth place and zero elsewhere. Then it follows that $\mathcal{P}(\bar{D} = \mathbf{e}_j) = P(Y_1 \in B_j) = \overline{\alpha}(B_j)$, $j = 1, \ldots, k$. From (2.1.7) we see that $\bar{P}$ satisfies the distributional equation

$$\bar{P} \stackrel{\text{st}}{=} \theta_1 \bar{D} + (1 - \theta_1) \bar{P},$$

where $\bar{D}$ is independent of θ_1 and the k-dimensional random vector $\bar{P}$ is independent of $(\theta_1, \bar{D})$. Now let us assume that the distribution of $\bar{P}$ on the right-hand side is the Dirichlet distribution $D(\alpha(B_1), \ldots, \alpha(B_k))$ and we know the distribution of $\bar{D}$ is Dirichlet $D(\mathbf{e}_j)$. Thus given $\bar{D} = \mathbf{e}_j$, the conditional distribution of $\theta_1 \bar{D} + (1 - \theta_1)\bar{P}$ by a property of the Dirichlet distribution [which says: *Let U and V be two independent k-*dimensional *vectors having Dirichlet distributions $D(\overline{\gamma})$ and $D(\overline{\delta})$, respectively, where $\overline{\gamma} = (\gamma_1, \ldots, \gamma_k)$, $\overline{\delta} = (\delta_1, \ldots, \delta_k)$ and independent of them, let $W \sim \text{Be}(\gamma, \delta)$, where $\gamma = \sum \gamma_i$ and $\delta = \sum \delta_i$. Then $WU + (1 - W)V \sim D\left(\overline{\gamma} + \overline{\delta}\right)$*] is $D\left((\alpha(B_1), \ldots, \alpha(B_k)) + \mathbf{e}_j\right)$. Multiplying by the marginal and using another property [which says: *If $\beta_j = \gamma_j/\gamma, j = 1, \ldots, k$, then $\sum \beta_j D\left(\overline{\gamma} + \mathbf{e}_j\right) = D(\overline{\gamma})$*] of the Dirichlet distribution we get $D((\alpha(B_1), \ldots, \alpha(B_k))$ verifying that this k-dimensional distribution satisfies the functional equation and that the solution is unique by his Lemma 3.3, thus completing the proof.

The weights in the above representation are related to the Poisson process (Ishwaran and Zarepour 2003) as follows. Noting that $M = \alpha(\mathcal{X})$,

$$P(\cdot) \stackrel{D}{=} \sum_{j=1}^{\infty} (e^{-\gamma_{j-1}/M} - e^{-\gamma_j/M}) \delta_{\xi_j}(\cdot), \tag{2.1.8}$$

with $\gamma_0 = 0, \gamma$'s as defined above. This can be seen by observing that $p_j = e^{-\gamma_{j-1}/M} - e^{-\gamma_j/M} = e^{-\varepsilon_1/M} \cdots e^{-\varepsilon_{j-1}/M}(1 - e^{-\varepsilon_j/M}) \stackrel{D}{=} \prod_{l=1}^{j-1}(1 - V_l) V_j$, since $e^{-\varepsilon_1/M} \sim \text{Be}(M, 1)$ and $1 - e^{-\varepsilon_1/M} \sim \text{Be}(1, M)$. Ishwaran and Zarepour (2003) give an interesting discussion of different representations of the weights.

A Limit Representation The DP can also be represented as a limit of finite dimensional Dirichlet distribution (Neal 2000; Ishwaran and Zarepour 2003). Let ξ_j's be iid $\mathfrak{X}$-valued random variables with common distribution $\overline{\alpha}(\cdot)$ and independent of ξ_j's, let $\mathbf{p} = (p_1, \ldots, p_N)$ be distributed as symmetric Dirichlet distribution with parameter $(\alpha(\mathfrak{X})/N, \ldots, \alpha(\mathfrak{X})/N)$, i.e., the vector $\mathbf{p}$ is assigned a symmetric Dirichlet prior. In view of a property of the Dirichlet distribution, p_i may also be defined as $p_i = Y_i / \sum_{j=1}^{N} Y_j$, where Y_i are iid $G(\alpha(\mathfrak{X})/N, 1)$. Define

$$\mathcal{D}_N(\cdot) = \sum_{j=1}^{N} p_j \delta_{\xi_j}(\cdot). \tag{2.1.9}$$

Ishwaran and Zarepour (2003) call $\mathcal{D}_N$ a *finite dimensional Dirichlet prior*. The authors formally prove that $\mathcal{D}_N$ converges to the DP with parameter α. Thus it can be used as an approximation to the DP. This construction is simple and a nice feature here is that the weights are exchangeable. The symmetric Dirichlet distribution has been previously used by Kingman (1975) in defining the Poisson–Dirichlet distribution (see Sect. 3.4.1), and Patil and Taillie (1977) in connection with the size-biased permutation of a vector of probabilities, among others. The

connection between Ferguson and Sethuraman weights and their relation to size-biased permutation are further explored in Sects. 3.4 and 3.5.

Polya Sequence Representation Blackwell and MacQueen (1973) also provide an alternative characterization of the DP. 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

$$\mathcal{P}(X_1 \in A) = \alpha(A)/\alpha(\mathfrak{X}) \text{ and}$$

$$\mathcal{P}(X_{n+1} \in A | X_1, \ldots, X_n) = \frac{\alpha_n(A)}{\alpha_n(\mathfrak{X})} = \frac{\alpha(A) + \sum_{i=1}^n \delta_{x_i}(A)}{\alpha(\mathfrak{X}) + n}, \tag{2.1.10}$$

where $\alpha_n = \alpha + \sum_{i=1}^n \delta_{x_i}$. A rule specifying the distribution of X_1 along with the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ for $n = 1, 2, \ldots$ is called a prediction rule for the sequence $X_1, X_2, \ldots$, and $\alpha_n(A)/\alpha_n(\mathfrak{X})$ as Polya urn distribution. The sequence $\{X_n : n \geq 1\}$, 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, its color is noted and is replaced by two balls of the same color. Given this prediction rule, they prove the following remarkable result:

Theorem 2.4 (Blackwell and MacQueen) *(a)The sequence $\alpha_n(\cdot)/\alpha_n(\mathfrak{X})$ converges with probability one as $n \to \infty$ to a limiting discrete measure P; (b) P is the Dirichlet process with parameter α; and (c) given P, $X_1, X_2, \ldots$ are independent with distribution P.*

The theorem shows that the sequence is exchangeable and that it is a sample from the DP. Furthermore, it gives us a way to sample from the DP as indicated later in the chapter. This result has received considerable interest in other areas that will be described later in Sect. 3.5. Thus the DP can be characterized in terms of its predictive rule. The predictive distribution in (2.1.10) can also be expressed as

$$\mathcal{P}(X_{n+1} \in \cdot | X_1, X_2, \ldots, X_n) = \sum_{i=1}^n \frac{1}{\alpha(\mathfrak{X}) + n} \delta_{x_i}(\cdot) + \frac{\alpha(\mathfrak{X})}{\alpha(\mathfrak{X}) + n} \overline{\alpha}(\cdot). \tag{2.1.11}$$

The interpretation of this expression is that X_{n+1} takes one of the previous n values with probability $1/(\alpha(\mathfrak{X}) + n)$ and with probability $\alpha(\mathfrak{X})/(\alpha(\mathfrak{X}) + n)$, a new value is picked randomly according to $\overline{\alpha}$.

The sequence $\{X_n : n \geq 1\}$ so constructed is exchangeable, by which we mean that for any $n \geq 1$, the joint distribution of the vector $(X_1, \ldots, X_n)$ is the same as that of any permutation of the same vector. Therefore, applying a de Finetti theorem it can be shown (Ghosh and Ramamoorthi 2003) that the mixing distribution in the theorem turns out to be the Dirichlet process. Basu and Tiwari (1982) have studied Blackwell–MacQueen definition in detail. Their paper clears up some measure theoretical details. Also see an interesting commentary by J. Sethuraman

("Commentary on a note on the Dirichlet process" in *Selected works of Debabrata Basu*, Ed: Anirban DasGupta, Springer, 2011).

The above expression enables us to construct an exchangeable sequence and once that is done, de Finetti theorem ensures the existence of an RPM providing the mixing distribution. Generating an exchangeable sequence via a prediction rule is discussed further in Sect. 3.5.

2.1.1.2 Samples from P

Ferguson defines what is meant by a sample from an RPM P, by providing the following definition:

Definition 2.5 (Ferguson) Let P be an RPM 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 $\mathfrak{X}$,

$$\mathcal{P}\{X_1 \in C_1, \ldots, X_n \in C_n | P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_n)\} = \prod_{j=1}^{n} P\left(C_j\right) \text{ a.s..}$$

This definition implies that given $P(C_1), \ldots, P(C_n)$, the events $\{X_1 \in C_1\}, \ldots, \{X_n \in C_n\}$ are independent of the rest of the process and are mutually independent, with $P\{X_j \in C_j | P(C_1), \ldots, P(C_n)\} = P\left(C_j\right)$ a.s. for $j = 1, \ldots, n$.

2.1.2 Properties

The Dirichlet process possesses certain interesting features. It is "rich" in the sense that it is flexible enough to incorporate any prior information or belief; it is closed in the sense that if the prior is a DP, so is the posterior, given a random sample; it has parameters which are easily interpretable; it is easy to evaluate expectation of simple functions with respect to the DP; and it is almost surely discrete. Its parameter α when expressed in terms of M and F_0 has interesting interpretations. Since $\mathcal{E}(P(\cdot)) = \alpha(\cdot)/\alpha(\mathfrak{X})$, one can choose the "shape" of α, F_0 to reflect one's prior guess at the shape of the distribution. As the posterior distribution is a linear combination of F_0 and F_n, the sample distribution function, the magnitude of M represents, in a sense, the degree of confidence in prior guess and treated as if it were a "prior sample" size. Independent of the shape, one can choose the magnitude of M to reflect his strength of conviction.

On the other hand, its discreteness seems to be inconsistent with the desired property of "richness" if one wants to place a prior on a class of continuous distributions. However, Ferguson has proved that the DP is rich in the sense that there is positive probability that a sample function will approximate as closely as desired with any fixed distribution which is absolutely continuous with respect to

the finite measure α, the parameter of the DP. As will be seen in the next chapter, the discreteness of the DP turns out to be an asset while carrying out Bayesian analysis of mixture and hierarchical models, in clustering problems where it is assumed that each observation may belong to only one distinct cluster, and in application of species sampling models.

Ferguson proved several important properties of the Dirichlet process and gave a few applications. Following this, various authors have proved additional properties. We state these properties without proofs.

There is a closed relationship between the parameter α and random probability P as the following properties reveal:

1. If $\alpha(A) = 0$, then $P(A) = 0$ a.s.; if $\alpha(A) > 0$, then $P(A) > 0$ a.s., and $\mathcal{E}(P(A)) = \alpha(A)/\alpha(\mathfrak{X})$.

2. If α is σ-additive, then so is P, i.e., for a fixed decreasing sequence of measurable sets $A_n \searrow \varnothing$, then $P(A_n) \to 0$ a.s. as $n \to \infty$. The converse is also true: If α is not σ-additive, then neither is P, a.s.

3. If Q is fixed probability measure on $(\mathfrak{X}, \mathcal{A})$ such that $Q \ll \alpha$, then, for any positive integer m and measurable sets $A_1, \ldots, A_m$, and $\epsilon > 0$,

$$\mathcal{P}\{|P(A_i) - Q(A_i)| < \epsilon \text{ for } i = 1, \ldots, m\} > 0.$$

It may seem from (**1**) that α and P are mutually absolutely continuous. But it is false since the null set outside of which the conclusion of (**1**) holds may depend upon the set A. The third property shows that the support of the Dirichlet process with parameter α contains the set of all probability measures absolutely continuous with respect to α, and that any measure Q not absolutely continuous with respect to α is not in the support of P.

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.

4. Let α_1 and α_2 be two nonatomic, non-null finite measures on $(\mathfrak{X}, \mathcal{A})$. If $\alpha_1 \neq \alpha_2$, then $\mathcal{D}(\alpha_1) \perp \mathcal{D}(\alpha_2)$.

5. Probabilities assigned to two disjoint sets are negatively correlated.

Ferguson demonstrated that various expectations of real valued functions Z's defined on $(\mathfrak{X}, \mathcal{A})$ can easily be derived.

6. If $\int |Z| \, d\alpha < \infty$, then $\int |Z| \, dP < \infty$ with probability one and $\mathcal{E} \int Z dP = \int Z d\mathcal{E}(P) = \int Z d\overline{\alpha}$.

7. 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 $\sigma_0^2 = \int Z_1^2 d\overline{\alpha} - \mu_1^2$.

8. 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

$$\mathcal{E}\,(\mu - \mu_0)^4 = \frac{6\mu_4 + 3M\sigma^4}{(M+1)(M+2)(M+3)},$$

where $\sigma^2 = \int (Z-\mu_0)^2 d\overline{\alpha}$, $\mu_3 = \int (Z-\mu_0)^3 d\overline{\alpha}$, and $\mu_4 = \int (Z-\mu_0)^4 d\overline{\alpha}$.

Some of these properties have been generalized in different directions. Let $g(x_1, \ldots, x_k)$ be a measurable real valued function defined on the k-fold product space $\left(\mathfrak{X}^k, \mathcal{A}^k\right)$ of $(\mathfrak{X}, \mathcal{A})$ and symmetric in $x_1, \ldots, x_k$. Assume that

$$\int_{\mathfrak{X}^m} \mid g(x_1, \ldots, x_1, x_2, \ldots, \; x_2, \ldots., x_m, \ldots, x_m) \mid d\overline{\alpha}(x_1) \cdots d\overline{\alpha}(x_m) < \infty, \tag{2.1.12}$$

for all possible combinations of arguments $(x_1, \ldots, x_1, x_2, \ldots, \; x_2, \ldots., x_m, \ldots, x_m)$ 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 (2.1.12) reduces to the simple condition

$$\int_{\mathfrak{X}^k} \mid g(x_1, \ldots, x_k) \mid d\overline{\alpha}(x_1) \cdots d\overline{\alpha}(x_k) < \infty \, . \tag{2.1.13}$$

An important property that has been used widely in solving some nonparametric Bayesian estimation problems is derived in Yamato (1977a,b, 1984) and Tiwari (1981, 1988).

9. Under the assumption (2.1.13),

$$\int_{\mathfrak{X}^k} \mid g(x_1, \ldots, x_k) \mid dP(x_1) \cdots dP(x_k) < \infty \quad \text{with probability one,} \tag{2.1.14}$$

and

$$\mathcal{E} \int_{\mathfrak{X}^k} g(x_1, \ldots, x_k)\, dP(x_1) \cdots dP(x_k) = \sum_{C(\sum i m_i = k)} \frac{k![\alpha\,(\mathfrak{X})]^{\Sigma m_i}}{\Pi_{i=1}^k [i^{m_i}\,(m_i)!]\alpha\,(\mathfrak{X})^{(k)}} \times \int_{\mathfrak{X}^{\Sigma m_i}} \Psi(\mathbf{x}) \Pi_{i=1}^k \Pi_{j=1}^{m_i} d\overline{\alpha}(x_{ij}), \tag{2.1.15}$$

where $\Psi(\mathbf{x}) = g(x_{11}, \ldots, x_{1m_1}, x_{21}, \ldots, \; x_{2m_2}, \ldots, x_{km_k}, \ldots, x_{km_k}), s^{(n)} = s\,(s+1) \cdots (s+n-1)$, and the summation $\sum_{C(\sum i m_i = k)}$ extends over all nonnegative integers $m_1, \ldots, m_k$ such that $\sum i m_i = k$. Taking g to be the indicator function, one can derive the marginal distribution of a sample $(x_1, \ldots, x_k)$ from P. Yamato (1977a,b) gives examples of various estimable

functions as special cases of g. For example, letting $g(x_1, \ldots, x_k) = x_1 x_2 \ldots x_k$, he derived the kth moment of the Dirichlet process $\mathcal{D}(\alpha)$.

Again, using the alternative definition of the Dirichlet process given by Ferguson mentioned before, namely $P(A) = \sum_{j=1}^{\infty} P_j \delta_{\xi_j}(A)$, $A \in \mathcal{A}$, Yamato (1977a,b, 1984) derived the joint moments of weights.

10. For any positive integers m, and any combination of positive integers $r_1, \ldots, r_m$, such that $\sum_{i=1}^{m} r_i = k$ and $M = \alpha(\mathcal{X})$,

$$\mathcal{E}\left(\sum\sum\cdots\sum_{j_1\neq j_2\neq\ldots\neq j_m} P_{j_1}^{r_1}\cdots P_{j_m}^{r_m}\right) = \frac{(r_1-1)!\cdots(r_m-1)!\cdot M^m}{M^{(k)}}. \tag{2.1.16}$$

As special cases we have

$$\mathcal{E}\left(\sum_{j=1}^{\infty} P_j^2\right) = 1/(M+1), \mathcal{E}\left(\sum_{i\neq j} P_i P_j\right) = M/(M+1),$$

$$\mathcal{E}\left(\sum_{j=1}^{\infty} P_j^3\right) = 2/[(M+1)(M+2)], \mathcal{E}\left(\sum_{i\neq j} P_i^2 P_j\right) = M/[(M+1)(M+2)],$$

$$\mathcal{E}\left(\sum_{i\neq j\neq k} P_i P_j P_k\right) = M^2/[(M+1)(M+2)], \mathcal{E}\left(\sum_{j=1}^{\infty} P_j^2\right)^2 = M(M+6)/M^{(4)},$$

$$\mathcal{E}\left(\sum_{j=1}^{\infty} P_j^2\right)^3 = M(M^2+18M+120)/M^{(6)}, \text{ etc..}$$

11. It is easier to compute the moments of p_j's using the Sethuraman (1994) representation. Let m_i's be as before. Then, for any combination $(m_1, \ldots, m_n)$ of $\{1, 2, \ldots, n\}$ Tiwari (1981, 1988) proved

$$\mathcal{E}\left(\Sigma^* p_{11}\cdots p_{1m_1} p_{21}^2\cdots p_{2m_2}^2\cdots p_{n1}^n\cdots p_{nm_n}^n\right) = \frac{\Pi_{i=1}^n[(i-1)!\alpha(\mathfrak{X})]^{m_i}}{\alpha(\mathfrak{X})^{(n)}}, \tag{2.1.17}$$

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

Yamato (1977a,b) also derived similar results using Ferguson's alternative definition. They both use Antoniak's (1974) result related to the previous property. For $n = 2$, the above result was established in Ferguson's (1973) paper.

Sethuraman and Tiwari (1982) have investigated the limits of prior distributions as the parameter α tends to various values. Using the representation (2.1.5), they have proved the following results.

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

$$\alpha_r(\mathfrak{X}) \to 0 \text{ and } \sup_A \mid \overline{\alpha}_r(A) - \overline{\alpha}_0(A) \mid \to 0 \text{ as } r \to \infty, A \in \mathcal{A},$$

where $\overline{\alpha}_r$ is a probability measure in Π. Then $\mathcal{D}(\alpha_r) \xrightarrow{w} \delta_{Y_0}$ and $\mathcal{D}(\alpha_r + \sum_{i=1}^n \delta_{x_i}) \xrightarrow{w} \mathcal{D}\left(\sum_{i=1}^n \delta_{x_i}\right)$ 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 point Y_0 selected according to $\overline{\alpha}_0$ distribution.

Various additional convergence properties of the Dirichlet process may be found in Ghosh and Ramamoorthi (2003) See also Lo (1983). The distribution of a linear functional of the Dirichlet process is studied in James (2006).

13. The distribution of $\mu(P)$, the mean of the process, can also be obtained from Hannum et al. (1981) [see Cifarelli and Regazzini (1979)] 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 $\exp\{-\int_R \log[1 - it\{g(t) - x\}]d\alpha(t)\}$. Using 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 $\alpha_n \xrightarrow{w} \alpha$ as $n \to \infty$, then under some mild regularity conditions $\int g dP_n$ converges in distribution to $\int g dP$.

14. Let X be a sample of size one from P. Then $\mathcal{P}(X \in A) = \alpha(A)/\alpha(\mathfrak{X})$ for $A \in \mathcal{A}$.

15. Let $X \sim P$. Then $\alpha(\mathfrak{X}) = \mathcal{E}[\text{Var}(X|P)]/\text{Var}[\mathcal{E}(X|P)]$. This property provides an interpretation of Baye's 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 an additional support for the interpretation of $\alpha(\mathfrak{X})$ as the prior sample size. However this interpretation is questioned by Walker and Damien (1998).

16. Let $P \sim \mathcal{D}(\alpha)$ and given $P = P$, let $X \sim P$. Then for any $A \in \mathcal{A}$, 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). That is knowing $X \in A$ does not add anything more to the process.

Multiplicities Since the DP is a discrete distribution, we expect samples from it to have duplications with probability one. This has some interesting consequences as first pointed out in Antoniak (1974). Let $X_1, \ldots, X_n$ be a random sample drawn from P, and $P \sim \mathcal{D}(\alpha)$. The marginal distribution of X_i is $\alpha(\cdot)/\alpha(\mathfrak{X})$. Normally we would expect them to be independent and therefore $\mathcal{P}\{X_j = X_i\}$ for any i and j is likely to be zero. However, this is not the case when P is a Dirichlet process. Assume α to be nonatomic and consider the case of $n = 2$. The conditional distribution of P given X_1 is $\mathcal{D}(\alpha + \delta_{X_1})$ (see Theorem 2.6), which shows that α is no longer nonatomic but has an atom of point mass 1 at X_1. Hence the probability of $X_2 = X_1$ given X_1 is in fact $1/(\alpha(\mathfrak{X}) + 1)$, independent of X_1, and the probability

of $X_2 \neq X_1$ is $\alpha(\mathfrak{X}) / (\alpha(\mathfrak{X}) + 1)$. Proceeding this way, one can argue, based on Blackwell–MacQueen characterization of the DP, that at the nth stage, probability of drawing a new distinct value is $\alpha(\mathfrak{X}) / (\alpha(\mathfrak{X}) + n - 1)$. This value is monotonically decreasing as n increases indicating that the probability of observing a new distinct value diminishes. Nevertheless it is noted that $k = k(n)$, the number of distinct values of observations among a sample of size n increases. Specifically,

17.

$$\mathcal{E}(k(n)) = \alpha(\mathfrak{X}) \sum_{m=1}^{n} 1/(\alpha(\mathfrak{X}) + m - 1) \approx \alpha(\mathfrak{X}) [\log(1 + n/\alpha(\mathfrak{X}))]. \tag{2.1.18}$$

Hence $\mathcal{E}(k(n)) \to \infty$ and $n \to \infty$. In fact Korwar and Hollander (1973) have shown that $k(n) \overset{\text{a.s.}}{\to} \infty$ and $n \to \infty$.

Thus, although new distinct values are increasingly rare, there are abundant of them. Since they are distributed as $\alpha(\cdot)/\alpha(\mathfrak{X})$, they provide information to estimate $\alpha(\cdot)$. On the other hand, the rate at which new distinct values appear depends on $\alpha(\mathfrak{X})$. It provides information on the magnitude of $\alpha(\mathfrak{X})$ and can be used to estimate it if unknown (see next property). Also, for a given sample size n, we would expect many more duplications when $\alpha(\mathfrak{X})$ is very small rather than when it is very large. Let the distinct values be denoted by $X_1^*, \ldots, X_k^*$, with their multiplicities $n_1, \ldots, n_k$, respectively, so that $n_1 + \cdots + n_k = n$.

18. Antoniak (1974) has derived the distribution of $k(n)$ as

$$\mathcal{P}\{k(n) = k\} = a_k^n \alpha(\mathfrak{X})^k / A_n(\alpha(\mathfrak{X})),$$

and given k, the joint distribution of multiplicities as

$$p(\mathbf{n}) = p_{n,K}(n_1, \ldots, n_k | k) = \frac{\alpha(\mathfrak{X})^k}{\alpha(\mathfrak{X})^{(n)}} \prod_{i=1}^{k} (n_i - 1)!. \tag{2.1.19}$$

Here $A_n(z) = z^{(n)} = z(z+1)\cdots(z+n-1)$ is the nth degree polynomial in z and a_k^n is the integer valued coefficient of the kth term. These are the absolute values of Sterling numbers of the first kind. $p(\mathbf{n})$ can be thought of as defining a prior distribution on the vectors $\mathbf{n} = (n_1, \ldots, n_k)$. See the section on species sampling models for further discussion on it.

19. Assuming α to be nonatomic and unknown, Korwar and Hollander (1973) established two interesting results:

(i) $k(n)/\log n \to \alpha(\mathfrak{X})$ a.s. as $n \to \infty$, 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.

Duplication of observations has given rise to two important combinatorial formulas. 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 $\sum_{i=1}^n i m_i = n$ and $\sum_{i=1}^n m_i = k$, the number of distinct observations. 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 $2m_2$ $X's, \ldots$, exactly m_n observations occur each n times and they are the last nm_n $X's$. For example, $m_1 = n$ and $m_i = 0$ for $i > 1$ means all observations are distinct. If $m_1 = \cdots = m_{n-1} = 0, m_n = 1$ means all observations are identical. Then Antoniak (1974) proved the following result.

20. Let α be nonatomic. Then

$$\mathcal{P}((X_1, \ldots, X_n) \in C(m_1, \ldots, m_n)) = \frac{n!}{\Pi_{i=1}^n i^{m_i}(m_i!)} \frac{\alpha(\mathfrak{X})^{\sum_{i=1}^n m_i}}{\alpha(\mathfrak{X})^{(n)}}. \tag{2.1.20}$$

This formula is discovered independently by Ewens (1972).

21. The marginal (predictive) distribution of $X_{n+1}|X_1, \ldots, X_n, \alpha$ is a rescaled version of the updated α^* measure, namely $\alpha_n^* = (\alpha + 1)/(\alpha(\mathfrak{X}) + n)$. It can be written as a mixture with $X_{n+1} \sim \alpha(\cdot)/\alpha(\mathfrak{X})$ with probability $M/(M+n)$ and X_{n+1} set equal to X_i, $i = 1, \ldots, n$, each with probability $1/(M+n)$ and $M = \alpha(\mathfrak{X})$. Since $X_1, \ldots, X_n|P$ are iid, they are exchangeable. Therefore one can rewrite the conditional distribution of $X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, \alpha$ as the mixture with $X_i \sim \overline{\alpha}$ with probability $M/(M+n-1)$ and $X_i = X_j, j = 1, \ldots, i-1, i+1, \ldots, n$, 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

$$X_{n+1}|X_1, X_2, \ldots, X_n \sim \frac{n}{M+n} \cdot \frac{1}{n} \sum_{j=1}^k n_j \delta_{X_j^*} + \frac{M}{M+n} \overline{\alpha}, \tag{2.1.21}$$

where as before X_j^* are $k \leq n$ distinct observations among n with frequency n_j, respectively. This implies that X_{n+1} will be a new observation with probability $M/(M+n)$ and will coincide with an existing observation with probability $n/(M+n)$, both of which do not depend upon k or on frequency n_j, important information is unutilized. This is remedied by the generalization of the Dirichlet process, namely the two-parameter Poisson–Dirichlet distribution (Pitman and Yor 1997. See Sect. 3.4.2).

22. The Dirichlet process induces an exchangeable distribution over partitions of N objects into K classes.

2.1.2.1 Characterization

Let $\mathcal{C}$ denote the class of all RPMs P such that either (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. Then Doksum (1974) proved the following characterizations:

1. If $P \notin \mathcal{C}$ is tailfree (see Chap. 5) 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 (see Chap. 4) 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 $X_1, \ldots, X_n$.

$$\begin{aligned} \mathcal{P}\left(X_{n+1} \in A | X_1, \ldots, X_n\right) &= \frac{\alpha(A) + \sum_{i=1}^{n} \delta_{x_i}(A)}{\alpha(\mathfrak{X}) + n} \\ &= p_n \frac{\alpha(A)}{\alpha(\mathfrak{X})} + (1 - p_n) \frac{\sum_{i=1}^{n} \delta_{x_i}(A)}{n}, \end{aligned} \tag{2.1.22}$$

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

2.1.3 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 nonparametric Bayesian problems—the motivator for developing the Dirichlet process in the first place. It forms the basis of various applications reported in Chaps. 6 and 7.

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

Theorem 2.6 (Ferguson) *Let $P \sim \mathcal{D}(\alpha)$ 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 $\alpha + \delta_x$. 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 2.7 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 proof of this theorem for the case $n = 1$, which can be extended to arbitrary n, depends upon his two propositions. (1) Let X be sample of size one from P, then the marginal distribution of $X \in A, A \in \mathcal{A}$ is $\alpha(A)/\alpha(\mathfrak{X})$; and (2) for any partition $(B_1, \ldots, B_k)$ of $\mathfrak{X}$ and $A \in \mathcal{A}$, the joint distribution of $X \in A$ and $P(B_1), \ldots, P(B_k)$ is given by

$$\begin{aligned}&\mathcal{P}\{X \in A, P(B_1) \leq y_1, \ldots, P(B_k) \leq y_k\}\\ &\quad = \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{\alpha(\mathfrak{X})} D\left(y_1, \ldots, y_k | \alpha_1^{(j)}, \ldots, \alpha_k^{(j)}\right),\end{aligned}$$

where $D\left(y_1, \ldots, y_k | \alpha_1^{(j)}, \ldots, \alpha_k^{(j)}\right)$ is the distribution function of the Dirichlet distribution with parameters $\left(\alpha_1^{(j)}, \ldots, \alpha_k^{(j)}\right)$ with $\alpha_i^{(j)} = \alpha(B_i)$ if $i \neq j$ and $\alpha_i^{(j)} = \alpha(B_i) + 1$ if $i = j$. Now to prove that the posterior distribution of $(P(B_1), \ldots, P(B_k))$ given X is the Dirichlet distribution

$$D(y_1, \ldots, y_k | \alpha(B_1) + \delta_X(B_1), \ldots, \alpha(B_k) + \delta_X(B_k)), \tag{2.1.23}$$

it is sufficient to show that this conditional distribution when integrated with respect to the marginal distribution of X over the set A yields the above distribution. This follows immediately as

$$\begin{aligned}&\int_A D(y_1, \ldots, y_k | \alpha(B_1) + \delta_X(B_1), \ldots, \alpha(B_k) + \delta_X(B_k))\, d\alpha(A)/\alpha(\mathfrak{X})\\ &= \sum_{j=1}^{k} \int_{B_j \cap A} D\left(y_1, \ldots, y_k | \alpha_1^{(j)}, \ldots, \alpha_k^{(j)}\right) d\alpha(A)/\alpha(\mathfrak{X})\\ &= \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{\alpha(\mathfrak{X})} D\left(y_1, \ldots, y_k | \alpha_1^{(j)}, \ldots, \alpha_k^{(j)}\right),\end{aligned} \tag{2.1.24}$$

completing the proof.

The implication of this theorem is that in obtaining the posterior distribution, all one has to do is to update the parameter α. It was noted in the characterization above that the DP is the only process not in $\mathcal{C}$ where the posterior distribution of $P(A)$ depends upon the number of observations that fall in A and not where they fall. This makes the DP easy to use, but otherwise not desirable. Lijoi and Prünster (2010) 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 distributions. Neutral to the right process is an example of this. The first implies the second, but not the other way around.

An extension of this theorem is proved in Antoniak (1974).

Theorem 2.8 (Antoniak) *Let $P \sim \mathcal{D}(\alpha)$ and θ be a sample of size one 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_x, which is the conditional distribution of θ given $X = x$, and transition measure $\alpha_u(\theta, \cdot) = \alpha(\cdot) + \delta_\theta(\cdot)$. In symbols, if $P \sim \mathcal{D}(\alpha)$, $\theta|P \sim P$, and $X|P, \theta \sim F(\theta, \cdot)$, then $P|X \sim \int_\Theta \mathcal{D}(\alpha + \delta_\theta) dH_x(\theta)$.*

In applications this model is generally referred to as a mixture of DPs model. Usually, it is difficult to find the posterior distribution $H_x(\theta)$. Therefore Gibbs sampling procedures are developed by which one can sample a posterior distribution in order to carry out Bayesian analysis. This is described later in this chapter.

2.1.3.1 Sampling a Dirichlet Process

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 $P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot)$ (2.1.5), 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 $\xi_j \overset{\text{iid}}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$. There are two approaches to generate the sample. 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 (CRP) presented formally in Sect. 4.8.1. It involves the process of generating an exchangeable sequence of random variables. The CRP is obtained by integrating out the P and thus it describes the marginal distributions in terms of random partitions determined by k tables in a restaurant. In both the cases, the sample is generated by generating the sequences $\{p_j\}$ and $\{\xi_j\}$ and then using $P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot)$ to produce a realization of P.

In the first approach, we need to generate a sequence $\{p_j, \xi_j\}_{j=1}^{\infty}$ taking values in $([0, 1] \times \mathfrak{X})^{\infty}$. Since p_j's are not independent, it is difficult to generate such

a sequence directly. Instead an alternative sequence $\{V_j, \xi_j\}_{j=1}^{\infty}$ taking values in $([0,1] \times \mathfrak{X})^{\infty}$ is generated, where each $V_j \overset{\text{iid}}{\sim} \text{Be}(1, \alpha(\mathfrak{X}))$ and $\xi_j \overset{\text{iid}}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$. Then we set $p_1 = V_1$ and $p_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$ for $j \geq 2$. 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 Blackwell and MacQueen (1973) described earlier in this section. 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. We have $X_1 \sim \alpha/\alpha(\mathfrak{X})$ 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 as noted earlier, can be written equivalently as

$$X_{n+1}|X_1, X_2, \ldots, X_n \sim \sum_{i=1}^{n} \frac{1}{\alpha(\mathfrak{X}) + n} \delta_{x_i} + \frac{\alpha(\mathfrak{X})}{\alpha(\mathfrak{X}) + n} \overline{\alpha}. \tag{2.1.25}$$

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_j be the number of times the color X_j^* is repeated, $j = 1, 2, \ldots, k, n_1 + \cdots + n_k = n$. Then the above expression can be written in terms of k distinct colors, as

$$X_{n+1}|X_1, X_2, \ldots, X_n, k \sim \sum_{j=1}^{k} \frac{n_j}{\alpha(\mathfrak{X}) + n} \delta_{X_j^*} + \frac{\alpha(\mathfrak{X})}{\alpha(\mathfrak{X}) + n} \overline{\alpha}. \tag{2.1.26}$$

This process is continued indefinitely. It has been interpreted in practical terms and popularized in culinary metaphor by the catchy name, CRP (attributed to Jim Pitman and Lester Dubins by Griffiths and Ghahramani 2006). The color of $(n+1)$th draw being the jth color can be interpreted as seating the $(n+1)$th patron at jth table in a restaurant. It forms the basis of many algorithms to generate posterior distributions when closed form is unobtainable.

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 $\alpha(\cdot)/\alpha(\mathfrak{X})$. 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 jth table (or orders dish X_j^*) with probability $n_j/(\alpha(\mathfrak{X})+n)$, where n_j is the number of customers already occupying the table (or enjoying the dish X_j^*), $j=1,2,\ldots,k$ i.e., $X_{n+1}=X_j^*$. 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_j which is a function of n, and not on the order in which the customers arrive. Thus a realization of p_j is obtained by

$$p_j=\lim_{n\to\infty}\frac{n_j(n)}{\alpha(\mathfrak{X})+n} \tag{2.1.27}$$

and $\xi_j \overset{\text{iid}}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$.

The distinction between the two methods is that in the stick-breaking method the weights generated are exact, whereas 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. 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 $(q_1,\ldots,q_N)\sim D(\alpha(\mathfrak{X})/N,\ldots,\alpha(\mathfrak{X})/N)$ and then define $P_N(\cdot)=\sum_{j=1}^N q_j\delta_{\xi_j}(\cdot)$ with $\xi_j \overset{\text{iid}}{\sim} \alpha(\cdot)/\alpha(\mathfrak{X})$. It can be shown that $P_N\to P$ in distribution, as $N\to\infty$.

The above procedures are for sampling from a DP. For the Bayesian analysis to move forward, we need to sample from the posterior of the DP. If the posterior is also a DP, as is the case with iid observations, we can adapt the above procedures by simply updating the parameter α. However in many applications, as will be seen later on in other sections, the posterior under different sampling models turns out to be complicated and the above procedures are no longer workable. In such cases, simulation methods are used to generate a sample from the posterior. This topic on computation will be presented sparingly later on and references will be given for further exploration.

2.1.4 Extensions and Applications of Dirichlet Process

The remarkable feature of the Dirichlet process, as the chart in Fig. 1 shows, is that it has led to the development of many extensions/generalizations and/or the Dirichlet process being a particular case of such processes. It may rightly be considered as a "base" giving rise to other prior processes. Its relation to various processes discussed in later sections is as follows.

The Dirichlet process is obviously a particular case of the *Dirichlet Invariant* and *mixtures of Dirichlet processes* introduced in Sects. 2.2 and 2.3, respectively. When

defined on the real line, the DP is also a *neutral to the right* process discussed in Sect. 4.2. A certain transformation of the *beta process* of Sect. 4.5 yields the DP and the DP is also a particular case of the *beta-Stacy* process presented in Sect. 4.7. It is a *tailfree* process of Sect. 5.1 with respect to every tree of partitions, and when the parameters of the *Polya tree* process of Sect. 5.2 are subjected to certain constraints, it yields the DP. When the discount parameter of the *two-parameter Poisson–Dirichlet* process of Sect. 3.4.2 is set to zero, the process reduces to the DP.

The Sethuraman countable mixture representation of the DP was originally used mainly for proving various properties of the DP. However in recent years, as pointed out in Sect. 1.3 and further detailed in Chap. 3, its use as an instrument in developing several related processes has exploded. This representation has four ingredients and by tempering them, a number of new extensions of the DP have been proposed 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 *finite discrete distribution priors* (Ongaro and Cattaneo 2004). If the weights defined by a one-parameter beta distribution Be $(1, \alpha(\mathfrak{X}))$ are replaced by those defined by a *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), and *Latent stick-breaking processes* (Rodriguez et al. 2010), 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). 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, such as the popularly known *Chinese restaurant* and *Indian buffet processes* discussed in Sects. 4.8.1 and 4.8.2, respectively. They have applications in nontraditional fields such as word documentation, machine learning, and mixture models.

A significant drawback in application of the above processes is that there are no convenient expressions for the posterior distributions available and hence the Bayesian analyses have to rely on simulation methods. For this purpose, however, fast running algorithms are available in the literature. The volume of literature on these processes is exhaustive and it is impossible to do a reasonable justice to them in this book. However, they are briefly introduced in Sect. 2.4.1. References are given for further exploration if interested.

All these processes belong to a family we call *Ferguson–Sethuraman* processes, discussed in the next chapter, to emphasize the importance of the countable mixture representation of the DP.

The above processes may be considered as special cases of a class of models proposed by Pitman (1996a,b) and called *species sampling models*,

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot) + \left(1 - \sum_{j=1}^{\infty} p_j\right) Q(\cdot), \tag{2.1.28}$$

where Q is the probability measure corresponding to a continuous distribution G, $\xi_j \overset{\text{iid}}{\sim} G$, and weights p_j's are constrained by the condition, $\sum_{j=1}^{\infty} p_j \leq 1$. However their use in mainstream statistics is very limited if not absent.

2.1.4.1 Applications of the Dirichlet Process

In view of its conjugacy property, the DP was convenient to use as prior in nonparametric Bayesian estimation of various parameters and functionals of an unknown distribution function F. This path was pursued in 1970s and 1980s and the beauty is that most of these results were obtained in closed form. There is a vast literature on these topics and we provide a summary of these results in Chaps. 6 and 7. Nevertheless, extensive references are included for the reader to pursue further if interested. Among the applications, Bayesian estimation of a CDF, symmetric CDF, empirical Bayes estimation, sequential estimation, and minimax estimation of a distribution function are derived in closed form. For the functionals of a CDF, estimation of the mean, median, variance, qth quantile, and location parameters have been given. Other applications, such as a regression problem, estimation of a density function, estimation of covariance and concordance coefficient, a bioassay problem, and a hypothesis testing problem, are discussed and end results are given in closed form in Chap. 6. Many of these applications are also derived for the right censored data in Chap. 7. For example, a nonparametric Bayesian analog of the popular Kaplan–Meier Product Limit (PL) estimator is obtained and shown that when the total mass of the prior goes to zero, it reduces to the PL estimator. Results obtained for other type of sampling, such as modified censoring scheme, progressive censoring, and left truncation, are also included in this chapter.

The current applications of the Dirichlet process are focused on Bayesian analysis of large and complex data. It is found to be very effective in making inferences in connection with hierarchical and group data modeling, dynamic mixtures, spatial modeling, and clustering of observations. In view of the advances in computational power, close form of results is no longer necessary. This has made deeper and more sophisticated treatment of data within the reach of data analysts. This will be evident in Chap. 3.

2.2 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.

Definition The Dirichlet Invariant process (DIP) is defined on the same line as the DP. 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 2.9 (Dalal) A $\mathcal{G}$-invariant RPM 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 $P \sim \mathcal{DGI}(\alpha)$.

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

If $\mathcal{G}$ consists of only one element, namely, the identity transformation, then as one would expect, the DIG corresponds to the 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. 6.

Tiwari (1981, 1988) extends the Sethuraman's (1994) alternative representation of the Dirichlet process to the DIP. 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 $j \geq 2$, $p_j = V_j \Pi_{i=1}^{j-1}(1 - V_i)$ and $V_j \sim \text{Be}(1, \alpha(\chi))$. Then the RPM P given by

$$P(A) = \sum_{j=1}^{\infty} p_j \frac{1}{k} \sum_{i=1}^{k} \delta_{g_i \xi_j}(A), \quad A \in \sigma(\mathcal{X}) \tag{2.2.1}$$

is a DIP with parameter α.

2.2.1 Properties

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

1. Let $P \sim \mathcal{DGI}(\alpha)$, and let $A \in \sigma(\mathcal{X})$. Then $P(A) = 0$ with probability one if and only if $\alpha(A) = 0$.

2. Let $P \sim \mathcal{DGI}(\alpha)$, 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 DIP are defined in the same way as for the Dirichlet process.

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

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

4. Let $P \sim \mathcal{DGI}(\alpha)$ 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

$$\mathcal{P}(X \in A, P(B_1) \le y_1, \ldots, P(B_m) \le y_m) = \sum_{i=1}^{m} \left[\frac{\alpha(A \cap B_i)}{\alpha(\mathcal{X})} D^*(y_1, \ldots, y_m) \right],$$

where $D^*(y_1, \ldots, y_m) = D(y_1, \ldots, y_m | \alpha(B_1), \ldots, \alpha(B_i) + 1, \ldots, \alpha(B_m))$.

5. If $P \sim \mathcal{DGI}(\alpha)$, it is also discrete with probability one, like the Dirichlet process.

6. The main property of the DIP is the following conjugacy property.

Theorem 2.10 (Dalal) *Let $P \sim \mathcal{DGI}(\alpha)$, 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 $\mathcal{DGI}(\alpha + \sum_{i=1}^{n} \delta^g_{X_i})$, where $\delta^g_{X_i} = (1/k) \sum_{j=1}^{k} \delta_{g_j X_i}$, $i = 1, \ldots, n$.*

Using the alternative definition of Tiwari (1981, 1988) for the Dirichlet Invariant process, Yamato (1986, 1987) and Tiwari prove properties of an estimable parameter φ that are similar to the properties 9 and 10 stated for the Dirichlet process, and extend a weak convergence result (property 12) 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.

Bayesian estimation of a symmetric distribution function and estimation of its functionals, such as a median, the location parameter, and other functionals are obtained in closed forms by these authors and are provided in the application chapters.

2.2.2 *Symmetrized Dirichlet Process*

Doss (1984) provides an interesting 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 $(0, \infty)$, 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 the 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 further extends the construction of symmetric Dirichlet priors to symmetric neutral to the right priors as follows. Let F_1 and F_2 be two independent neutral to the right distribution functions on $[0, \infty)$ and construct a random distribution function F as $F(t) = \frac{1}{2}F_1(t) + \frac{1}{2}(1 - F_2(-t^-)), t \in R$. Then it is a mixture of two neutral to the right distribution functions labeled 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.

2.3 Mixtures of Dirichlet Processes

There are situations where the Dirichlet process is inadequate. For example, consider the following bioassay problem (Ferguson 1974; Antoniak 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 nondecreasing 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 nonparametric Bayesian approach by choosing a Dirichlet process prior with parameter α for F, $F \sim \mathcal{D}(\alpha)$. However, it turns out that the posterior distribution of F given the data is not a 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 $F \sim \mathcal{D}(\alpha)$. Here again the posterior distribution is not the DP.

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. Motivated by the fact that the DP does not cover many of the situations encountered in Bayesian analysis, Antoniak (1974) introduced the concept of mixtures of DPs. A mixture of Dirichlet processes, roughly speaking, is a Dirichlet process where the parameter α is itself treated as random having a certain parametric distribution. The simplest mixture of Dirichlet processes would be the one that chooses P from $\mathcal{D}(\alpha_1)$ with probability π, and from $\mathcal{D}(\alpha_2)$ with probability $1 - \pi$. In this sense it is a parametric mixture of the Dirichlet processes. It chooses a continuous distribution with probability one. It is essentially an hierarchical modeling and allows greater flexibility by having the DP centered around some known parametric distribution.

The study of MDPs was pursued in detail in 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 using the MDPs model 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. As will be seen later on these mixtures of Dirichlet processes are extensively useful in modeling large scale high dimensional data. It allows one to proceed in a parametric Bayesian fashion. This approach essentially represents as a compromise between purely parametric and purely nonparametric models in applications. In this section the main ideas of MDPs, their important properties, and relevant theorems drawn from Antoniak's paper are presented. MDPs have been encountered in many applications in modeling data involving hierarchical structure, covariate data and spatial data. These applications and computing aspect for sampling from the posterior distribution to carry out Bayesian analysis will be discussed in Sect. 3.3. Its application in solving bioassay, empirical Bayes, and regression problems is covered in Chap. 6.

The mixture of Dirichlet processes should not be confused with the other type of mixtures mentioned in Sect. 3.3. For example, consider $f(x) = \int K(x, u)\, dG(u)$, where $K(x, u)$ is a known kernel and $G \sim \mathcal{D}(\alpha)$. Here the parametric functions are mixed with respect to a nonparametric mixing distribution.

2.3.1 *Definition*

Before a formal definition can be given, we need the following definition of a transition measure:

Definition 2.11 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, nonnegative, non-null measure on $(\Theta, \sigma(\Theta))$.
(b) For every $A \in \sigma(\Theta)$, $\alpha(\cdot, A)$ is measurable on $(U, \sigma(U))$.

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 $\alpha_u(\cdot)$.

An MDP is defined in the same way as the DP was defined. Recall that in defining the DP, Ferguson took a finite dimensional vector of P, $(P(A_1), \ldots, P(A_k))$ having a Dirichlet distribution with parameter $(\alpha(A_1), \ldots, \alpha(A_k))$. Now we replace this parameter vector with a bivariate extension of α vector, $(\alpha(u, A_1), \ldots, \alpha(u, A_k))$ and integrate u with respect to its distribution $H(u)$ to obtain a mixture of DPs. Formally,

Definition 2.12 *P is said to be a mixture of Dirichlet processes on $(\Theta, \sigma(\Theta))$ with mixing distribution H defined on $(U, \sigma(U))$ and transition measure α_u, if for any positive integer k 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*

$$\int_U D(\alpha(u, A_1), \ldots, \alpha(u, A_k)))\, dH(u),$$

i.e.,

$$\mathcal{P}(P(A_1) \le y_1, \ldots, P(A_k) \le y_k) = \int D(y_1, \ldots, y_k \mid \alpha_u(A_1), \ldots, \alpha_u(A_k))\, dH(u)$$

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

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, $u \sim H$, $P|u \sim \mathcal{D}(\alpha_u)$. The resulting marginal distribution of P is $\int_U \mathcal{D}(\alpha_u) dH(u)$. As an example, let $\alpha(u, A) = \alpha(A) + \delta_u(A)$, $u \sim H$, $P|u \sim \mathcal{D}(\alpha + \delta_u)$, then the resulting process is an MDP. Also, if $A_1, \ldots, A_k$ is any partition of Θ, then

$$(P(A_1), \ldots, P(A_k)) \sim \sum_{i=1}^{k} H(A_i) D(\alpha_u(A_1), \ldots, \alpha_u(A_i) + 1, \ldots, \alpha_u(A_k)). \tag{2.3.1}$$

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

Definition 2.13 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 $A_1, \ldots, A_m$, $C_1, \ldots, C_n$,

$$P\{\theta_1 \in C_1, \ldots, \theta_n \in C_n | u, P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_n)\} = \prod_{j=1}^{n} P(C_j) \text{ a.s.}$$

2.3.2 Properties

The following are important characteristics and properties of the MDP:

1. If $P \sim \int \mathcal{D}(\alpha_u(\cdot))dH(u)$ and θ is a sample of size one from P, then for any measurable set $A \in \sigma(\Theta)$, the marginal for θ is

$$\mathcal{P}(\theta \in A) = \int_U \frac{\alpha_u(A)}{\alpha_u(\Theta)} dH(u)$$

2. Let $P \sim \mathcal{D}(\alpha)$ and θ be a sample of size one from P, and let $A \in \sigma(\Theta)$ 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 index space $(A, \sigma(\Theta) \cap A)$, transition measure $\alpha_u = \alpha + \delta_u$ for $u \in A$, and mixing distribution $H_A(\cdot) = \alpha(\cdot)/\alpha(A)$. In symbols,

$$P|\theta \in A \sim \int_A \mathcal{D}(\alpha + \delta_u) dH_A(u).$$

If $A = \Theta$, then it reduces to the Dirichlet process. This should come as no surprise since, given $\theta \in \Theta$ means no information is given and therefore the posterior is same as the prior. If instead θ itself is observed, then the posterior distribution is not a mixture of DPs, but $\mathcal{D}(\alpha + \delta_\theta)$ as seen is Sect. 2.1.

3. The MDPs satisfy the conjugacy property. That is, if the prior is an MDP, so is posterior. However, there is an added complexity now since we have to compute the posterior distribution $H_{\boldsymbol{\theta}}(u)$ of u given $\boldsymbol{\theta}$. For this we need some additional conditions. This theorem is a special case of the Theorem (16) given below.

Theorem 2.14 (Antoniak) *Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be a sample of size n from P, $P \sim \int_U \mathcal{D}(\alpha_u(\cdot))dH(u)$. Suppose there exists a σ-finite, σ-additive measure μ on $(\Theta, \sigma(\Theta))$ such that for each $u \in U$, (i) α_u is σ-additive and absolutely continuous with respect to μ, and (ii) the measure μ has mass one at each atom of α_u. Then*

$$P|\boldsymbol{\theta} \sim \int_U \mathcal{D}\left(\alpha_u + \sum_{i=1}^{n} \delta_{\theta_i}\right) dH_{\boldsymbol{\theta}}(u), \tag{2.3.2}$$

where $H_{\boldsymbol{\theta}}$ is the conditional distribution of u given $\boldsymbol{\theta}$.

In applications it is usually difficult to compute $H_{\boldsymbol{\theta}}(u)$, since $\boldsymbol{\theta}$ will have duplications. A partial solution is

$$H_{\boldsymbol{\theta}}(u) = \frac{\frac{1}{\alpha_u(\Theta)^{(n)}} \prod_{i=1}^{k} \alpha_u'(\theta_i^*) \left(m_u\left(\theta_i^*\right) + 1\right)^{(n_i-1)} dH(u)}{\int_U \frac{1}{\alpha_u(\Theta)^{(n)}} \prod_{i=1}^{k} \alpha_u'(\theta_i^*) \left(m_u\left(\theta_i^*\right) + 1\right)^{(n_i-1)} dH(u)}, \tag{2.3.3}$$

where $\theta_1^*, \ldots, \theta_k^*$ are the k distinct observations, n_i is the frequency of θ_i^*, $\alpha_u'(\theta_i^*)$ denotes the Radon–Nikodym derivative of $\alpha_u(\cdot)$ with respect to μ, and $m_u\left(\theta_i^*\right) = \alpha_u'(\theta_i^*)$ if θ_i^* is an atom of α_u, zero otherwise.

It is worth noting that the observations not only affect the each component of the mixture as one would expect, but also alter the relative weighting of the components via the conditional distribution of u given $\boldsymbol{\theta}$.

4. Using the earlier mentioned result (of $C(\mathbf{m})$), and integrating $u \in B$ with respect to $H(u)$ Antoniak obtained the following result:

Proposition 2.15 *Let $P \sim \int_U \mathcal{D}(\alpha_u) dH(u)$, where α_u is nonatomic for all $u \in U$, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be a sample of size n from P. Then the posterior distribution of u given $(\theta_1, \ldots, \theta_n) \in C(m_1, \ldots, m_n) \in \sigma(\Theta)$ is determined by*

$$\mathcal{P}(u \in B | \boldsymbol{\theta} \in C(\mathbf{m})) = \frac{\int_B (\alpha(u, \Theta)^{\sum_{i=1}^{n} m_i} / \alpha(u, \Theta)^{(n)}) dH(u)}{\int_U (\alpha(u, \Theta)^{\sum_{i=1}^{n} m_i} / \alpha(u, \Theta)^{(n)}) dH(u)}. \tag{2.3.4}$$

Note that $m_1, \ldots, m_n$ appear here only through their sum which is $k(n)$, the number of distinct observations in the sample. The implication of assuming α to be nonatomic is that if $\alpha(u, \Theta) = M$, a constant independent of u, then the event $\boldsymbol{\theta} \in C(\mathbf{m})$ provides no information about u. But if α is atomic for some or all u, the posterior distribution of u given $\boldsymbol{\theta}$ will depend on whether any of the observations coincide with any of the atoms of α_u. For further elucidation his paper should be consulted.

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.

Theorem 2.16 (Antoniak) *Let $P \sim \int_U \mathcal{D}(\alpha_u) dH(u)$, θ be a sample of size one from P. If X is a random variable such that the conditional distribution of X given P, u and θ is $F(\theta, \cdot)$, then the conditional distribution of P given X is a mixture of Dirichlet processes with mixing distribution H_x, i.e., $P|X = x \sim \int_{\Theta \times U} \mathcal{D}(\alpha_u + \delta_\theta) dH_x(\theta, u)$, 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 2.17 *Let $P \sim \mathcal{D}(\alpha)$ on $(\Theta, \sigma(\Theta))$, and θ be a sample of size one 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_x, 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 $P \sim \mathcal{D}(\alpha)$, $\theta|P \sim P$, and $X|P, \theta \sim F(\theta, \cdot)$, then $P|X \sim \int_\Theta \mathcal{D}(\alpha + \delta_\theta) dH_x(\theta)$.*

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 2 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 2.18 (Blum and Susarla) *Let $P \sim \mathcal{D}(\alpha)$, and $\theta_1, \ldots, \theta_k$ be a sample from P. The conditional distribution of P given $\theta_i \in A_i$ for $i = 1, .., k$ 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*

$$\mu_l(A) = \frac{\alpha\left(A \cap A_l \cap A_{l-1}^c\right)}{\alpha(A_l) + l - 1} + \sum_{j=1}^{l-1} \frac{\alpha\left(A \cap A_j \cap A_{j-1}^c\right)}{\alpha\left(A_j\right) + j - 1} \prod_{i=j}^{l-1} \frac{\alpha(A_i) + i}{\alpha(A_{i+1}) + i}, \tag{2.3.5}$$

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

2.4 Dirichlet Mixture Models

The mixture models provide a statistical tool to model and analyze grouped data consisting of known or unknown number of groups, each with similar characteristics which can be captured by the assumption of a common underlying distribution. The distribution is assumed to be nonparametric allowing greater flexibility. Dirichlet process mixture models are especially useful tools in Machine Learning area; see, for example, Shahbaba and Neal (2009), Hannah et al. (2011), Wade et al. (2014), Blei and Frazier (2011) and Blei et al. (2003). In this section, we present some different kinds of mixture models encountered in practice. To carry out statistical inference, we need to obtain the posterior distribution which will be undoubtedly complicated and some simulations will need to be done. For this purpose, some simulation methods will be described briefly first. Following that we will present

some hierarchical and hierarchical mixture models and computational algorithms to carry out the necessary simulations. There is an extensive literature on mixture models based on Sethuraman representation. They are included in the next chapter. Finally, some generalizations are mentioned in the end.

In nonparametric Bayesian analysis, we usually encounter two types of mixtures. The first is parametric mixture of last section discussed in Antoniak (1974), 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. That is, $P(\cdot) = \int \mathcal{D}(\alpha_u) H(du)$, H a distribution function. The RPM so obtained was called a mixture of Dirichlet processes.

Discreteness of the DP causes problems when the unknown distribution is known to be continuous and may result in inconsistent estimators (Diaconis and Freedman 1986). A simple recourse is to introduce a convolution with a known kernel. Finding that in density estimation the DP is inadequate, Lo (1984) is the first one to follow this path. He considered the second type of mixture—the kernel mixture of Dirichlet processes—to serve as a prior on the space of density functions. He (Lo 1984) defines a kernel representation of the density function as, $f(x|G) = \int_R K(x,u)G(du)$, where G is a distribution function on R. Here $K(x,u)$ is a kernel defined on $(\mathcal{X} \times \mathcal{R})$ into R^+ such that for each $u \in R$, $\int_{\mathcal{X}} K(x,u)dx = 1$ and for each $x \in \mathcal{X}$, $\int_R K(x,u)\alpha(du) < \infty$. By treating G to be random with a $\mathcal{D}(\alpha)$ prior, a random mixing distribution (density) is defined. Here 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. Lo (1984) considers in his treatment a broad class of kernels, such as histogram, normal with location and/or scale parameters, symmetric and unimodal densities, and decreasing densities (see Sect. 6.5.4). Ferguson (1983), West (1992) and Escobar and West (1995) considered Dirichlet mixture of normals.

The difference between the two types is that the mixing components and mixing weights are interchanged. To distinguish it from the MDPs introduced by Antoniak, we will call this mixture as *kernel mixture of Dirichlet processes* (KMDP) or simply *Dirichlet Process Mixtures (DPM)*, as is known in certain publications. However they are related in a way. Consider, for example, the following basic mixture model: $y_i|\theta_i \sim f(\cdot|\theta_i)$, $\theta_i|G \overset{\text{iid}}{\sim} G$, $G|\alpha \sim \mathcal{D}(\alpha)$. Then $G|\boldsymbol{\theta} \sim \mathcal{D}(\alpha + \sum \delta_{\theta_i})$ and $G|\mathbf{Y} \sim \int \mathcal{D}(\alpha + \sum \delta_{\theta_i}) H(\boldsymbol{\theta}|\mathbf{Y})$ are an MDP with mixing distribution H. Thus the posterior distribution of G in a mixture model is an MDP.

A typical mixture is the normal mixture where the kernel is taken to be a normal density (say with known variance), $f(x) = \int N\left(x|\mu,\sigma^2\right) dG(\mu)$, where $N(x|\mu,\sigma)$ is the normal density with mean μ and variance σ^2. In general, this type of model may be stated as $f(x) = \int p(x|\theta)\, dG(\theta)$, and if we use the Sethuraman representation of $G = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}$, then we have $f(x) = \sum_{i=1}^{\infty} p_i p(x|\xi_i)$. This is known as a mixture model.

Ferguson (1983) considered a related but different formulation of the density function. He modeled it as a countable mixture of normal densities: $f(x) = \sum_{i=1}^{\infty} p_i N(x|\mu_i,\sigma_i)$. 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 $f(x) = \int N(x|\mu, \sigma)\, dG(\mu, \sigma)$, 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 mutually independent; $p_1, p_2, \ldots$ are the weights SBW(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 $\xi_i \overset{\text{iid}}{\sim} G_0$.

In both the cases, given a sample $x_1, \ldots, x_n$ of size n from a distribution with density $f(x) = \int N(x|\theta)\, dG(\theta)$, the posterior distribution of G given $x_1, \ldots, x_n$ is a mixture of Dirichlet processes . The Bayesian estimate of density function f is pursued in Sect. 6.5.4.

Normal mixtures also turn up in Escobar (1994) and Escobar and West (1995). Escobar considered the following model. Let $y_i|\mu_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$, $\mu_i|G \overset{\text{iid}}{\sim} G$, 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. Escobar also uses 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 the 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 as Dirichlet process with a normal base measure.

Here the authors assume $G \sim \mathcal{D}(M, G_0)$, 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 21 of Sect. 2.1. They proceed on the line of Ferguson and 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 $N(\mu_i, \sigma_i^2)$. Following that it is shown that the unconditional predictive distribution is given by $y_{n+1}|y_1, \ldots, y_n \sim \int P(y_{n+1}|\boldsymbol{\nu})\, dP(\boldsymbol{\nu}|y_1, \ldots, y_n)$; see Sect. 2.4.2 for further details.

In the above normal mixture model, the base distribution G_0 could be continuous. However, in certain applications such as modeling group data, we desire to have G_0 to be discrete instead. Rather than take it a priori a discrete distribution, it seems preferable to put another DP prior on G_0 itself. This has led to a class of hierarchical models discussed later. This may be viewed as a third kind of mixture. In all these

type of models, the posterior distribution is complicated and one has to resort to simulation process.

2.4.1 Sampling the Posterior Distribution

Density estimation problems described above are typical which reveal how difficult it is to compute the posterior distribution. Similarly, in models such as hierarchical, dependent DP, and spatial DP, the posterior distribution is not so simple. Earlier efforts dealt only with evaluation of posterior mean and variance. To carry out a full Bayesian analysis, we need to know the posterior distribution. Therefore one has to resort to simulation. This is accomplished by generating a sequence of values from the posterior distribution via simulation procedures. Based on these values, one can compute the various quantities such as moments. In fact a full Bayesian analysis can be carried out utilizing these simulated values.

However, we need to simulate a sufficiently large sequence of these samples so that the law of large numbers would assure us their consistency. During the last two decades various simulation methods are proposed and algorithms developed. They are mostly based on Markov chain Monte Carlo (MCMC) procedures using the Gibbs sampler. In recent years tremendous progress is made in improving these methods. Needless to say, there is an extensive literature on sampling methods. For a good starting point is to consult the papers contained in the book edited by Dey et al. (1998), Gelfand and Smith (1990) and for further improved methods, see the cross references in Walker (2007) and Papaspiliopoulos and Roberts (2008). We will first give a general outline of these procedures and later give relevant algorithmic steps while discussing specific models.

2.4.1.1 Gibbs Sampler

Basically, suppose we wish to draw a sample from the posterior distribution $p(\theta_1, \ldots, \theta_k|Y)$ of vector $\theta = (\theta_1, \ldots, \theta_k)$ given observations $Y = (y_1, \ldots, y_n)$. Let $p(\theta_i|\theta_{-i}, Y)$ denote the induced full (meaning all variables other than θ_i included) conditional distribution of each component θ_i given the other components $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$, for $i = 1, \ldots, k$, and known information, Y. Since the joint distribution can be written in terms of full conditional distributions, the procedure is to start with some initial values $\theta^0 = \left(\theta_1^0, \ldots, \theta_k^0\right)$ and generate $\theta^1 = \left(\theta_1^1, \ldots, \theta_k^1\right)$ by drawing θ_i^1, one by one, from the updated distribution in which we replace θ_i^0 with already drawn θ_i^1, until we replace all k θ's. Order does not matter. This is the first iteration from $\theta^0 = \left(\theta_1^0, \ldots, \theta_k^0\right)$ to $\theta^1 = \left(\theta_1^1, \ldots, \theta_k^1\right)$. Proceeding this way generates a sequence $\theta^0, \theta^1, \ldots, \theta^m, \ldots$ from $p(\theta_1, \ldots, \theta_k|Y)$, which is a realization of a Markov chain (MC), with transition probability kernel

from θ^m to θ^{m+1},

$$K\left(\theta^m, \theta^{m+1}\right) = \prod_{l=1}^{k} p(\theta_l^{m+1} | \theta_j^m, j > l, \theta_j^{m+1}, j < l, Y).$$

If the posterior is not of a known form, which would be the case if the prior is not conjugate, one uses the Metropolis–Hastings algorithm. It is a generalization of the acceptance-rejection rule. However, we have other methods that have proved to be more efficient.

2.4.1.2 Sampling Strategies

Generally speaking there are two different Gibbs sampling strategies: *marginal* and *conditional*. In the marginal, the strategy is to integrate out analytically the unknown distribution function G and then exploit the Polya urn characterization of Blackwell and MacQueen (1973) for drawing samples from the posterior distribution. This path was initiated by Escobar (1994) and pursued by Escobar and West (1995), MacEachern (1998) and West et al. (1994), among others. It can be applied to any mixture model for which the predictive distribution induced by G is known. This strategy is easily carried out in the case when we have conjugate prior. The process involves updating components one-at-a-time and is slow and there is tendency for the process to get stuck at the few distinct values or clusters formed earlier and new clusters are less frequently formed. In the absence of conjugacy, it is difficult and Metropolis–Hastings method needs to be used. But there again picking the right proposal distribution that would enable the Markov chain to converge well is difficult.

The conditional strategy introduced by Ishwaran and Zarepour (2000, 2003) and further developed in Ishwaran and James (2001, 2003) is to retain G and exploit its infinite sum representation. In this approach, the infinite sum is truncated at a suitable level and Gibbs sampling is carried out from the joint posterior distribution of weights p's, atoms ξ's, and cluster indicator s's, defined below. In this way, simultaneous updating of large subsets of the variables is done while focusing on finding appropriate ways for sampling a finite but large number of atoms of G. It is readily adaptable to more general stick-breaking measures and to many extensions of the DP. In addition, theoretically it allows inference for G as well. Under this approach, a number of procedures have been proposed. They include blocked Gibbs sampler (Ishwaran and Zarepour 2000; Ishwaran and James 2001), slice sampling (Neal 2000; Walker 2007), and retrospective sampling (Papaspiliopoulos and Roberts 2008). In the blocked Gibbs sampling, the infinite sum is truncated to a positive integer N. The convergence of the Markov chain (MC) is affected by the way N is chosen. In contrast, the latter two do not require truncation and the components in the infinite sum representation are added as needed.

Thibaux (2008) has noted that the Gibbs sampling procedure is too slow for large scale applications of DP mixtures. Therefore, he develops new Monte Carlo algorithms which are based on split-and-merge algorithms of Jain and Neal (2004).

2.4.1.3 Marginal Approach

The marginal approach of Escobar (1994) exploits the Polya urn characterization of the DP in constructing the MCMC using Gibbs sampler. We start with the following basic mixture model (West et al. 1994). Let

$$y_i|\theta_i, \sigma \overset{\text{ind}}{\sim} f\left(\cdot|\theta_i, \sigma\right), \theta_i|G \overset{\text{iid}}{\sim} G, i = 1, \ldots, n, G|\alpha \sim \mathcal{D}\left(\alpha\right), \tag{2.4.1}$$

where θ_i is a vector of parameters associated with index i, σ is a vector common to all i's, and $\alpha = MG_0$. The functional form of F with density $f\left(\cdot|\theta_i, \sigma\right)$ is assumed to be known. It may also depend on i as in the case of regression model (for example, $f_i\left(\cdot|\theta_i, \sigma\right) = N\left(\mathbf{X}_i\theta_i, 1\right)$). Here σ represents a vector of parameters associated with the distribution f, as was the case of normal mixture with unknown variance (West 1992). In hierarchical model we will have an additional set of hyperparameters, η. For example, if we assume $G \sim \mathcal{D}\left(M, G_0\right)$, $\eta = \left(M, G_0\right), G_0 = E\left(G\right)$ is the base prior, and M is the precision parameter as before. In such models, we further place priors on these parameters. However, in our treatment here we suppress σ and η, with the understanding that the following presentation is all conditional on both σ and η. Any realization of θ_i generated from G yields a set $\theta^* = \left(\theta_1^*, \ldots, \theta_k^*\right), k \leq n$ of distinct values and is marginally a random sample from G_0. Given k, n values of θ_i are selected from the set θ^* according to a uniform multinomial distribution.

In the Polya urn characterization of the DP, we integrate out G and get the joint distribution of θ's as

$$\pi\left(d\theta_1, \ldots, d\theta_n\right) = G_0\left(d\theta_1\right) \prod_{i=2}^{n} \left\{ \sum_{j=1}^{i-1} \frac{1}{M+n-1} \delta_{\theta_j}\left(d\theta_j\right) + \frac{M}{M+n-1} G_0\left(d\theta_i\right) \right\}$$

and the conditional predictive distribution as

$$\theta_n|\theta_{n-1}, \ldots, \theta_1 \sim \sum_{j=1}^{k^-} \frac{n_j}{M+n-1} \delta_{\theta_j^*} + \frac{M}{M+n-1} G_0, \tag{2.4.2}$$

where $\theta_1^*, \ldots, \theta_{k^-}^*$ are k^- distinct values with frequencies $n_1, \ldots, n_{k^-}$, respectively, among $\theta_1, \ldots, \theta_{n-1}$. That is, the posterior distribution of θ_n given $\theta_{n-1}, \ldots, \theta_1$ is a mixture of a discrete distribution with weights on other θ's, and a known distribution G_0. Since the sequence generated is exchangeable, we can replace θ_n in the above expression by θ_i conditional on $\theta_{-i} = \left(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n\right)$.

Multiplying the above with likelihood $f(y_i|\theta_i)$, we obtain the conditional posterior distribution of θ_i as

$$\begin{aligned}\theta_i|\theta_{-i}, \mathbf{y} &\propto \sum_{j=1}^{k^-} n_j^- f\left(y_j|\theta_j^{*-}\right)\delta_{\theta_j^{*-}} + Mf(y_i|\theta_i)\, G_0(\theta_i) \\ &= \sum_{j=1}^{k^-} n_j^- f\left(y_j|\theta_j^{*-}\right)\delta_{\theta_j^{*-}} + \left(M\int f(y_i|\theta_i)\, dG_0(\theta_i)\right) f(\theta_i|y_i, G_0) \\ &= \sum_{j=1}^{k^-} q_j^*\,\delta_{\theta_j^{*-}} + q_0^* f(\theta_i|y_i, G_0)\,, \text{ say,} \end{aligned} \tag{2.4.3}$$

where superscript $^-$ represents quantities we get when θ_i is excluded from the set and $f(\theta_i|y_i, G_0)$ is the posterior distribution of singleton θ_i based on y_i alone and $M\int f(y_i|\theta_i)\, dG_0(\theta_i)$ is the marginal of y_i. This leads to a Gibbs sampler for θ_i in which it takes one of the previously observed values θ_j^{*-} with probability proportional to $n_j^- f\left(y_j|\theta_j^{*-}\right)$ or a new value drawn from $f(\theta_i|y_i, G_0)$ with probability proportional to $M\int f(y_i|\theta_i)\, dG_0(\theta_i)$. That is,

$$\theta_i|\theta_{-i}, \mathbf{y} \sim \begin{cases} \theta_j^{*-} & \text{w.p. } \kappa n_j^- f\left(y_j|\theta_j^{*-}\right), j = 1, \ldots, k^- \\ \theta_{k^-+1}^* & \text{w.p. } \kappa M\int f(y_i|\theta_i)\, dG_0(\theta_i) \end{cases} \tag{2.4.4}$$

where κ is a normalizing constant.

The conditional distribution (2.4.2) clearly suggests that it is not necessary to generate all n $\theta's$ since some of them are expected to be the same, a consequence of the DP prior, but only fewer $\theta^{*\prime}s$ and identifiers s which attach θ to θ^*. This led MacEachern (1994) to suggest an improved approach in which he reduces the dimension of the state space and hence of the transition kernel, by considering $\theta^{*\prime}s$ instead of $\theta's$. This was refined in West et al. (1994).

Given k, let $s_i = j$ if $\theta_i = \theta_j^*$, $j = 1, \ldots, k$, so that, given $s_i = j$ and θ^*, $y_i \sim f\left(\cdot|\theta_j^*\right)$. The vector $S = (s_1, \ldots, s_n)$ determines grouping of $Y = (y_1, \ldots, y_n)$ into k distinct groups or clusters with $n_j = \#\{s_i = j\}, j = 1, \ldots, k$. Let I_j be the set of indices of observations in group j, i.e., $I_j = \{i : s_i = j\}$ and let $Y_j = \{y_i : s_i = j\}$ be the group of observations in cluster j. Since θ_j^* are a random sample from G_0, the posterior analysis reduces to a collection of k independent analyses, i.e., θ_j^* are conditionally independent with posterior density

$$\pi\left(\theta_j^*|Y, S, k\right) \equiv \pi\left(\theta_j^*|Y_j, S, k\right) \propto \prod_{i\in I_j} f\left(y_i|\theta_j^*\right) dG_0\left(\theta_j^*\right),\ j = 1, \ldots, k. \tag{2.4.5}$$

Since $\theta_i \sim G$ and $G \sim \mathcal{D}(M, G_0)$, we have conditional distribution

$$\theta_i|\theta_{-i}, s_{-i}, k_{-i} \sim \sum_{j=1}^{k_{-i}} \frac{n_j^-}{M+n-1}\delta_{\theta_j^{*-}} + \frac{M}{M+n-1}G_0. \tag{2.4.6}$$

This shows that θ_i can be generated such that with probability proportional to n_j^-, pick θ_j^{*-}, and with probability $M/(M+n-1)$ pick a new value distributed according to G_0. Knowledge of θ is theoretically equivalent to the knowledge of k, S, and θ^*. Now as before multiplying both sides by the likelihood, we get the conditional posterior distribution as

$$\theta_i|Y, \theta_{-i}, s_{-i}, k_{-i} \sim \sum_{j=1}^{k_{-i}} q_{ij}\delta_{\theta_j^{*-}} + q_{i0}G_{i0} \tag{2.4.7}$$

where

$$q_{i0} \propto Mh_i(y_i), q_{ij} \propto n_j^- f\left(y_i|\theta_j^*\right), j = 1, \ldots, k^-,$$

G_{i0} denotes the posterior, namely $dG_{i0}(\theta_i) \propto f(y_i|\theta_i)\, dG_0(\theta_i)$ and $h_i(y_i) = \int f(y_i|\theta_i)\, dG_0(\theta_i)$ is the marginal density of y_i evaluated at the realized datum under the base prior for θ_i. Since $s_i = j$ iff $\theta_i = \theta_j^{*-}$, Eq. (2.4.7) implies

$$\mathcal{P}(s_i = j|Y, \theta_{-i}, s_{-i}, k_{-i}) = \mathcal{P}\left(\theta_i = \theta_j^{*-}|Y, \theta_{-i}, s_{-i}, k_{-i}\right) = q_{ij},\ j = 0, 1, \ldots, k^-. \tag{2.4.8}$$

In other words, $s_i = j|Y, \theta_{-i}, S_{-i}, k_{-i}$ means θ_i takes value θ_j^{*-} with probability q_{ij}, $j = 1, \ldots, k^-$, $s_i = 0$, draw a new θ_i from G_{i0} of (2.4.7) and this would generate configuration θ^*. Now we may successfully sample sets of values k, S and $\theta^* = \left(\theta_1^*, \ldots, \theta_k^*\right)$, k will also vary depending upon how many distinct values are among θ_{-i}. The procedure at any stage m of the algorithm is

(a) Given current θ^* (hence k) and S, we generate a new configuration by sequentially sampling indicators one by one, from the posterior distribution (2.4.8), successively simulating and substituting $s_1, s_2, ..$; and mindful that for any index i such that $s_i = 0$, to draw a new θ_i from G_{i0} of (2.4.7). This will also yield a new k, the number of clusters. Next we need to pick random θ^*'s. Ishwaran and James (2001) suggest that it is simpler to use the current value of θ from which to compute current S and then update current θ^* given S.

(b) Given k and S, generate a new set of parameters θ^* by sampling each new θ_j^* from the relevant component $\prod_{r \in I_j} f_r\left(y_r|\theta_j^*\right) dG_0\left(\theta_j^*\right)$ in (2.4.5).

Successive simulated values will be the realized values of a sample from the joint posterior distribution $\pi(k, S, \theta|Y)$. Now inference can be based on these values. Convergence issues are discussed in MacEachern and Müller (1998).

Going back to the basic model, when σ is unknown and has a prior distribution $\pi(\sigma)$, then the posterior conditional distribution of σ given the data is

$$\pi(\sigma|Y,\theta,S,k) \propto \pi(\sigma)\prod_{i=1}^{n} f_i(y_i|\theta_i,\sigma) = \pi(\sigma)\prod_{j=1}^{k}\prod_{r\in I_j} f_r\left(y_r|\theta_j^*,\sigma\right). \tag{2.4.9}$$

Then the above sampling scheme would include an additional step.

(c) Simulate a value of σ from this distribution, given the current simulated values of θ, S, and k, at each step.

When we do not have conjugacy, evaluation of the integral in $h_i(y_i)$ becomes difficult. West et al. (1994) suggested replacing the integral with an average of integrals, average over draws θ' (in place of θ_i) from the base prior G_0. MacEachern and Müller (1998) pointed out that although this method does provide an approximation to the posterior, the accuracy is difficult to ascertain. Therefore, they suggest the so-called no gap algorithm. However Neal (2000) pointed out that this approach is potentially inaccurate since the posterior distribution $M\int f_i(y_i|\theta_i)\,dG_0(\theta_i)$ based on y_i alone will be considerably more concentrated than G_0, and as a consequence the process is slower in convergence but also may lead to the wrong stationary distribution, and is inefficient due to the reduced probability of assigning an observation to a new cluster.

Interestingly, Neal (2000) expresses this basic model (suppressing σ) as a finite mixture model with k components which is equivalent to the original model as $k \to \infty$. It is formulated in terms of grouping variables $s_i, i = 1,\ldots,n$, which identifies θ_j^*, the value θ_i assumes. Let $s_i \in \{1,..,k\}$ denote the class observation y_i belongs to

$$\begin{aligned} y_i|s_i,\theta^* &\sim f\left(\theta_{s_i}^*\right) \\ s_i|p &\sim \text{Discrete}\,(p_1,\ldots,p_k) \\ \theta_s^* &\sim G_0 \text{ and } p \sim D\left(\frac{\alpha}{k},\ldots,\frac{\alpha}{k}\right), \end{aligned} \tag{2.4.10}$$

where $p=(p_1,\ldots,p_k)$, D the Dirichlet distribution and α a positive real number.

By integrating over the mixing proportion p, we can obtain the conditional distribution

$$\begin{aligned} \mathcal{P}(s_i = j|s_1,\ldots,s_{i-1}) &= \frac{n_j^- + \alpha/k}{i-1+\alpha} \to \frac{n_j^-}{i-1+\alpha} \text{ as } k\to\infty \\ \text{and } \mathcal{P}\left(s_i \neq s_j \text{ for all } j<i|s_1,\ldots,s_{i-1}\right) &= \frac{\alpha}{i-1+\alpha}, \end{aligned} \tag{2.4.11}$$

where n_j^- is the number of s's among $s_1, \ldots, s_{i-1}$ that are equal to j. Now if we let $\theta_i = \theta^*_{s_i}$, we see that the limit of this model is equivalent to the original DP mixture model with α as the parameter of the DP.

2.4.1.4 Conditional Approach

Escobar and West and MacEachern's methods, as pointed out by Ishwaran and Zarepour (2000), suffer from two limitations: one, by marginalizing G, the Markov chain tends to mix slowly because the Gibbs sampler uses one coordinator at a time to update; two, it has the undesirable effect of allowing posterior inference to be based only on the values of Y. To circumvent these problems, they suggest replacing the DP prior P by its finite dimensional approximation P_N defined as

$$P_N = \sum_{j=1}^{N} p_j \delta_{\xi_j}, 1 \leq N < \infty, \tag{2.4.12}$$

where p's are random variables such that $0 \leq p_j \leq 1$ with $p_1 + \cdots + p_N = 1$, and independent of p's, $\xi_j \overset{\text{iid}}{\sim} H$. A key difference here is the weights p_j's need not be constructed using the SB construction. This will yield the posterior distribution also with such a representation and reduce the problem to a finite dimension allowing the model to be expressed in terms of a finite number of random variables. This will make implementation of the Gibbs sampler easier. This is the main idea.

It is similar in spirit to the Neal's model with one difference. The grouping variable s_i in Neal's model matches θ_i with cluster representative θ_j^*. Now the indicator variable k_i associates θ_i with $\xi_j, k_i \in \{1, \ldots, N\}$. That is, $\theta_i = \xi_{k_i}$. They both are, however, related. This can be seen by introducing another identifier $t_j = h$ iff $\theta_j^* = \xi_h$. In that case, $t_{s_i} = k_i$. In this approach also, k_i will be assumed to have some discrete distribution. This facilitates the Gibbs sampler to update blocks of parameters at a time, and generate a sequence (p^m, ξ^m, k^m), $m \geq 1$ of vectors $p = (p_1, \ldots, p_N), \xi = (\xi_1, \ldots, \xi_N)$, and $k = (k_1, \ldots, k_n)$. From this we can construct the posterior random measures $P_N^m = \sum_{j=1}^{N} p_j^{(m)} \delta_{\xi_j^{(m)}}$, which would be draws from the posterior measure $P_N(\cdot|Y)$, and based on them statistical inference can be performed. They named the procedure as *blocked Gibbs sampler.*

Blocked Gibbs sampler

The model (2.4.1) can equivalently be expressed as

$$y_i|\xi, k, \sigma \overset{\text{ind}}{\sim} f(y_i|\xi_{k_i}, \sigma), i = 1, \ldots, n$$

$$k_i|p \overset{\text{iid}}{\sim} \sum_{j=1}^{N} p_j \delta_j$$

$$(p, \xi) \overset{\text{ind}}{\sim} \pi(p)\,\pi(\xi)$$

$$\sigma \sim \pi(\sigma). \tag{2.4.13}$$

There are different possibilities to assign a distribution to p, such as a symmetric Dirichlet distribution with parameter α; a truncated beta two-parameter distribution; or a generalized Dirichlet distribution $\mathcal{GD}(a, b)$ with parameters $a = (a_1, \ldots, a_N)$ and $b = (b_1, \ldots, b_N)$.

To sample from the posterior distribution $P_N(\cdot|Y)$, we draw iteratively from the following conditional distributions:

$$\begin{aligned} &(\xi|k, Y, \sigma) \\ &(k|\xi, p, Y, \sigma) \\ &(p|k) \\ &(\sigma|\xi, k, Y). \end{aligned} \tag{2.4.14}$$

Each draw of (ξ, k, p, σ) defines an RPM $P(\cdot) = \sum_{j=1}^{N} p_j \delta_{\xi_j}(\cdot)$, which yields a draw from the posterior $P_N(\cdot|Y)$. Note that $\xi_j \overset{\text{iid}}{\sim} H$ and $k|p$ follow a discrete distribution. The auxiliary parameter σ will have some known distribution. Thus we need only to assign a prior for the vector p. This is done in such a way that the RPM P_N approximates the DP, P. The algorithm is as follows. Let $\{k_1^*, \ldots, k_m^*\}$ be the set of current distinct values.

(a) Sample $\xi_j \overset{\text{iid}}{\sim} H$ for each $j \in k \setminus \{k_1^*, \ldots, k_m^*\}$; and for $j = 1, \ldots, m$, draw $\xi_{k_j^*}$ from the density

$$H\left(\xi_{k_j^*}|k, \sigma, Y\right) \propto H\left(d\xi_{k_j^*}\right) \prod_{\left\{i:k_i=k_j^*\right\}} (fy_i|\xi_{k_j^*}, \sigma),\ j = 1, \ldots, m.$$

Conjugacy makes the draw exact. Otherwise, Metropolis–Hastings methods may be used. Hierarchical extensions can be implemented in a straight forward manner.

(b) Sample k_i from

$$(k_i|\xi, p, \sigma, Y) \overset{\text{ind}}{\sim} \sum_{j=1}^{N} p_{j,i} \delta_j(\cdot),\ i = 1, \ldots, n,$$

where

$$(p_{1,i},\ldots,p_{N,i}) \propto (p_1 f(y_i|\xi_1,\sigma),\ldots,p_N f(y_i|\xi_N,\sigma)).$$

(c) For p, Ishwaran and Zarepour (2000) propose three different choices for a prior.

(i) The first one is the commonly used approximation of the DP, the symmetric Dirichlet distribution, in which $p|\alpha \sim \mathcal{D}(\frac{\alpha}{N},\ldots,\frac{\alpha}{N}), \alpha > 0$. Sample the posterior $p|\alpha,k \sim \mathcal{D}(\frac{\alpha}{N}+n_1,\ldots,\frac{\alpha}{N}+n_M)$ for p, where $n_1,\ldots,n_M$ are the multiplicities of distinct values in the vector k. This choice is justified by noting that it approximates the DP for sufficiently large N. It leads to an easy update for the conditional distribution of p. Selection of N is discussed in their paper.

(ii) The second one is to assume $p|(\mathbf{a},\mathbf{b}) \sim GD(\mathbf{a},\mathbf{b})$, a generalized Dirichlet distribution with parameters $\mathbf{a},\mathbf{b}$. This results in exact draws from the posterior of SB ratios V_j^* and then computing p_j:

$$V_j^* \overset{\text{ind}}{\sim} \text{Be}\left(a_j+n_j, b_j+\sum_{l=j+1}^{N} n_l\right), \text{ for } j=1,..,N-1, \text{ and}$$

$$p_1 = V_1^*, p_j = V_j^* \prod_{i=1}^{j-1}(1-V_i^*). \tag{2.4.15}$$

This procedure requires simulation of $N-1$ beta random variables and is very efficient.

(iii) Another possibility is to use truncated beta two-parameter process, in which case $a_j = a$ and $b_j = b$ for all $j \geq 1$, in the above distribution of V_j^*'s. But then one has to be careful in choosing N appropriately.

(d) Conditional for σ : Noting that $y_i = \xi_{s_i}$ sample σ from the density

$$\pi(\sigma|\xi,S,Y) \propto \pi(d\sigma)\prod_{i=1}^{n} f(y_i|\xi_{s_i},\sigma).$$

Slice Sampling

This method introduced by Walker (2007), builds upon Neal's (2003) slice sampler, does not marginalize over G and employs the Sethuraman representation of the DP as well, $G \sim \mathcal{D}(M,G_0)$. However, the beauty is that it does not require truncation but reduces the infinite sum to a finite sum by introducing certain latent variables

$u_1, \ldots, u_n$, $0 \le u_i \le 1$, $i = 1, \ldots, n$, so that

$$\pi\left(y_i|\{p_j\}, \{\xi_j\}\right) = \int f\left(y_i|\theta_i\right) dG\left(\theta_i\right) = \sum_{j=1}^{\infty} p_j f\left(y_i|\xi_i\right) \tag{2.4.16}$$

is augmented to

$$\pi\left(y_i, u_i|\{p_j\}, \{\xi_j\}\right) = \sum_{j=1}^{\infty} I[p_j > u_i] f\left(y_i|\xi_i\right).$$

By integrating out u_i's it reduces to the above expression (2.4.16) again. By conditioning on u_i, the infinite mixture is transformed to a finite mixture with a fixed number of components $N_{u_i} = \sum_{j=1}^{\infty} I[p_j > u_i]$. Let $A_p\left(u\right) = \left\{j : p_j > u\right\}$. Now introducing further atom identifiers, $k_i \in \{1, 2, \ldots\}$, $i = 1, \ldots, n$, the augmented model becomes

$$\pi\left(y_i, u_i, k_i|\{p_j\}, \{\xi_j\}\right) = I[p_{k_i} > u_i] f\left(y_i|\xi_{k_i}\right). \tag{2.4.17}$$

Integrating over u_i and k_i, it again reduces to the original expression (2.4.16). The joint distribution of $\{y_i\}$, $\{u_i\}$ and $\{k_i\}$ is then

$$\pi\left(\{y_i, u_i, k_i\}_{i=1}^n|\{p_j\}, \{\xi_j\}\right) = \prod_{i=1}^{n} I[p_{k_i} > u_i] f\left(y_i|\xi_{k_i}\right). \tag{2.4.18}$$

Note that indicator k_i matches y_i with atom ξ_{k_i}, i.e., $k_i = j$ iff $y_i = \xi_j$, as before.

To implement the Gibbs sampler, we need to update p_j, ξ_j, u_i, and k_i, for which the key steps are (details may be found in Walker 2007, Müller et al. 2015):

1. Weights p_j are updated via the stick-breaking ratios V_j by sampling them from the updated beta distribution,

$$V_j| \ldots \sim \text{Be}\left(1 + n_j, M + \sum_{i>j} n_i\right),$$

   where $n_j = \sum_{i=1}^{n} I[k_i = j]$.
2. Atoms ξ_j are sampled from the posterior distribution proportional to $G_0\left(\xi_k\right) \prod_{\{i:k_i=k\}} p\left(y_i|\xi_k\right)$. If there are no $k_i = j$, sample atoms from G_0 directly.
3. Latent variables u_i are sampled from the (posterior) uniform distribution $U\left(0, p_{k_i}\right)$.
4. Indicators k_i are sampled from the distribution

$$\mathcal{P}\left\{k_i = k\right\} \propto I\left[p_k > u_i\right] f\left(y_i|\xi_k\right) = I\left[k \in A_p\left(u_i\right)\right] f\left(y_i|\xi_k\right).$$

Since only a finite number of components satisfy the constraint $p_j > u_i$, the normalizing constant can easily be evaluated as $\sum_{j\geq 1:p_j>u_i} f\left(y_i|\xi_j\right)$.

By introducing latent variables u_i, the process is simplified. Without them, we may have to sample an infinite number of ξ's, which is impossible. Papaspiliopoulos and Roberts (2008) also attempt to circumvent this problem by developing a retrospective sampling procedure, in which ξ's are generated retrospectively as the need arise. However, the procedure is not so simple and therefore is not included here.

2.4.2 Hierarchical and Mixture Models

In this section we will introduce various hierarchical and mixture models and processes that have appeared in the literature during the last decade in connection with modeling group and complex data. Besides hierarchical and mixture processes, they also include among others, nested and dynamic Dirichlet processes, dependent and spacial Dirichlet processes, and time-varying Dirichlet processes. They all exploit the discreteness property of the Dirichlet process to borrow information across observations as well as groups. It is clear that the Dirichlet process in their treatment can also be replaced by more general two-parameter Poisson–Dirichlet and beta processes. Computational methods described in the previous section are used for simulating posterior distributions to carry out inferential procedures. For details on computational and inferential procedures, and applications to real and simulation data, the reader is advised to refer to respective individual papers.

Generally in a nonparametric Bayesian modeling we have $y_1, \ldots, y_n \overset{\text{iid}}{\sim} G$ and $G \sim \mathcal{P}$, a certain prior on the distribution function G. That is an iid sample is drawn from a single realization of $\mathcal{P}$. In mixture modeling, we add an additional layer in between. We assume $y_i|\theta_i \overset{\text{ind}}{\sim} F_{\theta_i}$, a known parametric distribution, and place a prior on parameters $\theta_i|G \overset{\text{iid}}{\sim} G$ with $G \sim \mathcal{P}$. In group data modeling, we replace F_{θ_i} with a nonparametric distribution function G_i and assume $G_i \overset{\text{iid}}{\sim} \mathcal{P}$, i.e., each observation is drawn from a different realization of $\mathcal{P}$. It can also be extended to a mixture model in an obvious manner.

In the previous section we considered the following basic mixture model:

$$y_i|\theta_i, \sigma \overset{\text{ind}}{\sim} f\left(\cdot|\theta_i, \sigma\right), \theta_i|G \overset{\text{iid}}{\sim} G, i = 1, \ldots, n, G|\alpha \sim \mathcal{D}\left(\alpha\right). \tag{2.4.19}$$

The parameters θ_i's are conditionally independent given G, and the observations are conditionally independent given the parameters θ_i. A prior is placed on the nonparametric distribution G of parameters of the mixture model. When the prior is a DP, the model is referred to as a *Dirichlet Process mixture* (DPM) model or *Dirichlet mixture Model* (DMM).

It is obvious that the mixture models need not be restricted to the DP priors. Other priors such as the gamma process or two-parameter Poisson–Dirichlet process priors may also be used. In a recent publication, Favaro and Teh (2013) describe normalized random measure mixture models where the distribution function G is replaced by a normalized CRM with parameter ρ—a Levy measure, and base measure μ_0.

This model can alternatively be expressed in terms of the stick-breaking (SB) representation, order reflecting convenience for simulation purposes.

$$G|M, F_0 \sim \mathcal{D}(MF_0), \mathbf{p}|M \sim \text{SBW}(M), k_j|\mathbf{p} \sim \mathbf{p},$$
$$\xi_k|F_0 \overset{\text{iid}}{\sim} F_0, y_j|(k_j, \{\xi_k\}_{k=1}^{\infty}) \sim F\left(\cdot|\xi_{k_j}\right). \tag{2.4.20}$$

Also, $G = \sum_{k=1}^{\infty} p_k \delta_{\xi_k}$ and $\theta_j = \xi_{k_j}$ and $\mathbf{p} = (p_1, p_2, \ldots)$. This model can also be derived as the limit of a sequence of finite mixture models, where the number of mixture components tends to infinity, as we saw earlier.

In the above model, the distribution of the parameters of the function f was assumed to be nonparametric and a prior was placed on it. A different type of mixture is the one which leads to hierarchical models where the parameters of the prior distributions themselves are considered as random and assigned priors with hyper parameters. It has a long history of applications in parametric and semiparametric set ups. Its adaptation to the case of nonparametric (infinite dimensional parameters) may be framed as follows. In the usual nonparametric Bayesian models we assume $\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} F$ and F a Dirichlet process with parameters M and F_0. The baseline distribution F_0 is generally taken to be a parametric distribution. However we can go one step further—M or F_0, or both may be treated as random and priors may as well be assigned to them. For example, we may have for model,

$$\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} F, F|M, F_0 \sim \mathcal{D}(M, F_0) \text{ and } F_0|M^*, G_0 \sim \mathcal{D}\left(M^*, G_0\right), \tag{2.4.21}$$

each conditionally independent. A hierarchical model is thus basically a prior distribution over a set of RPMs or distributions whose parameters themselves are assigned priors. This type of models are referred to as *HDP* models or in general, *hierarchical models.*

2.4.2.1 Group Data Models

In the group data model, we associate with each subgroup an RPM G_j (RPM P replaced by G_j), and assume that we have variables θ_{ji} drawn from G_j, i.e., $\theta_{ji}|G_j \overset{\text{iid}}{\sim} G_j, i = 1, \ldots, n_j, j = 1, \ldots, J$, the objective being to find clusters that capture certain structure inherent within groups, the number of clusters being unknown, as well as to link groups in order to borrow information across groups. For the purpose of borrowing information, one may consider pooling the data into

one group and assign a single prior in which case the individual characteristics of the subgroups are lost. The other extreme case is to assign different priors to each subgroup in which case no information is shared. HDP model is a compromise between the two extremes in which dependence is implemented by sharing part of the probability mass across different groups. This is to be accomplished by sharing clusters among related groups. In such cases, the discreteness of the Dirichlet process turns out to be an asset. These models are applicable in the cases where the observations are assumed to be exchangeable within each group and across groups, groups themselves are exchangeable. They have proved to be especially useful in genetics and Information Retrieval fields.

Let G_0 be a global RPM. G_j are conditionally independent given G_0 and distributed according to a DP with base measure G_0 and concentration parameter M^*, i.e., $G_j|M^*, G_0 \overset{\text{ind}}{\sim} \mathcal{D}(M^*, G_0)$. Thus a reasonable model for group data may be stated as

$$\theta_{ji}|G_j \overset{\text{iid}}{\sim} G_j, i = 1, \ldots, n_j, j = 1, \ldots, J, G_j|M^*, G_0 \overset{\text{ind}}{\sim} \mathcal{D}\left(M^*, G_0\right). \tag{2.4.22}$$

The simplest way to borrow strength across groups is through the base measure G_0 and Sethuraman representation of G_j. G_0 may itself be taken as a parametric distribution and information may be shared through its parameters. But the choice of such a distribution may be considered as a limitation. Another possibility is to induce dependency among G_j's by integrating out the parameters of G_0. But they will not ensure any common atoms for some G_j and thus will not permit sharing of clusters between groups defeating the purpose on hand. Alternatively, some authors assume each G_j distributed as group-specific DP, say, $\mathcal{D}(M_j^*, G_{0j})$ and link the groups through M_j^* and/or G_{0j}. In such cases, however, each G_{0j} will have different sets of atoms and linking of groups via atoms will not be achievable.

2.4.2.2 Hierarchical/Mixture Models

Note that generally, G_0 is taken to be a nonatomic measure yielding distinct atoms with probability one in its SB representation. However, if we want clustering within groups, it is essential to force G_0 to be discrete so that a priori there are ties among its atoms. One could take it to be a discrete distribution to start with, but that would be too restrictive. Thus the proposed solution by Teh et al. (2004, 2006) is to not only force G_0 to be discrete by treating G_0 itself to be a draw from a DP, $\mathcal{D}(\gamma, H)$, but also treat each G_i to be drawn from the same DP, namely $\mathcal{D}(M^*, G_0)$, each G_i thus sharing the atoms of G_0. This does not cause any technical problem since the DP is defined on a general separable metric space. Thus we have the following model:

$$\theta_{ji}|G_j \overset{\text{iid}}{\sim} G_j, i = 1, \ldots, n_j, G_j|M^*, G_0 \overset{\text{iid}}{\sim} \mathcal{D}\left(M^*, G_0\right), G_0|\gamma, H \sim \mathcal{D}(\gamma, H). \tag{2.4.23}$$

This model is called a *HDP* model for group data. This can easily be extended to multiple levels of hierarchical modeling.

The hierarchical structure may be extended to mixture models involving group data, by placing the HDP prior over factors in the mixture model resulting into the following model:

$$y_{ji}|\theta_{ji} \overset{\text{ind}}{\sim} F\left(\cdot|\theta_{ji}\right), \theta_{ji}|G_j \overset{\text{iid}}{\sim} G_j, i = 1, \ldots, n_j, \; j = 1, \ldots, J,$$
$$G_j|\left(M^*, G_0\right) \overset{\text{iid}}{\sim} \mathcal{D}\left(M^*, G_0\right), G_0|\gamma, H \sim \mathcal{D}\left(\gamma, H\right), \tag{2.4.24}$$

where θ_{ji} is a factor corresponding to a single observation X_{ji} and F is assumed to be known. As usual all variables are assumed to be conditionally independent. The parameters $\left\{\theta_{ji}\right\}_{i=1}^{n_j}$ are likely to assume the atoms $\left\{\theta_k^*\right\}_{k=1}^{\infty}$—which are known as cluster parameters—of G_0 because of the SB representation of G_j (all G_j have the same set of atoms $\left\{\theta_k^*\right\}_{k=1}^{\infty}$). Here y_{ji} and $y_{ji'}$ belonging to the same group share the same atom θ^* of G_j, but also the observations across different groups may share them as a consequence of the discrete nature of G_0. This model is referred to as a *Hierarchical Dirichlet process mixture (HDPM) model*.

A formal study of such models in the context of nonparametrics described here was reported in Teh et al. (2006) where they develop these models. This was followed by other related extensions such as *nested DP* (Rodriguez et al. 2008), *dynamic DP* (Dunson 2006), *time-varying DP* (Caron et al. 2007), and *hierarchical dynamic DP* (Ren et al. 2008). The interest in these models stems from a need to model data which is subdivided into a set of groups, as in the case of multi-centric studies, where often there are group similarities forming group-clusters and groups themselves may show some clustering tendencies within each group as well, and the desire is to not only borrow strengths from within groups but also across the groups.

These models fall in the general framework of dependent nonparametric processes formulated by MacEachern (1999) and others and discussed later in Chap. 3. In this formulation $\mathcal{F}_\chi = \{F_x : x \in \chi\}$ is defined to be a class of related random probability measures indexed by $x \in \chi$, with the objective to borrow information across related/dependent RPM to strengthen inference procedures. $\mathcal{F}_\chi$ is treated as a stochastic process and the goal in Bayesian analysis is to place a joint distribution prior on $\mathcal{F}_\chi$. In covariate and spatial models χ is treated as a continuous space, usually a subset of finite dimensional Euclidean space; in the case of sequential and time-varying models, χ is considered as a countable infinite set of time points, $\chi = \{t_1, t_2, \ldots, \}$; and in case of group data, it takes a finitely many values, $\chi = \{1, \ldots, J\}$. Clearly the DP falls in this formulation when $\mathcal{F}_\chi$ has a single element and the distribution on $\mathcal{F}_\chi$ is the DP. Thus the dependent nonparametric processes and DDPs are extensive generalization of the DP to treat collectively related or dependent RPMs. This approach has generated lot of interest and a number of models have been proposed in the literature. We will describe some of them here and others later in the book under the heading Dependent nonparametric processes.

The Stick-Breaking Construction of HDPM

The stick-breaking construction for the above models may be framed as follows (Teh et al. 2006). The SB representation of the base measure G_0 and group measures G_j may be expressed as

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \text{ and } G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}, \tag{2.4.25}$$

where as usual β's are the SB weights, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots\} \sim \text{SBW}(\gamma), \phi_k \overset{\text{iid}}{\sim} H$. G_0 has support at the atoms $\{\phi_k\}_{k=1}^{\infty}$ and so necessarily G_j also has support at those atoms. Denote $\boldsymbol{\pi}_j = \{\pi_{j1}, \pi_{j2}, \ldots\}$. Weights $\boldsymbol{\pi}_j$ are independent given $\boldsymbol{\beta}$, since G_j's are independent given G_0. It can be seen that $\boldsymbol{\pi}_j$'s are related to β 's in the following way:

$$\pi_{jk} = \overline{\pi}_{jk} \prod_{m=1}^{k-1} \left(1 - \overline{\pi}_{jm}\right), \overline{\pi}_{jk} \sim \text{Be}\left(M^* \beta_k, M^* \left(1 - \sum_{l=1}^{k} \beta_l\right)\right). \tag{2.4.26}$$

In fact, $\boldsymbol{\pi}_j \sim \mathcal{D}(M^*, \boldsymbol{\beta})$, where $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}$ are viewed as discrete probability distributions over the positive integers. This may be seen as follows. Any finite partition $(A_1, \ldots, A_k)$ of $\mathfrak{X}$ induces a corresponding partition $(N_1, \ldots, N_k)$ of $\mathbb{N}$ such that $N_j = \{i : \phi_i \in A_j\}, j = 1, \ldots, k$. Therefore,

$$\begin{aligned} &\left(G_j(A_1), \ldots, G_j(A_k)\right) \sim D\left(M^* G_0(A_1), \ldots, M^* G_0(A_k)\right) \\ &\Rightarrow \left(\sum_{k \in N_1} \pi_{jk}, \ldots, \sum_{k \in N_k} \pi_{jk}\right) \sim D\left(M^* \sum_{k \in N_1} \beta_k, \ldots, M^* \sum_{k \in N_k} \beta_k\right), \end{aligned} \tag{2.4.27}$$

for every finite partition of $\mathbb{N}$. Thus $\boldsymbol{\pi}_j \sim \mathcal{D}(M^*, \boldsymbol{\beta})$ for each j independently.

Since each θ_{ji} is distributed according to G_j, it takes on the values ϕ_k with probability π_{jk}. As before for the mixture model, let k_{ji} be an indicator variable such that $\theta_{ji} = \phi_{k_{ji}}$. Given k_{ji}, we have $y_{ji} \sim F\left(\cdot|\phi_{k_{ji}}\right)$. Thus an equivalent representation of the *HDPM* model may be stated in terms of conditional distributions as follows:

$$\begin{aligned} &\boldsymbol{\beta}|\gamma \sim \text{SBW}(\gamma), \boldsymbol{\pi}_j|(M^*, \boldsymbol{\beta}) \sim \mathcal{D}\left(M^*, \boldsymbol{\beta}\right), k_{ji}|\boldsymbol{\pi}_j \sim \boldsymbol{\pi}_j, \\ &\phi_k|H \sim H, \; y_{ji}|(k_{ji}, \{\phi_k\}_{k=1}^{\infty}) \sim F\left(\cdot|\phi_{k_{ji}}\right). \end{aligned} \tag{2.4.28}$$

Chinese Restaurant Franchise

The CRP characterization of the DP is extended to the hierarchical DP where multiple restaurant share a set of common dishes. The authors name it as

Chinese Restaurant Franchise. Here restaurants correspond to groups, customers to parameters θ_{ji}, ith customer in jth restaurant. Also, let $\phi_1, \ldots, \phi_K$ denote K iid random variables distributed according to H, representing common dishes across restaurants. Let ψ_{jt} represent the table-specific dish (only one) served at table t in restaurant j, $\psi_{jt} \in \{\phi_1, \ldots, \phi_K\}$, t_{ji} be the index of the ψ_{jt} associated with θ_{ji} and let k_{jt} be the index of ϕ_k associated with ψ_{jt}. That is, $\theta_{ji} = \psi_{jt_{ji}}$ and $\psi_{jt} = \phi_{k_{jt}}$. In the Chinese Restaurant franchise metaphor, customer i in restaurant j sits at table t_{ji} whereas table t in restaurant j serves dish k_{jt}. Also, let n_{jtk} denote the number of customers in restaurant j sitting at table t and consuming dish k. Marginal totals will be denoted by dots. Thus $n_{jt\cdot}$ denotes the number of customers in restaurant j at table t, $n_{j\cdot k}$ denotes the number of customers in restaurant j consuming dish k, m_{jk} denotes the number of tables in restaurant j serving dish k, $m_{j\cdot}$ denotes the number of tables in restaurant j, $m_{\cdot k}$ denotes the number of tables serving dish k, and finally, $m_{\cdot\cdot}$ denotes the total number of tables occupied. By integrating out G_j, the conditional distribution of θ_{ji} given $\theta_{j1}, \ldots, \theta_{ji-1}, M^*, G_0$, is obtained as

$$\theta_{ji}|\theta_{j1}, \ldots, \theta_{ji-1}, M^*, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i-1+M^*}\delta_{\psi_{jt}} + \frac{M^*}{i-1+M^*}G_0. \tag{2.4.29}$$

This is a mixture with the usual interpretation of a draw from this mixture. Patron θ_{ji} will be seated at previously occupied table serving table-specific dish ψ_{jt} (i.e., θ_{ji} will take value ψ_{jt}) with probability $n_{jt\cdot}/(i-1+M^*)$ and with probability $M^*/(i-1+M^*)$ will choose a new table-specific dish ψ_{jt} according to the distribution G_0. Integrating out G_0 next, the conditional distribution of ψ_{jt} is obtained as

$$\psi_{jt}|\psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{jt-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot}+\gamma}H. \tag{2.4.30}$$

Table-specific dish ψ_{jt} will take value ϕ_k, $k = 1, \ldots, K$ with probability $m_{\cdot k}/(m_{\cdot\cdot}+\gamma)$ and will select with probability $\gamma/(m_{\cdot\cdot}+\gamma)$ a new dish $\phi_{k'}$ according to the distribution H. To obtain samples of θ_{ji}, proceed as follows. For each j and i, first sample θ_{ji} using the first expression. If a new sample from G_0 is needed, then use the second expression to obtain a new sample ψ_{jt} and set $\theta_{ji} = \psi_{jt}$. It should be noted that in HDP, the values of the factors are shared within as well as between groups. The above result describes marginals under a hierarchical DP when G_0 and G_j are integrated out.

For statistical inference, we need to sample the posterior distribution. For this purpose, the marginal approach is used in which θ_{ji} and ψ_{jt} are generated using the Gibbs sampling scheme. As indicated earlier, rather than generating them directly, it is more efficient (Neal 2000) to sample their index variables t_{ji} and k_{jt} and ϕ_k, from which θ_{ji} and ψ_{jt} can be reconstructed. The following sampling scheme is provided by the authors:

1. Sampling $\mathbf{t}$. The likelihood of y_{ji} is

$$p\left(y_{ji}|\mathbf{t}^{-ji}, t_{ji} = t^{\text{new}}, \mathbf{k}\right) = \sum_{k=1}^{K} \frac{m_{.k}}{m_{..} + \gamma} f_k^{-y_{ji}}\left(y_{ji}\right) + \frac{\gamma}{m_{..} + \gamma} f_{k^{\text{new}}}^{-y_{ji}}\left(y_{ji}\right) \quad *$$

where $f_{k^{\text{new}}}^{-y_{ji}}\left(y_{ji}\right) = \int f\left(y_{ji}|\phi\right) h\left(\phi\right) d\phi$ is the prior density of y_{ji}.

Therefore the conditional distribution of t_{ji} is given by

$$\mathcal{P}\left(t_{ji} = t|\mathbf{t}^{-ji}, \mathbf{k}\right) \propto \begin{cases} n_{ji\cdot}^{-ji} f_{k_{jt}}^{-y_{ji}}\left(y_{ji}\right) & \text{if } t \text{ is previously used} \\ M^* p\left(y_{ji}|\mathbf{t}^{-ji}, t_{ji} = t^{\text{new}}, \mathbf{k}\right) & \text{if } \quad t = t^{\text{new}}. \end{cases}$$

If $t = t^{\text{new}}$, then we sample $k_{jt^{\text{new}}}$ according to *

$$\mathcal{P}\left(k_{jt^{\text{new}}} = k|\mathbf{t}, \mathbf{k}^{-jt^{\text{new}}}\right) \propto \begin{cases} m_{.k} f_k^{-y_{jt}}\left(y_{jt}\right) & \text{if } k \text{ is previously used} \\ \gamma f_{k^{\text{new}}}^{-y_{jt}}\left(y_{jt}\right) & \text{if } \quad k = k^{\text{new}}. \end{cases}$$

If $n_{jt\cdot} = 0$, then delete the corresponding k_{jt} and by doing this if some mixture component k becomes unallocated, then delete that mixture component as well.

2. Sampling $\mathbf{k}$. Sample k_{jt} according to:

$$\mathcal{P}\left(k_{jt} = k|\mathbf{t}, \mathbf{k}^{-jt}\right) \propto \begin{cases} m_{.k} f_k^{-y_{jt}}\left(y_{jt}\right) & \text{if } k \text{ is previously used} \\ \gamma f_{k^{\text{new}}}^{-y_{jt}}\left(y_{jt}\right) & \text{if } \quad k = k^{\text{new}}. \end{cases}$$

Two other computation methods and discussion of the relative merits of these procedures can be found in their paper.

2.4.2.3 Nested Dirichlet Process

In HDP models, the information is shared across groups represented by distributions $G_1, \ldots, G_J$ by sharing the atoms of the base measure G_0 which allows us to identify common clusters across groups. However, it does not reveal any potential clustering of group distributions with similar characteristics among $G_1, \ldots, G_J$. Rodriguez et al. (2008) introduced the *nested Dirichlet process* (nDP) which allow such clustering. Their motivation was that in multi-centric studies, different centers may have different outcome distributions but also some clustering among mutually similar centers may be possible. This is different from clustering of observations within and across centers discussed by Teh et al. (2006) in their HDP model. One can potentially cluster centers via having some common parametric distributions or parameters, but that would be restrictive by the choice of the unknown distributions. Rodriguez et al. offer an alternative which is more flexible.

Like HDP, nDP is also a hierarchical model involving two levels of DPs: $G_j \overset{\text{ind}}{\sim} \mathcal{D}\left(M^*, G_0\right)$ with G_0 assumed to be the DP, $\mathcal{D}\left(\gamma, H\right)$. That is unlike in the HDP

model, here the baseline measure G_0 for the first level DP is itself a DP instead of an RPM drawn from it. Think of it as follows. A DP is a prior distribution or probability measure say, Q, over the space of distributions or probability measures. Now consider the class of all such Q's and put a DP prior Q^* on the space of Q's, which is permissible since the original definition of DP is on any complete separable metric space under the weak topology. This Q^* being a DP has a SB representation, say $\sum_{k=1}^{\infty} \beta_k \delta_{G_k^*}$, where G_k^* are members of Q-space and they being DP measures themselves admit SB representation also. Just as DP is a distribution on distributions, nDP can be viewed as a distribution on the space of distributions of distributions, i.e., $G_j \sim \mathcal{D}(M^*, \mathcal{D}(\gamma, H))$. This does not create any problem. This translates to replacing random atoms in the SB representation of the DP, with RPMs drawn from a second DP.

Thus the proposed model may be stated as follows. Let y_{ij}, $i = 1, \ldots, n_j$ be the observations for different subjects within center $j = 1, \ldots, J$. Assume subjects are exchangeable within centers, with $y_{ij} \overset{\text{iid}}{\sim} F_j$, $j = 1, \ldots, J$. The goal is to borrow information and allow clustering among distributions $\{F_j; j = 1, \ldots, J\}$. Thus let $\{G_1, \ldots, G_J\}$ be a collection of distributions such that $G_j \sim \mathcal{D}(M^*, \mathcal{D}(\gamma, H))$, and let $F_j(\cdot|\phi) = \int_\Theta p(\cdot|\theta, \varphi) G_j(d\theta)$, where $p(\cdot|\theta, \varphi)$ is a distribution parametrized by the finite dimensional vectors θ and φ. For example, $\theta = \mu, \varphi = \sigma^2$ and $p(\cdot|\theta, \varphi) = N(\cdot|\mu, \sigma^2)$ yield a class that is dense on the space of absolutely continuous distributions (Lo 1984) defined on the real line. The proposed model is

$$y_{ij} \overset{\text{iid}}{\sim} F_j(\cdot|\phi) = \int_\Theta p(\cdot|\theta, \varphi) G_j(d\theta), G_j \sim \mathcal{D}\left(M^*, \mathcal{D}(\gamma, H)\right), j = 1, \ldots, J. \tag{2.4.31}$$

The collection $\{F_1, \ldots, F_J\}$ is said to have an *nDP mixture* model. The definition of nDP implies the following SB representation:

$$\begin{aligned} &G_j \sim \sum_{k=1}^{\infty} \beta_k \delta_{G_k^*}, j = 1, \ldots, J, \; G_k^* \equiv \sum_{l=1}^{\infty} \pi_{lk} \delta_{\phi_{lk}}, k = 1, 2, \ldots \\ &\beta_k \sim \text{SBW}\left(M^*\right), \pi_{lk} \sim \text{SBW}(\gamma), \text{ for each } k = 1, 2, \ldots \text{ and } \phi_{lk} \overset{\text{iid}}{\sim} H. \end{aligned} \tag{2.4.32}$$

The first level of the hierarchy generates a distribution on random probability measures with atoms corresponding to random distributions $G_1^*, G_2^*, \ldots$, which are taken to be nonparametric and are drawn from a common DP, $\mathcal{D}(\gamma, H)$.

Since each G_k^* is a.s. discrete, some G_j and $G_{j'}$ will coincide with G_k^* for some k, thus forming the cluster of centers. On the other hand, subjects i and i' from centers j and j' will be clustered together if and only if $G_j = G_{j'} = G_k^*$ and $\theta_{ij} = \theta_{i'j'} = \phi_{lk}$ for some l. Thus the model highlights clustering of similar distributions at the first level and then clustering of observations takes place at the second level only across already clustered together distributions. In comparison with HDP it is

noted that in HDP, one draw from a DP is used as the baseline measure G_0. This implies that $\{G_1, \ldots, G_J\}$ share the same atoms (of G_0), but assign them different weights. Therefore, $P\left(G_j = G_{j'}\right) = 0$ under HDP and clustering occurs only at the level of the observations. Under nDP, clustering is induced on both observations and distributions. This means that the different distributions have either the same atoms with the same weights if they belong to the same cluster of distributions or completely different atoms and weights if they don't.

It also induces a random partition of a set of random distributions. Since a priori $P\left(G_j = G_{j'}|H\right) = \frac{1}{1+M^*} > 0$, the model induces clustering in the space of distributions and thus creates a collection of dependent random distributions. For $A \in \mathcal{A}$, $G_j(A)$ and $G_{j'}(A)$ are random variables with correlation coefficient between them given by $corr\left(G_j(A), G_{j'}(A)|H\right) = \frac{1}{1+M^*}$ and the correlation coefficient between two draws is

$$\text{corr}\left(\theta_{ij}, \theta_{i'j'}\right) = \begin{cases} \frac{1}{1+\gamma} \text{ if } j = j' \\ \frac{1}{(1+M^*)(1+\gamma)} \text{ if } j \neq j'. \end{cases} \tag{2.4.33}$$

It would be seen from this that as $M^* \to \infty$, $\text{corr}\left(G_j(A), G_{j'}(A)|H\right)$ is 0, each distribution in the collection is assigned to a distinct atom. That is, distributions become independent a priori. On the other hand, as $M^* \to 0$, the correlation goes to 1 implying the a priori probability of assigning all of the distributions to the same atom G_k^* goes to 1.

For computing posterior distributions, the authors follow the truncation approximation method (Ishwaran and Zarepour 2000) in which they truncate the infinite sums in G_j and G_k^* to K and L, respectively. This method is closely related to the method proposed by Ishwaran and James (2001, 2003). The main steps of their computation algorithm are as follows:

1. Sample the center indicators ς_j ($\varsigma_j = k$ if $G_j = G_k^*$), $j = 1, \ldots, J$ from a multinomial distribution with probabilities

$$\mathcal{P}\left(\varsigma_j = k|\ldots\right) \propto \beta_k \prod_{i=1}^{n} \sum_{l=1}^{L} \pi_{lk} p\left(y_{ij}|\phi_{lk}, \varphi\right).$$

2. Sample the atom indicators η_{ij} ($\eta_{ij} = l$ if $\theta_{ij} = \phi_{lk}$), $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$ from another multinomial distribution with probabilities

$$\mathcal{P}\left(\eta_{ij} = l|\ldots\right) \propto \pi_{l,\varsigma_j} p\left(y_{ij}|\phi_{l\varsigma_j}, \varphi\right).$$

3. Sample β_k by generating its SB ratios from $\text{Be}\left(1 + m_k, M^* + \sum_{s=k+1}^{K} m_s\right)$, $k = 1, \ldots, K-1$, and setting the last component as one, where m_k is the number of distributions assigned to component k.

4. Sample π_{kl} by generating its SB ratios from Be $\left(1 + n_{lk}, \gamma + \sum_{s=l+1}^{L} n_{ls}\right)$, $l = 1, \ldots, L-1$; and setting the last component as one, where n_{lk} is the number of observations assigned to atom l of distribution k.
5. Sample atoms ϕ_{lk} from the posterior distribution

$$p\left(\phi_{lk}|\ldots\right) \propto \prod_{\{i,j|\varsigma_j=k,\eta_{ij}=l\}} p\left(y_{ij}|\phi_{lk},\varphi\right) h\left(\phi_{lk}\right),$$

where $h\left(\phi_{lk}\right)$ is the density corresponding to the base measure H.
6. Sample φ from its full conditional distribution.

$$p\left(\varphi|\ldots.\right) \propto \prod_{j=1}^{J}\prod_{i=1}^{n_j} p\left(y_{ij}|\phi_{\eta_{ij},\varsigma_j},\varphi\right) p\left(\varphi\right).$$

If the concentration parameters M^* and γ are also unknown, they can be sampled as well.

It is clear that nesting can not only be done with respect to DP only. (Jordan 2010) remarked that we can in a similar fashion define a nested beta process as $B \sim BeP(\sum_{k=1}^{\infty} \beta_k^* \delta_{B_k^*})$, where $B_k^* = \sum_{l=1}^{\infty} p_{kl}\delta_{\omega_{kl}}$. This defines a random measure B which is a collection of atoms, each of which is a beta process. This may be better suited for document modeling.

2.4.2.4 Dynamic Mixture and Hierarchical Dirichlet Models

In practical applications often the data is collected in sequential manner over time and it is assumed that a time evolution exists between adjacent data groups. To incorporate this type of time-dependence, Dunson (2006), Ren et al. (2008), and others have introduced what is known as *dynamic* models that is described now.

The data consists of J groups collected sequentially as $\{y_{1i}; i = 1, \ldots, n_1\}$, $\ldots, \{y_{Ji}; i = 1, \ldots, n_J\}$, each y_{ji} distributed independently according to a known distribution $F\left(\cdot|\theta_{ji}\right)$, $\theta_{ji} \overset{\text{iid}}{\sim} G_j, j = 1, 2, \ldots, J$. To account for dependency among G_j's, Dunson (2006) (see also Müller et al., 2004) proposed a Bayesian *dynamic mixture Dirichlet process* (DMDP) in which G_j shares features with G_{j-1} but additionally some innovation may also occur. He introduced this dependence by modeling

$$G_j = (1-\epsilon_{j-1})G_{j-1} + \epsilon_{j-1}H_{j-1} = w_{j1}G_1 + w_{j2}H_1 + \cdots + w_{jj}H_{j-1}, \qquad (2.4.34)$$

where $w_{jl} = \epsilon_{l-1}\prod_{m=l}^{j-1}(1-\epsilon_m)$, $l = 1, \ldots, j$ are weights on the different components in the mixture with $\epsilon_0 = 1$ and H_l is the innovation distribution, $l = 1, \ldots, j-1$ taken each to be a DP. As can be seen, this model randomly reduces the probability attached to the atoms of G_{j-1} by a factor of $(1 - \epsilon_{j-1})$

and adding new atoms drawn from the nonatomic base distribution of H_{j-1}. ϵ_{j-1} controls the amount of the expected change from $j-1$ to j, with $\epsilon_{j-1} \to 0$ signifying no change. By assuming G_1 to be drawn independently from $\mathcal{D}(M^*, G_0)$, where G_0 is nonatomic, and $H_l \sim \mathcal{D}\left(M_l^*, H_{0l}\right)$, $l = 1, \ldots, j-1$, Dunson derives a formula for the correlation coefficient between G_j and G_{j-1} and gives details of posterior computation via MCMC algorithm.

This model has the drawback that mixture components can only be added over time thus resulting in more components at later times. Also by assuming $H_l \sim \mathcal{D}\left(M_l^*, H_{0l}\right)$, one will end up in adding more and more atoms at each stage. Ren et al. (2008) combined the dynamic modeling as well as hierarchical structure and proposed an extension of the HDP model called the *dynamic hierarchical Dirichlet Process* (dHDP) which has dynamic mixture model and HDP model each as a special case. The rationale for this extension is as follows. The HDP model shares "statistical strength" across different groups of data through sharing the same set of atoms (may refer to as discrete parameters), but the mixing weights associated with different atoms are varied. Furthermore, it is assumed that the data groups are exchangeable. However in certain applications such as the above or seasonal market data analysis, data is collected in a time sequential manner which justifies introduction of dynamic models to reflect this important temporal information. In such a scenario, the group data exchangeability assumption is obviously no longer valid. Therefore, they modified Dunson's model as follows:

$$G_j = (1 - w_{j-1})G_{j-1} + w_{j-1}H_{j-1}, j = 2, \ldots, J, \tag{2.4.35}$$

where $G_1 \sim \mathcal{D}(M^*, G_0)$, the innovation distribution H_{j-1} is drawn from $\mathcal{D}\left(M_j^*, G_0\right)$ instead of $\mathcal{D}\left(M_j^*, H_{0j}\right)$ and $\overline{w}_{j-1} \sim \text{Be}\left(a_{w(j-1)}, b_{w(j-1)}\right)$. This way, G_j is modified from G_{j-1} by the introduction of a new H_{j-1} and the variable $\overline{w}_{j-1}$ controls the probability of innovation. By choosing the prior for H_{j-1} as $\mathcal{D}\left(M_j^*, G_0\right)$ they eliminate the prospect of adding new atoms at every stage, as was the case in Dunson's model, and the same atoms of G_0 are used but weights differ. Additionally to ensure G_0 is discrete, a hierarchy is introduced by assuming G_0 to be drawn from $\mathcal{D}(\gamma, H)$.

The distribution G_0 has SB representation $\sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$, where ϕ_k are the global atoms drawn iid H and $\{\beta_k\}_{k=1}^{\infty}$ are $SBW(\gamma)$ defined as usual. Hence we can rewrite $G_1 = \sum_{k=1}^{\infty} \pi_{1k}\delta_{\phi_k}$, $H_1 = \sum_{k=1}^{\infty} \pi_{2k}\delta_{\phi_k}, \ldots, H_{J-1} = \sum_{k=1}^{\infty} \pi_{Jk}\delta_{\phi_k}$, where the weights $\boldsymbol{\pi}_j$ are independent given $\boldsymbol{\beta}$ and are shown to be distributed in Teh et al. (2006) as $\boldsymbol{\pi}_j | M_j^*, \boldsymbol{\beta} \sim \mathcal{D}\left(M_j^*, \boldsymbol{\beta}\right)$. When substituted in the above expression, G_j can be written as

$$\begin{aligned} G_j &= \prod_{l=1}^{j-1}(1 - \overline{w}_l)G_1 + \sum_{l=1}^{j-1}\left\{\prod_{m=l+1}^{j-1}(1 - \overline{w}_m)\overline{\omega}_l H_l\right\} \\ &= w_{j1}G_1 + w_{j2}H_1 + \cdots + w_{jj}H_{j-1}, \end{aligned} \tag{2.4.36}$$

where $w_{jl} = \overline{w}_{l-1}\prod_{m=l}^{j-1}(1-\overline{w}_m)$, $l = 1,\ldots,j$ with $\overline{w}_0 = 1$. For each $\mathbf{w}_j = \left(w_{j1},\ldots,w_{jj}\right), \sum_{l=1}^{j} w_{jl} = 1$. $\mathbf{w}_j$ is the prior probability that the data in group j will be drawn from the mixture distribution of $G_1, H_1, \ldots, H_{j-1}$. If all $\overline{w}_j = 0$, then all the groups share the same distribution G_1 and the model reduces to a Dirichlet mixture model. On the other hand, if all $\overline{w}_j = 1$, the model reduces to the HDP. Therefore, this model is more general and encompasses both DP and HDP models. The correlation coefficient of the distributions between two successive groups is computed as, for $j = 2,..,J$,

$$\text{Corr}\left(G_{j-1}, G_j\right) = \frac{\sum_{l=1}^{j-1} \frac{w_{jl}w_{j-1l}}{1+M_l^*} \cdot \frac{M_l^*+\gamma+1}{\gamma+1}}{\left[\sum_{l=1}^{j} \frac{w_{jl}^2}{1+M_l^*} \cdot \frac{M_l^*+\gamma+1}{\gamma+1}\right]^{\frac{1}{2}} \left[\sum_{l=1}^{j-1} \frac{w_{j-1l}^2}{1+M_l^*} \cdot \frac{M_l^*+\gamma+1}{\gamma+1}\right]^{\frac{1}{2}}}.$$

Similar to the alternate form of the HDPM expressed in terms of conditional probabilities, dHDP can also be stated as

$$\begin{aligned} &\boldsymbol{\beta}|\gamma \sim \text{SBW}\left(\gamma\right), \boldsymbol{\pi}_j|(M_j^*, \boldsymbol{\beta}) \sim \mathcal{D}\left(M_j^*, \boldsymbol{\beta}\right), \ \overline{w}_j|a_{wj}, b_{wj} \sim \text{Be}\left(a_{wj}, b_{wj}\right), \\ &r_{ji}|\overline{\mathbf{w}}_j \sim \mathbf{w}_j, z_{ji}|\boldsymbol{\pi}_{1:j}, r_{ji} \sim \boldsymbol{\pi}_{r_{ji}}, \phi_k|H \sim H, \ y_{ji}|(z_{ji}, \{\phi_k\}_{k=1}^{\infty}) \sim F\left(\cdot|\phi_{z_{ji}}\right), \end{aligned} \tag{2.4.37}$$

where r_{ji} is a variable to indicate which mixture distribution is taken from $\boldsymbol{\pi}_{1:j} = \{\pi_k\}_{k=1}^{j}$ to draw the observation y_{ji} and z_{ji} is the parameter component indicator.

A modified block Gibbs sampler (Ishwaran and James 2001) is used to carry out the posterior computation as follows:

1. Generate $\overline{w}_l$ from the updated conditional distribution

$$\text{Be}\left(a_w + \sum_{j=l+1}^{J} n_{j,l+1}, \ b_w + \sum_{j=l+1}^{J}\sum_{h=1}^{l} n_{jh}\right),$$

   where $n_{jh} = \sum_{i=1}^{N_j} I\left[r_{ji} = h\right]$. For simplicity, a_{wj} and b_{wj} are taken as a_w and b_w, respectively, for all j.
2. Generate $\overline{\pi}_{lk}$ from the updated conditional distribution

$$\text{Be}\left(M_l^*\beta_k + A, \ M_l^*\left(1 - \sum_{l=1}^{k}\beta_l\right) + B\right),$$

   where

$$A = \sum_{j=1}^{J}\sum_{i=1}^{N_j} I\left[r_{ji} = l, z_{ji} = k\right], B = \sum_{j=1}^{J}\sum_{i=1}^{N_j}\sum_{k'=k+1}^{K} I\left[r_{ji} = l, z_{ji} = k'\right].$$

3. Update indicator variables r_{ji} and z_{ji} by generating samples from multinomial distributions

$$\mathcal{P}\left(r_{ji}=l|\ldots\right) \propto \overline{w}_{l-1}\prod_{m=l}^{j-1}\left(1-\overline{w}_m\right)\overline{\pi}_{lz_{ji}}\prod_{q=1}^{z_{ji}-1}\left(1-\overline{\pi}_{lq}\right).p\left(x_{ji}|\phi_{z_{ji}}\right), l=1,\ldots,J,$$

$$\mathcal{P}\left(z_{ji}=k|\ldots\right) \propto \overline{\pi}_{r_{ji},k}\prod_{k'=1}^{k-1}\left(1-\overline{\pi}_{r_{ji},k'}\right)p\left(x_{ji}|\phi_k\right), k=1,\ldots,K.$$

These posterior distributions are appropriately normalized—the first such that $\sum_{l=1}^{j}\mathcal{P}\left(r_{ji}=l\right)=1$ and the second by a constant $\sum_{k=1}^{K}\mathcal{P}\left(z_{ji}=k\right)$.

The remaining variables in the above expression are sampled in the same way as in HDP model (Teh et al. 2006).

2.4.2.5 Time-Varying Dirichlet Process

Caron et al. (2007) introduce a class of *time-varying Dirichlet process* mixtures which ensure that at each time point the random distribution follows a DPM model. Thus the distribution of observations evolves over time instead of being fixed.

Suppose that at each discrete time point $t=1,2,\ldots$, we have a sample of n observations denoted by $y_t=\{y_{t1},\ldots,y_{tn}\}$ which are iid samples from $F_t(\cdot)=\int_{\Theta}f(\cdot|\theta)\,dG_t(\theta)$, where $f(\cdot|\theta)$ is the pdf and G_t is a mixing distribution assumed to be, $G_t\sim\mathcal{D}(M^*,G_0)$, and has the following infinite sum representation $G_t=\sum_{k=1}^{\infty}\beta_{kt}\delta_{\phi_{kt}}$, where β_{kt}'s are SBW(M^*), and $\phi_{kt}\overset{\text{iid}}{\sim}G_0$. That is, it is the following hierarchical model:

$$G_t|M^*,G_0\sim\mathcal{D}\left(M^*,G_0\right),\ \theta_{kt}|G_t\overset{\text{iid}}{\sim}G_t,\ y_{kt}|\theta_{kt}\sim f\left(\cdot|\theta_{kt}\right). \tag{2.4.38}$$

We may also integrate out G_t and state

$$\phi_{kt}\overset{\text{iid}}{\sim}G_0 \text{ and } y_{kt}|\phi^*_{kt}\sim f\left(\cdot|\phi^*_{kt}\right), \tag{2.4.39}$$

where ϕ^*_{kt} are distinct values among ϕ_{kt}'s.

This model is based on a simple generalized Polya urn model adopted by changing the number and locations of clusters over time. However, they are computationally difficult to implement.

There are other dynamic models discussed in the literature.

2.4.2.6 Probit Stick-Breaking Processes

If in the SB representation of an RPM P, we replace $V_j \sim \text{Be}(1, M)$ by $V_j = \Phi(v_j)$ and $v_j \sim N(\mu, \sigma^2)$, we say that P follows a probit SB process with baseline distribution F_0 and shape parameters μ and σ, denoted by PSBP(μ, σ, F_0). That is, the beta distribution of V_j is replaced by a probit model. PSBP has been discussed by Rodriguez and Dunson (2011), nested DP by Rodriguez et al. (2008), and local DP by Chung and Dunson (2011). The support of PSBP with respect to the topology of pointwise convergence is the set of absolutely continuous measures with respect to the baseline measure F_0.

2.4.3 Some Further Generalizations

The Dirichlet process used in the above mixture and hierarchical models 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. Similarly the DP can be replaced with an RPM of the form (1.3.2) encountered in *species sampling models* developed by Pitman (1995, 1996a; 1996b). Since the DP is a normalized gamma CRM, it gives a clue to replace the same with a normalized CRM. Mixture models with normalized random measure priors are presented in Favaro and Teh (2013). 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 becoming feasible.

Parallel to the HDP, Thibaux and Jordan (2007) proposed a *hierarchical beta process* to be discussed in Sect. 4.6. In features modeling (Sect. 4.8), 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 in a subject. 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, $Z_i \overset{\text{iid}}{\sim} BeP(B)$ and $B \sim BP(c, B_0)$. In modeling documents by the set of words or phrases they contain, assume that the documents are classified into K categories, $A_1, \ldots, A_K$ and Z_{ij} associated with the jth document in ith category, $i = 1, \ldots, K$, is a vector of binary responses with p^i_ω being the probability (specific to the ith 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.

Remark 2.19 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 Chaps. 6 and 7. 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 the parameters of $PD(\alpha, \theta)$.

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.

2.5 Some Related Dirichlet Processes

In applications, some variants of the DP are introduced in the literature. Briefly, they are as follows.

Dirichlet-Multinomial Process

In the context of Bayesian inference for sampling from a finite population, Lo (1986) defines a finite dimensional process. Assume that $F \sim \mathcal{D}(\alpha)$ and given F, $X_1, \ldots, X_N$ is an 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 $F \sim \mathcal{D}(\alpha)$ 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 process as $N \to \infty$. It also appears in connection with Bayesian bootstrap (Lo 1987).

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 $P_j \sim \mathcal{D}(\alpha_j)$, $j = 1,\ldots,k$. 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 described in Sect. 7.5.3).

Generalized Dirichlet Process

Hjort (1990) defines a prior for the distribution function F via its cumulative hazard function H. Let H be a beta process prior with parameters $c(\cdot)$ and $H_0(\cdot)$, that is symbolically, $H \sim Be\{c(\cdot),\ H_0(\cdot)\}$, 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. 4.5. 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 $H \sim Be\{c(\cdot),\ H_0(\cdot)\}$, then $B = -\log(1-F)$ is a Lévy process (independent nonnegative increments process) with Lévy representation

$$\mathcal{E}(e^{-\theta B(t)}) = \left(\prod_{j:t_j\le t}\mathcal{E}(1-S_j)^{\theta}\right)\exp\left\{-\int_0^{\infty}\{\psi(c(s)+\theta)-\psi(c(s))\}\right\}c(s)dH_{0c}(s), \tag{2.5.1}$$

where $\psi(x)$ is the digamma function $\Gamma'(x)/\Gamma(x)$. Various properties of the Dirichlet process developed in Sect. 2.1 and applications may be investigated for the generalized Dirichlet process as well.

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

$$B(x;k,F)=\sum_{j=0}^{k}F\left(\frac{j}{k}\right)\binom{k}{j}x^{j}\left(1-x\right)^{k-j}. \tag{2.5.2}$$

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. Its derivative can be written as a mixture of beta densities $\sum_{j=1}^{k}w_{j,k}\mathrm{Be}(x:j,k-j-1)$ where

$$w_{j,k}=F\left(\frac{j}{k}\right)-F\left(\frac{j-1}{k}\right),j=1,2,\ldots,k,F(0)=0, \tag{2.5.3}$$

$w_{j,k}\geq 0$ 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 its 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, $F\sim\mathcal{D}(\alpha)$, 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.

Chapter 3
Ferguson–Sethuraman Processes

3.1 Introduction

In this chapter, we describe briefly several new processes which have their origin in Ferguson (1973) and Sethuraman (1994) countable sum mixture representations 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 they have revolutionized the nonparametric Bayesian approach to modeling and statistical inference.

Ferguson's formal definition of the DP, while elegant and concise had limited applicability. Given a random sample, the posterior is also a DP and all we had to do was to update the parameter. However, if the sample was distorted or right censored, the posterior was no longer a DP but a mixture of DPs as noted earlier. Similarly, it does not give any clue as to how to perform Bayesian analysis if we wanted to incorporate covariates, or have spatial data, or are dealing with complex mixture models for which the posterior does not have any simple form. His alternative definition defined the DP as a countable mixture of point masses where the weights were derived via normalized increments of a gamma process. But since it required evaluation of an infinite sum, it was not conducive for practical applications.

Sethuraman (1994) showed that the weights can be constructed without having to evaluate an infinite sum via the stick-breaking process involving a series of beta random variables and showed that the two representations are equivalent in distribution:

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot) \stackrel{d}{=} \sum_{j=1}^{\infty} P_j \delta_{\xi_j}(\cdot), \tag{3.1.1}$$

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1_3

where P_j's and p_j's are weights in Ferguson and Sethuraman representations, respectively, defined earlier in Sect. 2.1.1, independent of $\xi_i \overset{\text{iid}}{\sim} H$. This remarkable result has had far reaching impact on Bayesian analysis of complex models as will be seen later in this chapter. However, the infinite sum poses a problem in applications. Ishwaran and Zarepour (2000) provided a partial solution to this problem by truncating the sum to a finite number of terms N as an approximation to the DP, which makes sense since the weights p_j's decay rapidly. This idea of truncation was cast into a broader formulation by Ishwaran and James (2001) who define a larger class of priors, $\mathcal{P}_N(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{N} p_j^* \delta_{\xi_j}(\cdot)$, called *stick-breaking priors,* where N could be finite or infinite, and the weights constructed via the stick-breaking construction, but with one difference—the stick-breaking ratios V_j were taken to be independent beta random variables with parameters $a_i \geq 0$ and $b_i \geq 0$. When N is finite, V_N is taken to be one so that $\sum_{j=1}^{N} p_j^* = 1$. Besides the DP, this class includes some well-known processes to be described later.

Ferguson and Sethuraman representations provide a clue to consider a more general class of random probability measures (RPMs). Let P be a random probability measure defined on $(\mathfrak{X}, \mathcal{A})$ as

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot), \;\; p_j \geq 0, \, j = 1, 2, \ldots, \;\; \sum_{j=1}^{\infty} p_j = 1, \tag{3.1.2}$$

a countable mixture of unit masses placed at (random) points $\xi_1, \xi_2, \ldots$ with random weights, $p_1, p_2, \ldots$, which need not be constructed according to the stick-breaking construction (Ferguson's weights were not). We call P a *Ferguson–Sethuraman* process. (Note that we could replace ∞ in the above definition by N, $1 \leq N \leq \infty$ as in the class $\mathcal{P}_N$, but except for some finite dimensional priors, most of the well-known priors anyway are defined when $N = \infty$. Also, the mass at ξ_j need not be restricted to unity (with corresponding rebalance of weights), but such generalizations may not have much utility in practice. Therefore it seems hardly worthwhile to consider such generality.) Discrete random measures may also be defined with the condition $\sum_{j=1}^{\infty} p_j < \infty$.

The phrase, stick-breaking is generally 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), or normalized increments of independent increment processes (say a gamma process), where the p's need not be constructed using the stick-breaking construction and yet have the above representation, I 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 me as more appropriate. The remarkable feature is that it not only encompasses the $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ family but also includes many more additional processes. They all are discrete and some have SB construction representation as well. The manner in

which the weights are assigned or constructed and locations are defined, determine different processes, as will be seen in this chapter.

A Ferguson–Sethuraman process 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 distributions have emerged to serve as priors. In fact, most of the prior processes currently defined in the literature belong to this family. 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 Dirichlet processes; and hierarchical and mixture processes such as the Kernel based stick-breaking processes. These and other processes are the subject matter of this chapter.

Two immediate types of Ferguson–Sethuraman processes 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 Sethuraman representation. By truncating the sum to a finite number of terms, it yields a class of discrete distribution priors (see, for example, Ishwaran and James 2001; Ongaro and Cattaneo 2004) and provides a way to approximate the Dirichlet process for implementing MCMC algorithm in generating a sample from the Dirichlet process. This in turn has facilitated analysis of complex hierarchical and mixture models in practice.

Recall that Ferguson's countable mixture representation uses normalized increments of the gamma process to construct the weights p_j's. This could be generalized further by replacing them with normalized increments of completely random measures (see, for example, Regazzini et al. 2003; Favaro and Teh 2013). On the other hand, the Sethuraman representation is based on a series of beta $\mathrm{Be}(1, M)$ random variables. This suggests a natural generalization by taking a more flexible distribution of stick-breaking ratios V_i, the beta distribution with parameters a_i and b_i, i.e., $V_i \sim \mathrm{Be}(a_i, b_i), a_i > 0, b_i > 0, i = 1, 2, \ldots$. 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 such as one- and two-parameter Poisson–Dirichlet and stable processes. However, except for some special cases, there does not seem to be any meaningful interpretation of these parameters. During the last few years, Sethuraman's 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 types of Ferguson–Sethuraman processes 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 RPM studied by Pitman (1996a,b) 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

$$P(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}(\cdot) + \left(1 - \sum_{j=1}^{\infty} p_j\right) Q(\cdot), \tag{3.1.3}$$

where Q is a probability measure corresponding to a nonatomic distribution H, $\xi_j \overset{\text{iid}}{\sim} H$, and the 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 encountered in a sample drawn from a large population. Clearly, if $\sum_{j=1}^{\infty} p_j = 1$, it reduces to the model (2.1.5). However, this model is not popular in mainstream statistics.

The weights p_j's have some interesting features. Ferguson's weights were constructed using a gamma process and they were in decreasing order, $p_1 > p_2 > \cdots$. 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 vector of probabilities, the joint distribution of p_i's or of $\mathbf{p}$ determined by beta random variables V_i's 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 $\overline{\mathbf{p}} = \left(p_{(1)}, p_{(2)}, \ldots\right)$ of $\mathbf{p}$ is the Poisson–Dirichlet distribution (Kingman 1975). Further connections of these weights are discussed below.

The manner in which the four ingredients are specified determine the type of Ferguson–Sethuraman process generated. In this chapter various processes as identified by their originators are described. The selection of these processes was guided by their novelty, variety, and breadth, and not by their importance. Also the discussion is limited to a general introduction suggesting further consultation to the relevant papers for details, computational methods, and illustrative examples. The material of this chapter is broadly grouped under the headings:

(a) Discrete and finite dimensional priors
(b) Dependent Dirichlet Processes
(c) Poisson–Dirichlet processes
(d) Species sampling models

3.2 Discrete and Finite Dimensional Priors

3.2.1 Stick-Breaking Priors $P_N(\mathbf{a}, \mathbf{b})$

Motivated by the Sethuraman representation and construction of two-parameter Poisson–Dirichlet process, Ishwaran and James (2001) define a broad class of priors, $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$, called *stick-breaking (SB) priors,* as follows. An RPM P is said to a *stick-breaking prior* if it has the following a.s. representation:

$$P(\cdot) = \sum_{j=1}^{N} p_j \delta_{\xi_j}(\cdot),\ 1 \le N \le \infty, \tag{3.2.1}$$

where

$$p_1 = V_1, p_i = V_i \prod_{j=1}^{i-1} (1 - V_j), i \ge 2,$$

$$V_i \overset{\text{ind}}{\sim} \text{Be}(a_i, b_i), i = 1, 2, \ldots, (\mathbf{a}, \mathbf{b}) = (a_1, a_2, \ldots, b_1, b_2, \ldots), \tag{3.2.2}$$

and independent of V_i's, $\xi_i \overset{\text{iid}}{\sim} H$, H nonatomic. It is a special case of Pitman's species sampling model (3.1.3), the weights being constructed via the SB construction. In order that the family $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ is well defined, we have to ensure that $\sum_{j=1}^{N} p_j = 1$ a.s. In the case of $N < \infty$, $(\mathbf{a}, \mathbf{b}) = (a_1, \ldots, a_{N-1}, b_1, \ldots, b_{N-1})$ and V_N must necessarily be 1. To check that it is well defined in the case of $N = \infty$, Ishwaran and James (2001) provide the following condition that must be satisfied.

$$\sum_{k=1}^{\infty} p_k = 1 \text{ a.s. iff } \sum_{k=1}^{\infty} \mathcal{E}(\log(1 - V_k)) = -\infty \text{ or equivalently } \sum_{k=1}^{\infty} \log\left(1 + \frac{a_k}{b_k}\right) = \infty. \tag{3.2.3}$$

The expectation and variance of P can easily we calculated, based on independence of V_i, and are as follows (Griffin and Steel 2006).

$$\mathcal{E}(P_N(B)) = \mathcal{E}\left(\sum_{j=1}^{N} p_j\right) \mathcal{E}\left(\delta_{\xi_j}(B)\right) = H(B),\ B \in \mathcal{A}$$

$$\text{Var}(P_N(B)) = H(B)(1 - H(B)) \sum_{i=1}^{N} \frac{a_i(a_i + 1)}{(a_i + b_i)(a_i + b_i + 1)} \prod_{j<i} \frac{b_j(b_j + 1)}{(a_j + b_j)(a_j + b_j + 1)}.$$

If we assume $a_i = 1$ and $b_i = M$ for all i, and $N = \infty$, we recover the variance formula for the Dirichlet process: $\text{Var}(P_N(B)) = H(B)(1 - H(B))/(M + 1)$. The

stick-breaking priors have been used in certain models, such as order-based SB processes, spatial processes, and latent SB processes, discussed later in the chapter.

As stated earlier, by varying the parameters of the beta distribution leads to the following processes. A two-parameter generalization is introduced by Ishwaran and Zarepour (2000) by replacing the distribution $\text{Be}(1,\alpha)$ of V_i in the Sethuraman representation of the DP, with $\text{Be}(a,b), a > 0, b > 0$, that is, if $a_i = a$ and $b_i = b$ for all $i \geq 1$. They call it a *beta two-parameter process.* It is discussed in connection with the approximation of the DP by truncating the infinite sum (3.1.2) 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. 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 $\text{PD}(\alpha,\theta)$ (Pitman and Yor 1997). By setting $\alpha = 0$ and $\theta = \mu$ in $\text{PD}(\alpha,\theta)$, the Dirichlet process $\mathcal{D}(\mu)$ is recovered; $\text{PD}(\alpha,0)$ yields a stable-law process; and $\text{PD}(0,\theta)$ yields the (*one parameter*) *Poisson–Dirichlet* distribution (Kingman 1975). One- and two-parameter Poisson–Dirichlet distributions are discussed in greater detail later as they have emerged as useful distributions with different applications. Other than these special cases, any meaningful interpretation of parameters $(a_i, b_i), i = 1,2,\ldots$ of $\mathcal{P}_N(\mathbf{a},\mathbf{b})$ is not clear if not difficult.

Similar to the Polya urn characterization of the DP, Polya urn characterization of the family $\mathcal{P}_\infty(a,b)$ may be obtained as follows. Suppose we have a sample of size n from the above process, that is, $\theta_i|P \overset{\text{iid}}{\sim} P$, $i = 1,2,\ldots,n$, $P \sim \mathcal{P}_\infty(a,b)$. In addition, assume that $\xi_i \overset{\text{iid}}{\sim} H$, H nonatomic. The sample will have some ties. Let $\theta_1^*,\ldots,\theta_k^*$ be the k distinct observations among the sample and $n_1,\ldots,n_k$ be their multiplicities, respectively, so that $n_1+\cdots+n_k = n$. Integrating out P, the marginal distribution of θ_i's will satisfy the following prediction rule (Pitman 1995)

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \sum_{i=1}^{k} \frac{n_i - a}{b+n}\delta_{\theta_i^*} + \frac{b+ka}{b+n}H. \tag{3.2.4}$$

From this it is clear (Pitman 1995, 1996a,b) that the process can be characterized in terms of a generalized Polya urn scheme. Given $\theta_1,\theta_2,\ldots,\theta_n$, choose θ_{n+1} at the $(n+1)$-th step to be a new observation drawn from H with probability $(b+ka)/(b+n)$ and equal to a previous observation θ_i^* with probability $(n_i - a)/(b+n), i = 1,2,\ldots,k$. Its special case, $a = 0$ and $b = \alpha(\chi)$ corresponds to the Dirichlet process with parameter α and yields the Blackwell and MacQueen (1973) predictive rule.

3.2.2 *Finite Dimensional Dirichlet Priors*

Alternate to the SB priors, a class of finite dimensional priors may be defined by truncating the sum in (3.1.2) above to a finite $N < \infty$, i.e., as $P(\cdot) = \sum_{j=1}^{N} p_j \delta_{\xi_j}(\cdot)$, where the weights need not be constructed via SB construction. If the random weights are defined as above, then the vector $\mathbf{p}_N = (p_1, \ldots, p_N)$ has generalized Dirichlet distribution GD $(\mathbf{a}, \mathbf{b})$, with density given by

$$\left(\prod_{i=1}^{N-1} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\,\Gamma(b_i)}\right) \prod_{j=1}^{N-1} p_j^{a_j - 1} p_N^{b_{N-1}-1} \times \prod_{j=1}^{N-2} \left(1 - \sum_{k=1}^{j} p_k\right)^{b_j - (a_{j+1}+b_{j+1})}, \quad (3.2.5)$$

and it constitutes a conjugate prior to multinomial sampling. However if the random weights are defined as having a Dirichlet distribution, $\mathbf{p}_N \sim D(a_1, \ldots, a_N)$, it can also be represented as having GD $(\mathbf{a}, \mathbf{b})$ with $\mathbf{a} = (a_1, \ldots, a_{N-1})$ and $\mathbf{b} = \left(\sum_{k=2}^{N} a_k, \sum_{k=3}^{N} a_k, \ldots, a_N\right)$. Thus P defines a class of priors named as *finite dimensional Dirichlet priors* by Ishwaran and Zarepour (2000). In particular, if the parameters of the Dirichlet distribution taken as $a_1 = \cdots = a_N = \alpha/N$ for some $\alpha > 0$ (symmetric Dirichlet distribution), then $\mathbf{a} = (\alpha/N, \ldots, \alpha/N)$. P is called a finite dimensional Dirichlet prior with parameter α. This symmetric prior has been used by other authors (for example, Kingman 1975 and Patil and Taillie 1977) in constructing prior distributions. The finite dimensional priors may also be constructed by truncating the infinite sum at N and discarding all terms beyond N and setting $p_N = 1 - \sum_{k=1}^{N-1} p_k$.

Another example of finite dimensional Dirichlet priors is where the weights are constructed using the gamma random variables. Let $p_j = Y_j/Y$, where $Y_j \overset{\text{iid}}{\sim} G(\alpha/N, 1), j = 1, \ldots, N, Y = \sum_{k=1}^{N} Y_k$, and independent of ξ_j, $\xi_j \overset{\text{iid}}{\sim} H$, and define

$$P_N(\cdot) = \sum_{k=1}^{N} \frac{Y_k}{Y} \delta_{\xi_k}(\cdot). \quad (3.2.6)$$

It is a good approximation of the $\mathcal{D}(\alpha H)$ since for a fixed value of $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$, $P_N|\boldsymbol{\xi} \sim \mathcal{D}(\alpha F_0(\boldsymbol{\xi}, \cdot))$, where $F_0(\boldsymbol{\xi}, \cdot) = \frac{1}{N} \sum_{k=1}^{N} \delta_{\xi_k}(\cdot)$ is the empirical measure of $\xi_1, \ldots, \xi_N$. It can also be expressed as a mixture of DPs because $F_0(\boldsymbol{\xi}, \cdot)$ is a random measure. Conditioning on $\boldsymbol{\xi}$, and then integrating we see that

$$P_N(\cdot) \overset{d}{=} \int \cdots \int \mathcal{D}(\alpha F_0(\boldsymbol{\xi}, \cdot))\, dH(\xi_1) \cdots dH(\xi_N).$$

In fact Ishwaran and Zarepour (2000) prove a stronger result that for any real valued measurable function g, $P_N(g) \xrightarrow{D} P(g)$, where $P = \mathcal{D}(\alpha H)$. An equivalent

representation of P_N is

$$P_N(\cdot) = \sum_{k=1}^{N} \frac{Y_k}{Y} \delta_{\xi_k}(\cdot) \stackrel{D}{=} \sum_{j=1}^{N} \left\{ V_j \prod_{i=1}^{j-1} (1 - V_i) \right\} \delta_{\xi_j}(\cdot) \tag{3.2.7}$$

where $V_j \stackrel{\text{ind}}{\sim} \text{Be}(1 + \alpha/N, \alpha(1 - j/N))$ for $j < N$ and $V_N = 1$.

If we take $a = -\alpha/N$, $N \geq n$ and $b = \alpha > 0$ in (3.2.4), it would yield

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \sum_{i=1}^{k} \frac{n_i + \alpha/N}{b + n} \delta_{\theta_i^*} + \frac{\alpha\left(1 - \frac{k}{N}\right)}{b + n} H, \tag{3.2.8}$$

and would represent a sample from P_N, an example of $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$.

Muliere and Tardella (1998) introduced an approximation to the DP by introducing what they call ϵ-*Dirichlet Process* for which, given $\epsilon > 0$, they terminate the infinite sum at N^ϵ defined as $N^\epsilon = \inf\left\{n \in \mathbb{N} : \sum_{j=1}^{n} p_j \geq 1 - \epsilon\right\}$. Now the RPM is defined as

$$P^\epsilon(\cdot) = \sum_{j=1}^{N^\epsilon} p_j \delta_{\xi_j}(\cdot) + \left(1 - \sum_{j=1}^{N^\epsilon} p_j\right) \delta_{\xi_{N^\epsilon+1}}(\cdot).$$

That is the remaining weight is placed on the $(N^\epsilon + 1)$-th atom. An alternative is to distribute the remaining weight on all N^ϵ atoms. They show that

$$\text{Sup}_{A \in \mathcal{A}} \{|P(A) - P^\epsilon(A)|\} \leq \epsilon.$$

Thus as $\epsilon \to 0$, ϵ-DP converges to the usual DP. They also show that $N^\epsilon \sim$ *Poisson* $(-M \log \epsilon)$. Rather than truncate the infinite sum in Sethuraman representation by a finite N as in Ishwaran and Zarepour (2000, 2003), these authors let it guide by the amount of closeness from a particular distribution of interest is desired in advance.

Finite dimensional Dirichlet priors serve as an approximation to $\mathcal{D}(\alpha)$ and is used in computation of posterior distributions. Finite dimensional Dirichlet processes, such as Dirichlet-multinomial and Dirichlet-multivariate, were introduced towards the end of last chapter. They were defined via appropriate modification of the basic definition of the DP. Here they are defined via truncation of the infinite sum representation.

Blocked Gibbs sampler discussed in Sect. 2.4 can be applied for posterior computations in a mixture model (2.4.1) under the prior $P_N(a, b)$ with minor changes. In this efforts an interesting observation of Sethuraman (see Paisley et al. 2010) could prove a boon. He has shown that one could sample $\pi \sim \text{Be}(a, b)$

according to the following stick-breaking construction:

$$\pi = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1} (1 - V_j)\, I[Y_i = 1] \tag{3.2.9}$$

with $V_i \overset{\text{iid}}{\sim} \text{Beta}(1, a+b)$ and $Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(a/(a+b))$ and where $I[\cdot]$ is the indicator function.

3.2.3 Discrete Prior Distributions

It is clear that 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 unifying general class of *discrete prior distributions* Π_d as follows. Let H be a nonatomic probability measure on $(\mathfrak{X}, \mathcal{A})$. An RPM 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, $\xi_i \overset{\text{iid}}{\sim} H$ 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

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

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.

3.2.4 Residual Allocation Models

The weights in the finite case, when viewed as a vector of proportions $\mathbf{p}_N = (p_1, \ldots, p_N)$, 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:

$$v_1 = p_1, v_2 = \frac{p_2}{1-p_1}, v_3 = \frac{p_3}{1-p_1-p_2}, \ldots, v_N = \frac{p_N}{1-p_1-\cdots-p_{N-1}} = 1. \tag{3.2.10}$$

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 $\mathcal{P}(\lim_{n\to\infty}(1 - p_1 - p_2 - \cdots - p_n) = 0) = 1$. Examples of RAM are (Patil and Taillie 1977)

1. The symmetric Dirichlet distribution with parameters N and $\alpha > 0$ is an obvious RAM with $v_i \sim \text{Be}(\alpha, (N-i)\alpha)$.
2. Engen's (1975) model is also a RAM with $v_i \overset{\text{iid}}{\sim} \text{Be}(1, \alpha)$, $\alpha > 0$.
3. *Size-biased permutation* (defined later) of $\mathbf{p}_N$ having a symmetric Dirichlet distribution with parameter α is another RAM with $v_i \sim \text{Be}(\alpha + 1, (N-i)\alpha)$.

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

3.3 Dependent Dirichlet Processes

As is well known, the DPM models are one of the most versatile tools in modeling data. They are a generalization of finite mixture models to the case which allow infinite number of mixture components. One of the assumption is that each sample is generated independently from the same DP. However in practice this assumption may be called into question or considered to be an unrealistic constraint as the samples may come from different DPs. HDP provides a partial solution by assuming them to be drawn independently from a common super (parent) DP. However this is not sufficient when dealing with sequential or time-varying data, where RPMs at contiguous points of observation are assumed to be related. This and other reasons prompt us to consider a family $\mathcal{F}_\chi$ of nonparametric RPMs or random distribution functions, F_x, defined on a common measurable space $(\mathfrak{X}, \mathcal{A})$, and indexed by $x \in \chi$ for some space χ, such that they are mutually dependent. It will be denoted as $\mathcal{F}_\chi = \{F_x : x \in \chi\}$. It is recognized as a stochastic process defined on the space of probability measures over the domain χ. This makes it possible to share information among F_x, across x.

There is an advantage to treat RPMs F_x collectively in making inference. For example, while dealing with data which consist of subgroups, it is reasonable to assume that each subgroup has a distinct but related RPM, say the distributions of treatment outcomes in related health facilities in a multicenter clinical trial. In this case χ has finite number of elements. In sequential time-varying data, χ may have countably infinite number of values $\{t_1, t_2, \ldots\}$. In regression and random effect models, χ would be a set of covariates; and in spatial setting χ may be a subset of

d-dimensional Euclidean space, R^d. Models based on such collection of RPMs are collectively termed as *dependent nonparametric models*. They offer a compromise between two extreme choices: one in which all F_x are assumed to be identical and two, all are distinct and independent.

Earlier reference to dependent Dirichlet processes (DDP) may be found in Cifarelli and Regazzini (1979) and Muliere and Petrone (1993) who defined dependent nonparametric models across related RPMs, by introducing a regression for the base measure of marginal DP distributed RPM F_x, i.e., $F_x \sim \mathcal{D}(M, F_{0x})$. Muliere and Petrone assumed a specific form for F_{0x}, $F_{0x} = N(\beta x, \sigma^2)$.

MacEachern (1999) introduced a general concept to define DDP as an extension of the DP model to the class of dependent RPMs, $\mathcal{F}_\chi$. His motivation was that the traditional linear models are inappropriate for several reasons. One, the conditional distribution of response variable, given a particular explanatory variable need not follow a parametric distribution. Two, the usual assumption of error term following a normal distribution is not always justifiable and excludes distributions with other potential shapes. Compounding the problem is the assumption that it is independent of the explanatory variables. Consequently, he makes a case for a nonparametric approach in accommodating covariates with the error term distribution evolving with the changes in explanatory variables. Recall that in the context of survival data analysis, the most popular nonparametric approach to incorporate covariates is through the Cox model, and its Bayesian analog 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. 4.3).

In Bayesian approach to such modeling, it is desired to put a prior on $\mathcal{F}_\chi$ which allows us to borrow strength across index x, such as learning across various studies. It represents a generalization of placing a prior on a single member F_x. Learning across studies can also be achieved by employing hierarchical models, or in a hierarchical setup for group data, where the RPM for each subgroup is assumed to have been drawn from a single parent prior. However the present approach provides far greater flexibility. A subclass of models are termed as *dependent Dirichlet processes* (DDPs) when marginally each F_x is assumed to have a DP prior. Exception to this includes models when F_x may have other priors such as SB process priors or beta process priors.

Among the desirable properties of a prior on $\mathcal{F}_\chi$, indicated succinctly in Chung and Dunson (2011), are (i) dependence should increase between F_x and $F_{x'}$, as x and x' gets closer; (ii) expressions for the mean, variance for each F_x and covariance should be simple and easily interpretable; (iii) F_x marginally should be a DP; and (iv) posterior computation should be able to be carried out. Most of the models presented here achieve these properties, although (iv) is usually difficult to implement. However, adequate simulation procedures are available to address this challenge.

MacEachern's approach is to use the Sethuraman representation of a DP to accommodate covariate, $x \in \chi$, where $\chi \subset R^d$ is known as the covariate space.

Recall that under this representation, an RPM F can be expressed as $F = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}$, where the location atoms $\xi_j's$ and weights p_j's are defined as usual with mass parameter M, and base measure F_0 defined on $(\mathfrak{X}, \mathcal{A})$. The key idea behind his approach is to introduce dependence across F_x by assuming the distribution of point masses dependent across levels of x besides being independent across j. Accordingly, we replace each ξ_j by ξ_{jx}, and define $F_x = \sum_{j=1}^{\infty} p_j \delta_{\xi_{jx}}$. That is, each ξ_j is replaced by a realization of the stochastic process ξ_{jx}, which, for example, could be a Gaussian Process (GP), indexed by $x \in \chi$. Thus dependence is introduced by linking F_x through dependent location point masses. In this formulation, the locations of point masses have been indexed by the covariates, but the weights p_j are undisturbed. The distribution on $\mathcal{F}_\chi$ is now termed as a *Dependent Dirichlet process*. It is governed by three parameters: M, F_0, and a stationary stochastic process ξ_χ which has F_x as marginal for each x. The parameters F_0 and M may or may not depend on the covariate x. If they do, the marginal distribution of F_x will be a Dirichlet process with parameters M_x and F_{0x}. If they do not, ξ_χ will generally be a stationary stochastic process with continuous path and index set χ.

A simple application of the DDP model is in modeling the residual structure in the linear model. In the traditional linear model $y_i = x_i\beta + \epsilon_i$, the errors are assumed to be drawn iid from a normal distribution with mean 0 and variance σ^2, which does not allow them to evolve with the covariate x. In the DDP modeling, ϵ_i's are assumed to be independent and each $\epsilon_i \sim F_{x_i}$. This approach of inducing dependence is used in the development of ANOVA DDP models (DeIorio et al. 2004) where they considered the index x as a categorical variable to model an ANOVA type dependence of point masses; of spatial models (Gelfand et al. 2005; Duan et al. 2007) in which x is replaced with a stochastic process of random variables where each random variable is associated with a location in some given region; and in constructing latent stick-breaking process (Rodriguez et al. 2010). Having only the locations depend on x constitutes the first kind of extension.

The second kind of extension is to have the random weights p_j also vary with x. This can be accomplished by replacing the individual SB ratios V_j, used in building up p_j, by stochastic processes V_{jx}, $x \in \chi$. In doing so, note that the total mass M will have to be allowed to depend on x as well, since the distribution of V_j involves the parameter M. This will yield a modified parameter M_χ. This approach of inducing dependence through the weights is used in developing time-varying DDP (Nieto-Barajas et al. 2008); and in order-based SB process priors (Griffin and Steel 2006). In addition, Griffin and Steel (2006) use permutation of V_j and Reich and Fuentes (2007) supplement V_j with kernels, as discussed below.

Among the properties of the DDP model, it is pointed out that the prior distribution on $(F_{x_1}, \ldots, F_{x_d})$ has full support as long as the stochastic process ξ_χ is rich; the random marginal distribution follows a DP for each $x \in \chi$; the RPMs F_x are continuous in x, the feature that provides evolving distributions along with changes in covariate; it encompasses a wide spectrum of inference ranging from a nearly parametric ($M_x \to \infty$) to an inference showing a strong dependence between distribution near x; and finally such models are computationally amenable.

By construction, F_x's are discrete. To extend the models to absolutely continuous distributions, they all follow the usual recipe of supplementing the DDP model with a continuous distribution G on unobserved covariate, x. Let $F(y) = \int F_x(y)\, dG(x)$. Thus, although each of F_x is discrete, integrating over a covariate with G produces a continuous F.

It is clear that in place of the DP, other Ferguson–Sethuraman processes, such as SB priors (Ishwaran and James 2001) or Pitman–Yor process (Pitman and Yor 1997) may also be used, affording a great flexibility in modeling.

In this section, we will now describe several dependent nonparametric models. DeIorio et al. (2004) propose a dependent nonparametric model that describes dependence across related random distributions F_x in $\mathcal{F}_\chi$ in an *ANOVA-type* set up. Nieto-Barajas et al. (2012) proposed a *time-series DDP* model, in which dependence of RPMs is introduced over time while still marginally at each time point the RPM is assumed to be a DP. Griffin and Steel (2006) propose *order-based SB prior* models for continuous covariates in which dependence is introduced by allowing weights and locations to depend upon covariates indirectly. By allowing the locations x to be drawn from random surfaces, Gelfand et al. (2005) create random spatial processes. Duan et al. (2007) modify Gelfand et al. model so that the surface selection can vary with the choice of locations. Rodriguez et al. (2010) follow a different route. They no longer build different distributions F_x for each possible value of the index space, but instead construct a stochastic process where observations at different locations are dependent and have a common unknown marginal distribution. They call the resulting process a *latent stick-breaking process.* Chung and Dunson (2011) define a new generalization of the DP called the *Local Dirichlet Process* to allow predictor dependence which enables them to borrow information from neighboring components. Dunson and Park (2008) incorporate covariates through a kernel mixture, first used by Lo (1984) in density estimation. Savitsky and Paddock (2013) extend this approach and develop a DDP model for repeated measure data which is not included here.

All of the above models were defined based on Sethuraman representation and dependence was induced through weights and location atoms. Rao and Teh (2009) follow a completely different route and provide a general framework to construct DDP on arbitrary spaces by way of normalized gamma processes. It is well known that the DP is a normalized completely random measure (CRM) and that a CRM can be constructed via the Poisson process (PP) (Kingman 1967). Therefore it offer an alternative approach to the popular SB construction of the DP. Noting this connection, Lin et al. (2010) give a construction of dependent processes based on compound PPs. In particular, they develop a dynamic process they call as Markov Chain DP which is shown to be suitable as a prior in the case of evolving mixture models. But their approach is for the DDP models which is subsumed by a generalized procedure developed by Foti et al. (2012) for the construction of dependent processes based on *thinning* PPs. The nice thing is that this approach is not restricted to DDP only but can be used for all models that can be represented as CRMs. For example, IBP can be described in terms of a mixture of Bernoulli

processes where the mixing measure, the beta process (Thibaux and Jordan 2007), is a CRM.

The list obviously is not exhaustive and models are included for their novelty and variety. The respective authors provide adequate rationale for developing these models and describe procedures for carrying out analyses with practical examples. Needless to say that these extensions of the DP offer a great deal of flexibility in modeling statistical data.

Computational methods to simulate the posterior distribution which would unable us to carry out Bayesian inference are fairly well crafted and have become common tools during the last decade or two. Therefore, we will mention them briefly in passing and let the reader explore them for details in the original papers where the models were introduced. To facilitate this, we retain the original notations as much as possible.

3.3.1 Covariate Models

3.3.1.1 ANOVA DDP Model

For the case x being a categorical variate, DeIorio et al. (2004) propose a dependent nonparametric model that describes dependence across related random distribution functions F_x in $\mathcal{F}_\chi$ in an ANOVA-type setup. They use the DDP model to introduce dependence in such a way that marginally each random distribution F_x follows a DP. Therefore, each F_x has Sethuraman representation of a countable mixture of point masses. Now the strategy is to assume ANOVA model for these location point masses. The model developed can alternatively be considered as a mixture of ANOVA models with a DP prior on the unknown mixing measure. Here the categorical covariate x could be a k-dimensional vector belonging to the covariate space χ. As an example consider the case, when x is a bivariate variable, say, $x = (v, w)$, $v = 1, \ldots, V$ and $w = 1, \ldots, W$. Let $F_x = \sum_{k=1}^{\infty} p_k \delta_{\xi_{kx}}$ and assume the following additive structure on the locations ξ_{xk}:

$$\xi_{xk} \equiv \xi_{(v,w)k} = m_k + A_{vk} + B_{wk} \tag{3.3.1}$$

with $A_{1k} = B_{1k} \equiv 0$, $m_k \overset{\text{iid}}{\sim} p_m^0(\cdot)$, $A_{vk} \overset{\text{iid}}{\sim} p_{Av}^0(\cdot)$, and $B_{wk} \overset{\text{iid}}{\sim} p_{Bw}^0(\cdot)$, mutually independent across k, v, and w. This is a familiar random effect model in the analysis of variance setup. The weights in the F_x are independent of x and are computed as usual by the SB construction with mass parameter M. The base measure F_0 is a convolution of p_m^0, p_{Av}^0, and p_{Bw}^0, which, for example, could be taken as normal distributions with respective means μ_m, μ_{Av}, and μ_{Bw} and variances σ_m^2, σ_A^2, and σ_B^2. Dependence between two random distribution functions F_x and $F_{x'}$ is induced when x and x' share a common main effect. They refer to this probability model over $\mathcal{F}$ as *ANOVA DDP* model and denote it as $\{F_x : x \in \chi\} \sim \text{ANOVA DDP}(M, F_0)$.

As in the traditional ANOVA model, the structural relationships are defined here through the additive structure and the level of dependence is governed by the variances of p^0's. Marginally, for each $x = (v, w)$, F_x is distributed as DP with mass M and base measure F_x^0. This model allows us to incorporate the notion of main effects and interactions, and adapts the interpretation of the traditional ANOVA model of normal means to the model with unknown random distributions.

The above model can be extended to a k-dimensional categorical covariate $x = (x_1, \ldots, x_k)$ with main effects and interactions, in a straightforward manner and F_x need not be univariate. Other extensions are also indicated.

To counter the effect of DP being a discrete measure, they also introduce a kernel mixture to the ANOVA DDP model, similar to the kernel mixtures described elsewhere. A typical such model would be

$$y_i|(x_i = x) \sim H_x(y_i) \text{ where } H_x(y) = \int f(y|\theta)\, dF_x(\theta)$$
$$\text{and } \{F_x : x \in \chi\} \sim \text{ANOVA DDP } (M, F_0), \tag{3.3.2}$$

where the kernel $f(y|\theta) = N(y|\mu, S)$, with the covariance matrix S is typically used.

For posterior inference it is convenient to view this model as a mixture of ANOVA models, with mixing coefficients following a DP prior. Thus, let d_i denote a design vector to select the appropriate ANOVA model corresponding to x_i and $\alpha_h = \{m_h, A_{2h}, \ldots, A_{Vh}, B_{2h}, \ldots, B_{Wh}\}$ so that $\xi_{xh} = \alpha_h d_i$ for $x = x_i$. With the above specified parameters of prior base measure $\left(p_m^0, p_{Av}^0, p_{Bw}^0\right)$, the model can be rewritten as

$$y_i|(x_i = x) \sim H_x(y_i),\ H_x(y) = \int N(y|\alpha d_i, S)\, dF(\alpha), F \sim \mathcal{D}(M, F_0), \tag{3.3.3}$$

or for posterior simulation, by introducing latent variables α_i

$$y_i = \alpha_i d_i + \epsilon_i, \alpha_i \sim F, F \sim \mathcal{D}(M, F_0) \text{ and } \epsilon_i \sim N(0, S). \tag{3.3.4}$$

Let $\alpha_1^*, \ldots, \alpha_k^*$ be $k \leq n$ distinct elements among $\alpha_1, \ldots, \alpha_n$ with n_j multiplicities of $\alpha_j^*, j = 1, \ldots, k$. Define $s_i = j$ iff $\alpha_i = \alpha_j^*$, the cluster identifier, and let $\Gamma_j = \{i : s_i = j\}$, and η stand for some known, or possibly unknown, hyperparameters of F_0 and S. To implement the posterior distribution simulation, the marginal approach discussed in Sect. 2.4 is used. The conjugate nature of p^0 simplify the simulation. It can be implemented by reformulating the above model as a mixture of ANOVA models. That is, the data y_i are drawn from a mixture of ANOVA models with a DP prior on the unknown mixing distribution. Now the MCMC scheme used in the case of DP mixture models (for example, MacEachern 1998) indicated earlier can also

be extended for posterior simulation in ANOVA DDP models. The steps for Gibbs sampler are

1. Resample s_i from

$$\mathcal{P}\left(s_i = j | s_{-i}, y, \eta, S\right) \propto \begin{cases} n_j^- p\left(y_i | s_i = j, s_{-i}, y_{-i}, \eta, S\right), j = 1, \ldots, k^- \\ M \int N\left(y_i : \alpha_i d_i, S\right) dF\left(\alpha\right) \quad j = k^- + 1, \end{cases}$$

where a_{-i} denotes all elements of vector a except the i-th element a_i,

$$n_j^- = \begin{cases} n_j - 1 \text{ if } j = s_i \\ n_j \quad \text{if } j \neq s_i, \end{cases}$$

and k^- is the number of clusters with α_i removed. If $n_{s_i}^- = 0$, relabel the remaining clusters $j = 1, \ldots, k^- = k - 1$. After sampling s_i, set

$$k = \begin{cases} k^- \quad \text{if} \quad s_i \leq k^- \\ k^- + 1 \text{ if } s_i = k^- + 1. \end{cases}$$

2. Resample α_j^* from its posterior distribution given by

$$p\left(\alpha_j^* | s, y, \eta, S\right) \propto \left[\prod_{i \in \Gamma_j} N\left(y_i : \alpha_j^* d_i, S\right)\right] p^0\left(\alpha_j^* | \eta\right).$$

3. Resample η, if unknown, conditional on the current values of s, α^*, and k, as per standard posterior simulation for the DP mixture models.

This model can be extended to accommodate other features. For example, one could consider a non-linear model or introduce hierarchy in the model. Similarly, it can be extended to the data involving repeated measures.

In this model the weights of the point masses do not depend on the covariate x. More complex models can be constructed by allowing its dependence.

3.3.1.2 Time-Series DDP

Nieto-Barajas et al. (2012) introduced another variation of DDP model in which dependence of RPMs is induced over time while still marginally at each time point the RPM is assumed to be a DP. The dependence is induced by defining multivariate beta distributions for the SB ratio V_j's. This model is suitable to serve as a prior for a time series of RPMs. In this case χ consists of a finite number of values, $t_1, \ldots, t_J$, which without loss of generality, we take here as $1, \ldots, J$. A time-series model is proposed by defining a joint probability distribution on the collection $\mathcal{F} = \{F_t : t = 1, \ldots, T\}$ of RPMs representing time-varying distributions. This is

achieved by making the weights in the Sethuraman representation of F_t vary with time, but dependent, while the atoms remain the same over time. This allows them to specify different strengths of dependence at different stages. This model is different from the ANOVA DDP model where locations change but weights remain the same across t.

Let $F_t = \sum_{j=1}^{\infty} p_{jt}\delta_{\xi_j}$, where p_{jt} are weights specific to F_t and ξ_j are point masses assumed to be common across all t and distributed according to F_0. Now the dependence between F_t and F_{t+1} is implemented by introducing a sequence of latent binomial random variables

$$z_{jt}|V_{jt} \sim \text{Bin}\left(m_{jt}, V_{jt}\right), V_{j1} \sim \text{Be}\left(1, M\right), \tag{3.3.5}$$

and for $t > 1$, replacing the distribution of V_{jt} with

$$V_{jt}|z_{jt-1} \sim \text{Be}\left(1 + z_{jt-1}, M + m_{jt-1} - z_{jt-1}\right), t = 2, \ldots, T. \tag{3.3.6}$$

The joint probability model for $\{F_1, \ldots, F_T\}$ is referred to as *time-series DDP* model and written as $\{F_1, \ldots, F_T\} \sim \text{tsDDP}\left(M, F_0, \mathbf{m}\right)$, where $\mathbf{m} = \left\{m_{jt}\right\}_{j=1,t=1}^{\infty,T}$. The marginal distribution of $V_{jt} \sim \text{Be}\left(1, M\right)$ does not change. This together with $\xi_j \overset{\text{iid}}{\sim} F_0$, imply that $F_t \sim \mathcal{D}\left(M, F_0\right)$. So F_t remains unchanged as a DP. The role of latent variables z_{jt} is to introduce dependence between pairs V_{jt} and V_{jt+1}. Here the parameter m_{jt} governs the level of dependence, larger the value of m_{jt}, greater is the dependence. In fact the correlation between V_{jt} and V_{jt+1} is given by

$$\text{corr}\left(V_{jt}, V_{jt+1}\right) = \frac{m_{jt}}{1 + M + m_{jt}}$$

and is controlled by m_{jt}. As $m_{jt} \to \infty$, the correlation is 1, and $V_{jt} = V_{jt+1}$ with probability 1 and results in equal weights. Doing so for all j and t, we get all RPMs F_t's as identical with probability 1. On the other hand, if all $m_{jt} \to 0$, we get V_{jt} and V_{jt+1} as independent.

The authors also derive correlation between F_t and F_{t+1} and discuss its interpretation. From this it could be concluded that even if V_{jt} is independent of V_{jt+1}, for all j, F_t is still correlated with F_{t+1} due to common atoms ξ's they share. In special case $m_{jt} = m_t$, the correlation between $F_t\left(B\right)$ and $F_{t+1}\left(B\right)$ for $B \in \mathcal{A}$ simplify to

$$\text{corr}\left(F_t\left(B\right), F_{t+1}\left(B\right)\right) = \frac{1 + M}{1 + 2M\left(2 + M + m_t\right)/\left(2 + M + 2m_t\right)}.$$

The simulation of posterior distribution can be performed via the usual MCMC scheme. To sample from the posterior distribution of F_t, a collapsed Gibbs sampler which is similar to the blocked Gibbs sample of Ishwaran and James (2001) is used, in which the process is expressed as Sethuraman representation but truncated at finite number of terms N, and hence it approximates the posterior distribution of

the process. In contrast, the collapsed Gibbs sampler, however, does not involve truncation and is more efficient. It is similar to the retrospective sampling procedure developed by Papaspiliopoulos and Roberts (2008) and is based on marginalizing nonallocated point masses in the SB representation of F_t.

3.3.1.3 Order-Based Stick-Breaking Prior

In contrast to DeIorio et al. (2004), Griffin and Steel (2006) propose models for continuous covariates in which dependence between the distributions is induced by allowing both weights and locations to depend upon covariate indirectly. The rationale is that since an RPM is defined as a function of $\mathbf{V}$ and $\boldsymbol{\xi}$ via the SB construction, it is natural for a F_x defined at a point x to make $\mathbf{V}$ and $\boldsymbol{\xi}$ also depend on x. However, instead of having V_j 's and ξ_j 's depend directly on x, they induce dependence by ranking elements of $\mathbf{V}$ and $\boldsymbol{\xi}$ by an ordering π which depends upon x such that the distributions for similar covariate values are associated with similar ordering and thus will be close. At any covariate x, the random distribution used first is the *Stick-breaking* prior, but later their main focus is the DP prior, since it is not easy to define ranking for an infinite sequence. They define an *order-based stick-breaking prior* on the covariate space χ by a sequence $\left\{a_j, b_j\right\}_{j=1}^{N}$ (where N is potentially infinite), a centering distribution F_0, and a stochastic process $\{\pi(x)\}_{x\in\chi}$ such that the following holds:

$$\left\{\pi_1(x), \ldots, \pi_{n(x)}(x)\right\} \subseteq \{1, \ldots, N\} \text{ for some } n(x) \leq N; \pi_i(x) = \pi_j(x) \text{ iff } i = j.$$

$\pi(x) = \left\{\pi_1(x), \ldots, \pi_{n(x)}(x)\right\}$ with $n(x) \leq N$, is referred to as the *ordering at* x. The random variables ξ_j's and V_j's are independent, $\xi_j \overset{\text{iid}}{\sim} F_0$, and $V_j \sim \text{Be}\left(a_j, b_j\right)$. The ordering π associates each pair $\left(\xi_j, V_j\right)$ with a position in the covariate space and the ordering is defined through ranking of them by their distance from x, with the smallest distance first. In applications they use a Poisson point process Z with intensity λ to generate the ordering in combination with the permutation. The same distribution over the space χ is obtained if $\pi_i(x) = i$ for all $x \in \chi$ and $i = 1, \ldots, N$. The random distribution at $x \in \chi$ is defined as

$$F_x = \sum_{j=1}^{n(x)} p_j(x)\, \delta_{\xi_{\pi_j(x)}}, \tag{3.3.7}$$

where $p_j(x) = V_{\pi_j(x)} \prod_{i<j}\left(1 - V_{\pi_i(x)}\right)$ and for finite $n(x)$, $p_{n(x)}(x) = \prod_{i<n(x)}\left(1 - V_{\pi_i(x)}\right)$. They show $E[F_x(B)] = F_0(B)$ for $B \in \mathcal{A}$. Variance of $F_x(B)$ and correlation between $F_{x_1}(B)$ and $F_{x_2}(B)$ are also derived. If $a_j = 1$ and $b_j = M$ and that $N = \infty$, we recover a DP at any $x \in \chi$ if $n(x) = \infty$. This subclass of processes is called *order-based dependent Dirichlet processes* (πDDP) with parameter M, base distribution F_0, and stochastic process $\{\pi(x)\}_{x\in\chi}$. Since the

DP produces a discrete distribution, the kernel (usually taken as a normal kernel) mixture of DPs model provides a vehicle to generate an absolutely continuous distribution. This model can be used when x is time, space, or a covariate.

Mixtures of order-based dependent Dirichlet processes are also introduced in three different settings: curve fitting, modeling volatility in time-series data, and in spatial modeling. In the case of curve fitting, models used for density estimation (Lo 1984; Escobar and West 1995) can be extended by replacing the DP by πDDP.

As one would expect, there are difficulties in simulating the posterior distribution. The authors use the truncation method of Ishwaran and James (2001). An additional step required is to sample the point process Z and the intensity parameter λ in which Z is also truncated. In contrast to these authors, DeIorio et al. use the Gibbs sampler for the posterior distribution marginalized over the parameters V's, and where needed over the parameters ξ's as well. The updating of parameters follows from that of the DP described in Ishwaran and James (2001). Alternatively, one can use the methods proposed in Walker (2007) and Papaspiliopoulos and Roberts (2008) which alleviates the necessity of truncation.

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 $V_j(x)$ in the stick-breaking construction by a kernel $w_j(x) V_j$. This is similar to the approach of Dunson and Park (2008) discussed elsewhere.

3.3.2 Spatial Models

3.3.2.1 Spatial Dirichlet Process

Earlier $\mathcal{F}_\chi$ was considered as a collection of random probability measures indexed by x, which was treated as a covariate and χ a covariate space. For spatial models, x is replaced with a stochastic process of random variables where each random variable is associated with a location in some given region. Modeling in such spatial context is usually carried out by assuming a stationary Gaussian process for the spatial process. The parameters of the Gaussian process are unknown and hence the process is random. For the Bayesian analog, parameters are assigned certain prior distributions, otherwise the procedure is deemed to be a finite dimensional model.

Gelfand et al. (2005) instead replace the Gaussian process and allow the locations x to be drawn from random surfaces to create a random spatial process. They developed a nonparametric, non-stationary spatial process called the *spatial DP*, and then extend it to a *spatial DP mixture*. To replace the GP, a stochastic process of random variables, $Y(s)$, where each is associated with a location s in a given region $D \subseteq R^d$ is needed. As in the case of Gaussian process, the stochastic process is specified by its arbitrary finite dimensional distribution of the vector

$(Y(s_1), \ldots, Y(s_k))$, where $s^{(k)} = (s_1, \ldots, s_k)$ is a set of k locations in D. This resulting process is obviously non-stationary and the joint distribution need not be normal.

Gelfand et al. start developing the model for point-referenced data assuming the data collected as a sample from a realization of random process $Y_D \equiv \{Y(s) : s \in D\}$ at the specific locations $(s_1, \ldots, s_k)$, and that we have n replications of measurements at these locations. Thus the data is of form $\{y_t(s_1), \ldots, y_t(s_k)\}^T$, $t = 1, \ldots, n$. To mirror the GP, they define a random process F such that marginally at each $s \in D$, $F(Y(s))$ is a DP (think as $F_{Y(s)}$ instead of F_x). So to develop the spatial DP, they start with a DP with parameters, M, and F_0 which is assumed to be continuous, frequently a stationary Gaussian. A random distribution function arising from a $\mathcal{D}(M, F_0)$ has the usual representation $\sum_{j=1}^{\infty} p_j \delta_{\xi_j}$ where p_j's are the weights and ξ_j are iid F_0, and independent of p_j's. (ξ_j could be vectors in which case we obtain a multivariate DP.) Now to model Y's, atoms ξ_j are replaced with a realization of a random process $\xi_{j,D} = \{\xi_j(s) : s \in D\}, j = 1, 2, \ldots$ while leaving the weights undisturbed. The resulting random process F for Y_D has the form $\sum_{j=1}^{\infty} p_j \delta_{\xi_{j,D}}$ and the construction is referred to as a *spatial DP* model. This means that for $s^{(k)}$ as above, F induces a finite dimensional RPM $F^{(k)}$ on the space of distribution functions for the vector $\{Y_t(s_1), \ldots, Y_t(s_k)\}$. Because the weights p_j's are independent of locations $s \in D$, we have $F^{(k)} \sim \mathcal{D}\left(M, F_0^{(k)}\right)$, where $F_0^{(k)}$ is the k-variate distribution for $\{Y_t(s_1), \ldots, Y_t(s_k)\}$ induced by F_0. For example, if F_0 is a Gaussian process, then $F_0^{(k)}$ is a k-dimensional normal distribution. The representation of F indicates that it is a non-stationary process, centered around a stationary process F_0. Note the same surface is used for each realization of $\xi_{j,D}$ from F_0. Covariance structure is thus incorporated thereby pooling information from nearby locations.

In comparison with MacEachern (2000), the distinction highlighted is that there the distributions are dependent such that at each index value the distribution is univariate DP. Here, F induces a random distribution $F(Y(s))$ for each s, hence the set $\mathcal{F}_D \equiv \{F(Y(s)) : s \in D\}$ will be a DDP. (It is possible to get a richer spatial DP by allowing V_j in p_j as well to depend on s, parallel to a more general DDP.) The advantage here is that by the proximity of $\xi_{j,D}$'s, the random marginal distribution $F(Y(s_i))$ of $Y(s_i)$ and $F\left(Y\left(s_j\right)\right)$ of $Y\left(s_j\right)$ given F are such that the difference between them tends to zero as $\left\|s_i - s_j\right\| \to 0$. This way information from neighboring locations can be utilized. They show

$$E(Y(s)|F) = \sum_{j=1}^{\infty} p_j \xi_j(s), V(Y(s)|F) = \sum_{j=1}^{\infty} p_j \xi_j^2(s) - \left[\sum_{j=1}^{\infty} p_j \xi_j(s)\right]^2$$

and the covariance of a pair of sites

$$\mathrm{Cov}\left(Y(s), Y(s')|F\right) = \sum_{j=1}^{\infty} p_j \xi_j(s)\, \xi_j\left(s'\right) - \left[\sum_{j=1}^{\infty} p_j \xi_j(s)\right]\left[\sum_{j=1}^{\infty} p_j \xi_j\left(s'\right)\right].$$

To overcome the discreteness of F, they follow the usual practice and propose a kernel mixture type model in which a pure error process is mixed with respect to F to create a random process F^* that has continuous support. The above model can further be used for prediction of a random realization of the process at locations where Y_D is not observed. Properties of these processes are investigated.

To simulate the posterior distribution, the Gibbs sampling procedure is used.

3.3.2.2 Generalized Spatial Dirichlet Process

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 function for the spatial effects such that the surface selection can vary with the choice of locations and so can the joint selection of surfaces for the n-locations. The marginal distribution of the effect at each site still comes from a DP. Thus their model is a generalization of the above model in which p_j's are also made to vary with locations. They call it as a *generalized spatial Dirichlet Process*. Here an RPM F on the space of surfaces over D is defined such that its finite dimensional distributions is specified as follows. For any collection of sets $A_1, \ldots, A_k$ of $\mathcal{A}$

$$\mathcal{P}\left(Y(s_1) \in A_1, \ldots, Y(s_k) \in A_k\right) = \sum_{i_1=1}^{\infty} \cdots \sum_{i_k=1}^{\infty} p_{i_1,\ldots,i_k} \delta_{\xi_{i_1}(s_1)}(A_1) \ldots \delta_{\xi_{i_k}(s_k)}(A_k), \tag{3.3.8}$$

where i_j stands for $i\left(s_j\right)$, $j = 1, \ldots, k$, ξ_j are iid F_0, 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

$$\mathcal{S} = \left\{ p_{i_1,\ldots,i_k} \geq 0; \sum_{i_1=1}^{\infty} \cdots \sum_{i_k=1}^{\infty} p_{i_1,\ldots,i_k} = 1 \right\}. \tag{3.3.9}$$

The weights must also satisfy the consistency condition

$$p_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_n} = p_{i_1,\ldots,i_{k-1},\cdot,i_{k+1},\ldots,i_n} \equiv \sum_{j=1}^{\infty} p_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_n}$$

and a continuity property. The authors compute the first two moments of $Y(s)$ and the covariance of $Y(s_i)$ and $Y(s_j)$ conditional on the realized distribution F. They use this generalized process to model a random effect model of the type $Y(s) = m(s) + \beta(s) + \epsilon(s)$, where $m(s)$ is a constant term and $\epsilon(s)$ is a Gaussian pure random error term with mean zero and variance σ^2.

3.3.2.3 Latent Stick-Breaking Process

In the previous model, we considered a class $\mathcal{F}$ of RPMs F_x defined for each value x of the index space D and proposed priors on $\mathcal{F}$. Rodriguez et al. (2010) follow a different route. They construct stochastic processes on $D \subset R^d$ and propose prior distributions on these processes having constant marginals at each point $x \in D$, but allow samples from it to be dependent. That is, different distributions F_x for each possible value of the index space is no longer built, but instead construct a stochastic process where observations (random variables) 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 F_x (independent of x) across the index space, but allow samples from it to be dependent. The resulting process is called a *latent stick-breaking process.* It induces a partition of the index space.

Thus the goal in the case of univariate is to construct a stochastic process $Y_D \equiv \{Y(s) : s \in D\}, D \subseteq R^1$ such that, marginally, $Y(s)|F \sim F$ for all $s \in D$, for some unknown F, but $\text{cov}(Y(s), Y(s')|F) \neq 0$. It is sufficient for the purpose to provide a finite dimensional joint distribution of the vector $(Y(s_1), \ldots, Y(s_k))$. For this purpose, they define a stochastic process $U = \{U(s) : s \in D\}$ with uniform marginals defined, and then define $Y(s) = F^-(U(s))$, where F is a discrete probability distribution and F^- is its generalized inverse. This clearly shows that $Y(s) \sim F$ for any given s and $Y(s)$ and $Y(s')$ are correlated. To choose the process U and a prior F, they consider independent sequences $\{z(s) : s \in D\}, \{V_l\}_{l=1}^L$ and $\{\xi_l\}_{l=1}^L$. Here the *latent* Gaussian process z is taken such that marginally $z(s) \sim N(0,1)$ for all $s \in D$ and $\text{corr}(z(s), z(s')) = \gamma(s, s')$. Then $U(s)$ is taken as $\Phi(z(s))$, where Φ is the cumulative distribution of the standard normal variate. Sequences $\{V_l\}_{l=1}^L$ and $\{\xi_l\}_{l=1}^L$ are used to define a random distribution $F = \sum_{l=1}^L p_l \delta_{\xi_l}$, where

$$p_l = V_l \prod_{k<l} (1 - V_k), V_l \sim \text{Be}(a_l, b_l) \text{ for } l < L \text{ and } V_L = 1. \tag{3.3.10}$$

However, as a departure from the usual practice, they impose order restriction in constructing the sequence of atoms $\{\xi_l\}_{l=1}^L$, as follows. $\xi_1 \sim F_0$, and for $l > 1$, $\xi_l \sim F_{0l}$, where F_{0l} is the restriction of F_0 to the set $S_l = \{\xi : \xi > \xi_{l-1}\}$, i.e., $F_{0l}(B) = F_0(B \cap S_l)/F_0(S_l)$ for any measurable set $B \in \mathcal{A}$. This order restriction allows them to generate skewed mixing distributions. Now for any finite set of locations

$s_1, \ldots, s_n$ in D, the joint distribution is given by

$$\mathcal{P}\left(Y(s_1) = \xi_{l_1}, \ldots, Y(s_n) = \xi_{l_n} | \{V_l\}_{l=1}^{L} \text{ and } \{\xi_l\}_{l=1}^{L}\right)$$
$$= \mathcal{P}\left(z(s_1) \in [\Phi^{-1}(\pi_{l_1-1}), \Phi^{-1}(\pi_{l_1})), \ldots, z(s_n) \in [\Phi^{-1}(\pi_{l_n-1}), \Phi^{-1}(\pi_{l_n}))\right), \tag{3.3.11}$$

where $\pi_l = 1 - \prod_{k \le l}(1 - V_k)$ is the proportion of the unit stick assigned to the first l atoms, with $\pi_0 = 0$. The prior distribution induced this way on F is called a *latent stick-breaking process (LaSBP)*, and has parameters $(\{a_l\}_{l=1}^{L}, \{b_l\}_{l=1}^{L}, F_0, \gamma)$. It is written as

$$Y|F \sim F \text{ and } F \sim \text{LaSBP}_L(\{a_l\}_{l=1}^{L}, \{b_l\}_{l=1}^{L}, F_0, \gamma). \tag{3.3.12}$$

Here the joint distribution reflects closeness between $Y(s)$ and $Y(s')$ as opposed to between the distributions at s and s', as is the case in DDP.

In Gelfand et al. (2005) a random distribution function F was constructed for the (whole) process Y_D and it induced the distribution $F(Y(s))$ at a point $s \in D$. Here a single F is constructed for all points in D. Moreover, replication of observations is not required.

For inference purposes, an MCMC algorithm is developed in which $a_l = a$ and $b_l = b$ are set. For the case $L < \infty$, a blocked Gibbs sampler (Ishwaran and James 2001) is used and in case of $L = \infty$, a retrospective sampler of Papaspiliopoulos and Roberts (2008) is used. The algorithm proceeds by sequentially updating the parameters involved.

The *hybrid DP* (hDP) models proposed by Petrone et al. (2009) is related to LaSBP. The authors also use Gaussian copula processes to build dependence across locations. Their model is as follows. Let $Y_i \equiv \{Y(s) : s \in D\}$, $D \subseteq R^1$, $i = 1, \ldots, n$ be random curves defined on D and consider the following model: $\mathbf{Y}_i = \boldsymbol{\theta}_i + \epsilon_i$, $i = 1, \ldots, n$, where ϵ_i are independent realization of a Gaussian process GP $(\mathbf{0}, \sigma^2 I)$, or $\mathbf{Y}_i|\boldsymbol{\theta}_i \sim \text{GP}(\boldsymbol{\theta}_i, \sigma^2)$, where the mean functions $\boldsymbol{\theta}_i$ are specified so as to borrow strength in the estimation by introducing dependence across $\boldsymbol{\theta}_i$'s. They are assumed to be sampled from a probability measure G. Thus we have a mixture of Gaussian processes

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_n | G \sim \int \text{GP}(\cdot|\theta, \sigma^2)\, dG(\theta). \tag{3.3.13}$$

For the random curve $\theta \equiv \{\theta(s) : s \in D\}$ in R^D with probability measure G, denote by $G_{\mathbf{s}}$ the finite k-dimensional distribution of $(\theta(s_1), \ldots, \theta(s_k))$ at coordinates $(s_1, \ldots, s_k)$. It is assumed that the random curves are observed at the same coordinates so that the available data are $Y_i = (Y_i(s_1), \ldots, Y_i(s_k))$, $i = 1, \ldots, n$, with $Y_i(s_j) = \theta_i(s_j) + \epsilon_i(s_j)$. Let $\theta_i = (\theta_i(s_1), \ldots, \theta_i(s_k))$. Then the finite

dimensional characterization of the above model is

$$Y_i|\theta_i \overset{\text{ind}}{\sim} N_k\left(\theta_i, \sigma^2 I_k\right) \text{ and } \theta_i|G_{\mathbf{s}} \overset{\text{iid}}{\sim} G_{\mathbf{s}}, \tag{3.3.14}$$

where $N_k(\cdot,\cdot)$ is the k-variate normal distribution and I_k is the k-dimensional identity matrix. After integrating out θ's, the finite dimensional characterization can be stated as, $Y_i|G_{\mathbf{x}} \overset{\text{iid}}{\sim} \int N_k\left(\theta, \sigma^2 I_k\right) dG_{\mathbf{s}}(\theta)$, a mixture of normal densities. Now the problem becomes familiar—a prior is assigned to $G_{\mathbf{s}}$.

The authors choose a finite dimensional discrete SB prior (Ishwaran and James 2001) for G which has representation $G_N = \sum_{j=1}^{N} p_j \delta_{\theta_j^*}$, with $p = (p_1, \ldots, p_N) \sim D(\alpha_1, \ldots, \alpha_N)$ and independent of p_j's, θ^*'s are iid curve atoms in R^D distributed according to G_0, a nonatomic distribution on R^D. Using the corresponding k-dimensional representation, we have

$$Y_i|G_{\mathbf{x}} \overset{\text{iid}}{\sim} \sum_{j=1}^{N} p_j N_k\left(\theta_j^*, \sigma^2 I_k\right), \tag{3.3.15}$$

a finite dimensional mixture of k-variate normal distribution, a familiar mixture model. Note, here $\theta_j^* = \left(\theta_j^*(s_1), \ldots, \theta_j^*(s_k)\right) \overset{\text{iid}}{\sim} G_{0\mathbf{s}}$. If the Dirichlet distribution is taken to be symmetric with parameters, $\alpha_i = \alpha/N$, it is clear that in limit as $N \to \infty$, the prior G tends to the DP. Extending to the functional case, it can be said that the RPM G_N defined on R^D has a *functional* DP. If we take $N = \infty$, then G_∞ represents the Sethuraman representation of the DP and in that case p_j's are defined as usual using beta random variables and the resulting prior is identified as *functional* DP with parameters α and G_0 as usual.

The authors thus define a prior on the space of probability measures for a random curve by allowing local clustering. They discuss several models, called *hybrid* Dirichlet mixture models for functional data, using this prior and discuss their performances with real and simulated data sets.

However, in LaSBP scalar atoms are used instead of process realizations, which is simpler.

Details of multivariate extension and mixtures of LaSBP along with various properties are available in their paper. For computation purposes, an MCMC algorithm is developed and steps are given for updating the coordinates from full conditionals of the parameters. Application of the above model with practical examples is illustrated in the paper.

3.3.2.4 Local Dirichlet Process

Chung and Dunson (2011) define a new generalization of the DP called the *Local Dirichlet Process* to allow predictor dependence which enables them not only to borrow information from neighboring components, but also yields marginal

DPs. It is a prior distribution on a collection of *RPMs* indexed by predictors. It allows local adoptability while preserving many of the desirable properties of the DDP, (1)–(4) mentioned in the introduction. The RPM at a given predictor value is constructed using weights and atoms of the SB construction located in a neighborhood of the predictor, thus inducing dependence. Their strategy is first to define predictor-dependent SB priors. Define the global sets of locations, weights and atoms, respectively, as follows. $\mathbf{Y} = \{y_i : i = 1, 2, \ldots\}, \mathbf{V} = \{V_i : i = 1, 2, \ldots\}$ and $\boldsymbol{\xi} = \{\xi_i : i = 1, 2, \ldots\}$, where $y_i \overset{\text{iid}}{\sim} H$, a finite positive measure, $V_i \overset{\text{iid}}{\sim} \text{Be}(1, M)$, and $\xi_i \overset{\text{iid}}{\sim} F_0$, and H is a probability measure on $(\chi', \sigma(\chi'))$, χ' Lebesgue measurable subset of Euclidean space R^d. Let $\chi \subset \chi'$ be a given predictor space. For a given $x \in \chi$, let D_x^{ψ} denote a ψ -neighborhood of $x \in \chi$. That is $D_x^{\psi} = \{y : d(x, y) < \psi, y \in \chi'\}$, where $\psi > 0$, and $d(\cdot, \cdot) : \chi \times \chi' \to R^+$ is a given metric. Assume that $H\left(D_x^{\psi}\right) > 0$. Also, for $x \in \chi$, let L_x be a predictor-dependent set indexing the locations belonging to D_x^{ψ} defined by $L_x = \{i : d(x, y_i) < \psi, i = 1, 2, \ldots\}$. Define a set of local random components,

$$\mathbf{Y}(x) = \{y_i : i \in L_x\}, \mathbf{V}(x) = \{V_i : i \in L_x\} \text{ and } \boldsymbol{\xi}(x) = \{\xi_i : i \in L_x\}. \tag{3.3.16}$$

This implies that the sets $\mathbf{V}(x)$ and $\boldsymbol{\xi}(x)$ contain weights and atoms that are assigned to the locations $\mathbf{Y}(x)$ in D_x^{ψ}. Using the local components $\mathbf{Y}(x), \mathbf{V}(x)$, and $\boldsymbol{\xi}(x)$ define

$$F_x = \sum_{l=1}^{N(x)} p_l(x)\, \delta_{\xi_{\pi_l(x)}} \text{ with } p_l(x) = V_{\pi_l(x)} \prod_{j=1}^{l-1} \left(1 - V_{\pi_j(x)}\right), \tag{3.3.17}$$

where $N(x)$ is the cardinality of L_x and $\pi_l(x)$ is the l-th ordered index in L_x. Under the above setup and assumption, they prove (their Lemma 1) that for all $x \in \chi$, $N(x) = \infty$ and $\sum_{l=1}^{N(x)} p_l(x) = 1$ almost surely.

Thus F_x is well defined. Given M, F_0, H, ψ and choice of metric $d(\cdot, \cdot)$, the above formulation defines a *predictor-dependent stick-breaking prior (SBP)* for $\mathcal{F}_\chi = \{F_x : x \in \chi\}$ deemed as *local Dirichlet Process (lDP)*, i.e., an lDP prior with parameters M, F_0, H, ψ is assigned to $\mathcal{F}_\chi$. Symbolically, $\mathcal{F}_\chi = \{F_x : x \in \chi\} \sim \text{lDP}(M, F_0, H, \psi)$. They prove that marginally, F_x is a DP, $F_x \sim \mathcal{D}(M, F_0)$.

The parameter ψ controls the size of neighborhood D_x^{ψ}. If x and x' are close to each other, their respective neighborhoods might overlap and F_x and $F_{x'}$ may share some common atoms resulting in corresponding dependence of F_x and $F_{x'}$. The amount of dependence will decrease if x and x' are farther apart and may even tend to F_x and $F_{x'}$ being independent. While each F_x is a DP, lDP prior approach allows it to vary with predictors and borrow information across local regions of prediction space. Due to the discreteness property of each F_x, the lDP will induce local clustering of subjects according to their predictor values.

The authors compute the correlation coefficient between $F_x(B)$ and $F_{x'}(B)$ for $B \in \mathcal{A}$ and discuss some properties. For computational purposes, an MCMC algorithm based on the blocked Gibbs sampler (Ishwaran and James 2001) is developed and the necessary steps for sampling from the relevant conditional posterior distributions are provided. The paper also contains an application of the lDP to an epidemiologic study.

3.3.2.5 Kernel Based Stick-Breaking Processes

A different way of incorporating the covariates is through the kernel mixture first introduced by Lo (1984) in relation to density estimation (see Sect. 2.5.3). Dunson and Park (2008) follow this line of extension. In the problem of density estimation, Lo (1984) defines a random density function by setting $f(y) = \int k(y,u)\, dF(u)$, where $k(y,u)$ is a known kernel and F is taken to be a random distribution function. By assuming $F \sim \mathcal{D}(M, F_0)$, 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. Now the covariate can be accommodated via letting $f(y|x) = \int k(y,u)\, dF_x(u)$, where F_x is chosen according to the Dependent Dirichlet process discussed above.

Motivated by this possibility, 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, they replace the point mass at ξ_j by a nondegenerate probability measure G_j. Dependence is induced through weighted mixture of independent probability measures. That is, they construct predictor-dependent RPMs via the SB construction in which the weights are expressed as a kernel multiplied by beta random variables. Consider a sequence of mutually independent random components $\{\xi_j, V_j, G_j;\ j = 1, 2, \ldots\}$, where for each j, $\xi_j \sim F_0$ is a location parameter, $V_j \sim \text{Be}(a_j, b_j)$ 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 a *kernel stick-breaking process is* defined as

$$F_x = \sum_{j=1}^{\infty} U_{jx}(\xi_j, V_j) \prod_{i<j} (1 - U_{ix}(\xi_i, V_i))\, G_j. \tag{3.3.18}$$

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. By the argument of Ishwaran and James (2001), it can be concluded that F_x is well defined. The usual weights in SB construction get relatively higher proportion of probability at earlier locations. By replacing them with $K(x, \xi_j) V_j$, allows weights to be rebalanced by proper choice of K so that ξ_j close to x gets relatively higher proportion of SB probability. This way F_x and $F_{x'}$ will have similar probabilities allocated if x and x' are close. In this manner, the kernel SB process accommodates dependence.

There are some obvious special cases. For example, if the kernel $K(x, \xi) = 1$ for all $(x, \xi) \in \chi \times \mathfrak{X}$, so that

$$F_x \equiv F = \sum_{j=1}^{\infty} V_j \prod_{i<j} (1 - V_i) G_j. \tag{3.3.19}$$

Then if $G_j \sim \mathcal{D}(M, F_0)$, independently for each j, we get a stick-breaking mixture of DPs as a prior for F; and if $G_j = \delta_{\xi_j}$, $\xi_j \sim F_0$, then F is assigned a stick-breaking prior as defined by Ishwaran and James (2001). If in addition to $K(x, \xi) = 1$, $a_j = 1$ and $b_j = M$ for all $j = 1, 2, \ldots$, the prior for F is a DP mixture of DPs. Conditional on the sequence $\{\xi_j, V_j;\ j = 1, 2, \ldots\}$, expectation and variance of F_x, and covariance are derived. The authors discuss further properties of the process, clustering and prediction rules. For posterior computation they develop a conditional approach relying on a combination of MCMC algorithm that uses retrospective sampling and generalized Polya urn sampling steps.

3.3.3 Generalized Dependent Processes

All of the models presented in the previous sections were developed using the Sethuraman representation of the DP and different modifications were implemented in its SB construction to obtain various nonparametric dependent processes models. Realizing that the DP can also be constructed as a normalized gamma process, Rao and Teh (2009) pursue an alternate approach in developing DDP. Following this and also noting that the DP is a normalized completely random measure (CRM) (Kingman 1967) that can be constructed using a Poisson process (it may help to review Sect. 4.1 first), Lin et al. (2010) give a different construction of dependent Dirichlet processes based on compound PPs. In particular they constructed a *Markov DP*. Their approach, however, is for the DDP model which is subsumed by a generalized procedure developed by Foti et al. (2012) for the construction of dependent processes based on *thinning* PPs. The nice thing is that this approach is not restricted to DDP only but can be used for all models that can be represented as CRMs. For example, IBP can be described in terms of a mixture of Bernoulli processes where the mixing measure, the beta process (Thibaux and Jordan 2007) is a CRM. This means that Foti et al. method can also be used to construct a family of dependent beta processes (Ren et al. 2011). This interesting alternate approach for developing dependent processes will now be presented in this section.

3.3.3.1 Spatial Normalized Gamma Processes

Rao and Teh (2009) provide a general framework to construct DDP on arbitrary spaces by way of normalized gamma processes. The idea is to define a gamma

process $\mathcal{G}$ over an extended space, slice up the space into different regions and associate a DP with each region. Then define the DP by normalizing the restriction of the gamma process on the associated region. This produces a set of dependent DPs. Dependence is controlled by the amount of overlap among the regions. Since the gamma process is a completely randomized measure, it can be constructed by a Poisson process. Its realization can be expressed as an infinite sum of atomic measures with random weighted point masses, based on the mean measure of the Poisson process.

Let $(\Theta, \sigma(\Theta))$ be a measure space and let T be an index space. We construct the set $\mathcal{F}_T = \{F_t : t \in T\}$ of dependent random measures over $(\Theta, \sigma(\Theta))$ such that each F_t is marginally a DP as follows. Let χ be an auxiliary space and for each $t \in T$, let $\chi_t \subset \chi$ be a measurable set. Consider the product space $\chi \times \Theta$ and define a gamma process $\mathcal{G}$ on the product space with base measure α itself defined on the product space $\chi \times \Theta$. Let $\alpha_t(d\theta) = \int_{\chi_t} \alpha(dx, d\theta)$ be the restriction of α on the set χ_t. α_t is then a measure on Θ for each $t \in T$. Since the restriction of a gamma process is also a gamma process, define $\mathcal{G}_t(d\theta) = \int_{\chi_t} \mathcal{G}(dx, d\theta)$, which is a gamma process with base measure α_t. Now define $F_t = \mathcal{G}_t / \mathcal{G}_t(\Theta)$. Then clearly, $F_t \sim \mathcal{D}(\alpha_t)$. The resulting set $\mathcal{F}_T = \{F_t : t \in T\}$ of dependent DPs is called *Spatial normalized gamma processes*. The degree of dependence of Dirichlet processes, F_s and F_t depends upon the amount of overlapping of gamma processes, $\mathcal{G}_s$ and $\mathcal{G}_t$.

As an example, consider $T = \chi = R^1$, the real line, and $\chi_t = (t - L, t + L)$, interval of length $2L$, for $L > 0$. Also, let M be a concentration parameter and F_0, the base distribution defined on Θ. The base measure of the gamma process $\mathcal{G}$ is $\alpha(dx, d\theta) = dxMF_0(d\theta)/2L$ and yields $\alpha_t = MF_0(d\theta)$. Therefore, $F_t \sim \mathcal{D}(MF_0)$. Each atom in $\mathcal{G}$ has a time point y and a time-span $[y - L, y + L]$ and therefore will appear in the DP F_t as long as $t \in [y - L, y + L]$. Thus two DPs F_s and F_t will share more atoms if s and t are close to each other and no atoms if $|s - t| > 2L$. In fact, the dependence between F_s and F_t is a function of $|s - t|$ and they will be independent if $|s - t| > 2L$.

Although $\mathcal{F}_T$ is a collection of infinite number of DPs, in practice we will come across only a finite number of observations at times say $t_1, \ldots, t_n$.

Like the models covered earlier, this construction also yields marginal DPs. Chung and Dunson's (2009) model come close to this model where the dependence is introduced through spatially overlapping regions and use SB construction instead of normalized gamma processes. For posterior computation the authors give a MCMC procedure involving Gibbs sampler, and also Metropolis–Hastings procedures.

3.3.3.2 Dependent CRMs and Processes

Since the DP is a normalized CRM, and that a CRM can be constructed via the Poisson process (PP), Lin et al. (2010) give a different construction of dependent processes based on compound PPs instead of gamma processes. In particular, they develop a dynamic process they call as *Markov Chain DP* which is shown to be

suitable as a prior in the case of evolving mixture models such that the DP at each step is generated by modifying the previous one to include new information and delete the information no longer useful. But their approach is still for the DDP model which is subsumed by a generalized procedure developed by Foti et al. (2012) for the construction of dependent processes based on *thinning* PPs.

Recall that a CRM defined on the space $(\Theta, \sigma(\Theta))$ with Levy measure ν can be represented by a PP with mean measure ν on the space $\Theta \times R^+$. If ν has infinite mass, then there are infinite number of points of the PP. Then the CRM with Levy measure ν can be represented as $\mu = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$. In the same vein, dependent nonparametric processes can also be described by a PP defined on the augmented space $\chi \times \Theta \times R^+$, where χ is an auxiliary space in addition to Θ being a parameter space and R^+ the space of masses of points of the PP. Foti et al. (2012) use this framework as a basis for dealing with two covariate dependent models, namely the gamma and beta processes, and show that it improves prediction power in nonparametric Bayesian approach to modeling.

The augmented PP for the dependent processes may be written as follows: Let $\Pi = \{(x_i, \theta_i, \pi_i)\}_{i=1}^{\infty}$ be a PP on the space $\chi \times \Theta \times R^+$ with mean measure described by the positive Levy measure $\nu(dx, d\theta, d\pi)$. While the theory holds for this form of Levy measure, we consider the homogeneous case where $\nu(dx, d\theta, d\pi) = G(dx, d\theta)\, \nu_0(d\pi)$. That is, the masses attached to atoms are independent of the location of atoms in $\chi \times \Theta$. In this case the CRM corresponding to dependent processes can be represented as:

$$\mu = \sum_{k=1}^{\infty} \pi_k \delta_{(x_k, \theta_k)} \tag{3.3.20}$$

on the space $\chi \times \Theta$. Let T be some covariate space and let $\{p_x : T \to [0,1]\}_{x \in \chi}$ be a collection of functions induced by $x \in \chi$. For defining a family of measures μ_t dependent on $t \in T$, one proceeds as follows: For each point $(x_k, \theta_k, \pi_k) \in \Pi$, define a set of Bernoulli random variables $\{r_k^t\}_{t \in T}$ taking value 1 with probability $p_{x_k}(t)$, zero otherwise. The variable r_k^t serves as an identifier and indicates whether the atom k in the global measure is included in the local measure μ_t at covariate value t or not. By the marking theorem of PPs, the resulting *thinned* PP can be written as $\Pi_t = \left\{(x_k, \theta_k, \pi_k) \,|r_k^t = 1\right\}_{k=1}^{\infty}$ with mean measure

$$\nu(A, d\theta, d\pi) = \int_A p_x(t)\, \nu(dx, d\theta, d\pi) \text{ for } A \in \sigma(\chi), \tag{3.3.21}$$

and the corresponding infinite sum representation of μ_t is

$$\mu_t = \sum_{k=1}^{\infty} r_k^t \pi_k \delta_{\theta_k}. \tag{3.3.22}$$

μ_t is a CRM defined on Θ that varies with $t \in T$ and has levy measure

$$\nu_{\mu_t}(d\theta, d\pi) = \int_{\chi} p_x(t)\, \nu(dx, d\theta, d\pi). \tag{3.3.23}$$

It is named as *thinned* CRM. The thinning function $p_x(t)$ can take many forms including a bounded kernel and it controls the dependency between μ_t and $\mu_{t'}$. As an example in the case of a single-location thinned CRM they take $p_x(t) = f(|x-t|)$, where $T = \chi$ and $f : \chi \to [0,1]$ a unimodal function. Just as a DP can be constructed by normalizing a gamma process, the authors note that a DDP can be constructed by normalizing a thinned gamma process. If $\chi = T$ and the thinning probability given by $p_{x_k}(t) = \int_0^\infty I[|t-x_k| \le l] f(l)\, dl$ for some distribution $f(\cdot)$ over window size l, then after normalization, it yields the spatial normalized gamma process. Another example of this approach is the *kernel beta process* developed by Ren et al. (2011) and presented in the next chapter, Sect. 4.6.3. Thus it is clear that different Levy measures and thinning functions give rise to different dependent processes.

3.4 Poisson–Dirichlet Processes

The Dirichlet process was defined as an RPM P on quite a general and arbitrary measurable space $\mathfrak{X}$, and Π was the collection of all such measures. Its countably infinite mixture representation was expressed as $P = \sum_{j=1}^{\infty} p_j \delta_{\xi_j}$. 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 on $\mathbb{N}$ or some translation of $\mathbb{N}$. 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 $p_i \ge 0$, $i = 1, 2, \ldots$, 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.

Recall that earlier we have dealt with a sequence $\mathbf{p} = (p_1, p_2, \ldots)$, where the p_i's were constructed via the stick-breaking construction. Namely, $p_1 = v_1, p_n = v_n \prod_{i=1}^{n-1}(1 - v_i)$, $n \ge 2$ and v_i are independent and identically distributed on $[0,1]$. We shall identify this sequence as *SB* sequence. It seems to have first appeared in McCloskey (1965). If in particular, $v_i \overset{\text{iid}}{\sim} \text{Be}(1, \lambda)$, then the sequence is said to have a GEM (λ) distribution. In which case, p_i's were weights in the Sethuraman representation. Now we define two additional sequences, which are permutations of $\mathbf{p}$.

The first one is a rank-ordered permutation of $\mathbf{p}$ in which p_i's are arranged in descending order, and is denoted by $\overline{\mathbf{p}} = \left(p_{(1)}, p_{(2)}, \ldots\right)$. Recall that the weights in Ferguson's alternative definition (see Sect. 2.1) of an RPM P form 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 $\tilde{p}_1 = p_i = p_{i(1)}$ (i.e., $\tilde{p}_1 = p_i$ with probability p_i). Remove $p_{i(1)}$. Now pick p_i at random from the remaining components and set $\tilde{p}_2 = p_i = p_{i(2)}$ (i.e., $\tilde{p}_2 = p_i I[p_i \neq \tilde{p}_1]$ with probability $p_i I[p_i \neq \tilde{p}_1]/\left(1 - \tilde{p}_1\right)$). Continue this way. Then the resulting sequence $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots)$ is said to be a *size-biased random permutation* of $\mathbf{p}$. Formally,

Definition 3.1 Let $p = (p_1, p_2, \ldots)$ be a sequence of real numbers subject to restrictions $p_i \geq 0$, $i = 1, 2, \ldots$, and $\sum_{i=1}^{\infty} p_i = 1$ and define a new sequence $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots)$ as follows:

$$\mathcal{P}\left(\tilde{p}_1 = p_i\right) = p_i, i = 1, 2, \ldots \text{ and for } k \geq 2,$$

$$\mathcal{P}\left(\tilde{p}_{k+1} = p_j | \tilde{p}_1, \ldots, \tilde{p}_k, \mathbf{p}\right) = \frac{p_j I[p_j \neq \tilde{p}_i, \text{ for all } 1 \leq i \leq k]}{\left(1 - \tilde{p}_1 - \tilde{p}_2 -, \ldots, -\tilde{p}_k\right)}, \; j = 1, 2, \ldots. \tag{3.4.1}$$

Then $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots)$ is known as *size-biased permutation of* $\mathbf{p} = (p_1, p_2, \ldots)$.

The objective of this section is to describe the relationship between these sequences and to introduce the distribution of $\overline{\mathbf{p}}$ known as Poisson–Dirichlet distribution (Kingman 1975) involving one parameter, and its two-parameter extension (Pitman and Yor 1997) which is also a two-parameter generalization of the Dirichlet process. These distributions may be considered as priors on Π^*, a subset of Π such that $p_1 \geq p_2 \geq \cdots$. These distributions may also be constructed by the stick-breaking construction and thus are termed as Ferguson–Sethuraman type distributions (atoms are treated as fixed known constants).

3.4.1 One-Parameter Poisson–Dirichlet Process

In quest for the distribution of $\overline{\mathbf{p}}$, 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 nondegenerate 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 $\overline{\mathbf{p}} = \left(p_{(1)}, p_{(2)}, \ldots\right)$, where $p_{(1)} \geq p_{(2),} \geq \cdots$ and where $p_{(k)}$ for $k = 1, 2, \ldots$, represents the proportion of k-th dominant species η_k encountered in a sample draw.

Another way to look at the ordered values $p_{(1)} \geq p_{(2),} \geq \cdots$ 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 $(Y_{(1)}, Y_{(2)}, \ldots,)$. Or the sequence may also be viewed in terms of a Poisson process as follows. Let $T_1 > T_2 > \cdots$ be the points of a Poisson process on $(0, \infty)$ with mean measure $\lambda x^{-1} e^{-x} dx$. Let $P_i = T_i / \left(\sum T_i\right), \; i = 1, 2, \ldots$. Then $P_1 > P_2 > \cdots$. Kingman has shown that the ranked permutation vector $\overline{\mathbf{p}}$ does converge in distribution, but unfortunately its limiting distribution is intractable.

Let $\mathbf{p}_N = (p_1, \ldots, p_N)$ 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 $D(\alpha_1, \ldots, \alpha_N)$, the Dirichlet distribution with parameter vector $(\alpha_1, \ldots, \alpha_N)$, on the $(N-1)$-dimensional simplex. However, Kingman notes that if in particular all α's are the same and equal to α, then p_i's have an exchangeable symmetric distribution $D(\alpha, \ldots, \alpha)$, and $\mathcal{E}(p_i) = N^{-1}$. When α is large, the distribution tends to degenerate at $\left(N^{-1}, \ldots, N^{-1}\right)$, 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 led him to consider the asymptotic case resulting in a distribution called *Poisson–Dirichlet* distribution defined as:

Definition 3.2 (Kingman) The distribution of an infinite sequence $(p_1, p_2, \ldots)$ of real numbers with $p_1 \geq p_2 \geq \cdots > 0$ and $\sum_{i=1}^{\infty} p_i = 1$, a.s. depends only on a single parameter λ and is called *Poisson–Dirichlet* distribution, denoted as PD (λ).

It is difficult to derive PD (λ) directly. Therefore, consider 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 $N\alpha \to \lambda$, it can be shown that $\overline{\mathbf{p}}_N$ has a limiting distribution.

Theorem 3.3 (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 $\left(p_{(1)}, p_{(2)}, \ldots, p_{(n)}\right)$ *converges to the n-dimensional marginal of* $\mathbf{P}_\lambda$.

This theorem exhibits the limiting joint distribution of order statistics $p_{(1)} \geq p_{(2),} \geq \cdots$. Thus PD (λ) defines a prior on Π^* just as the Dirichlet process is a prior on Π. Unlike the DP, however, the finite dimensional distribution of PD (λ) is difficult to describe explicitly. 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 (λ) as long as n is large and α_i are small with their sum $\alpha_1 + \alpha_2 + \cdots + \alpha_n$ being close to λ.

Following Patil and Taillie (1977), Kingman (1993) gives a simple construction of PD (λ) in terms of size-biased sampling. Briefly, let the random probability vector $\mathbf{p}_N = (p_1, p_2, \ldots, p_N)$ have the symmetric Dirichlet distribution $D\ (\alpha, \alpha, \ldots, \alpha)$. Draw components of $\mathbf{p}_N$ by size-biased sampling so that $\mathbf{p}_N$ can be rearranged as a vector $\mathbf{q}_N$ with components expressed as (see the next theorem), $q_1 = v_1$, $q_i = v_i \prod_{j=1}^{i-1} \left(1 - v_j\right)$, $i \geq 2$ and v_i's are beta random variables with appropriate parameters. Now taking the limit $N \to \infty$ and $\alpha \to 0$ so that $N\alpha \to \lambda$, it is shown that $v_i \overset{\text{iid}}{\sim} \text{Be}\,(1, \lambda)$ and that the rank order permutation $\overline{\mathbf{q}} = \left(q_{(1)}, q_{(2)}, \ldots\right)$, has the PD (λ) distribution. This construction forms the basis of two-parameter Poisson–Dirichlet distribution in which the distribution of SB ratios is replaced by more general beta distribution.

Interestingly, the random sequences $\mathbf{p}$, $\overline{\mathbf{p}}$, and $\tilde{\mathbf{p}}$ are in a way related and the stick-breaking representation plays a critical role as the following theorem attest (Pitman 1996a).

Theorem 3.4 (McCloskey) *Let* $\tilde{\mathbf{p}}$ *be a size-biased permutation of a sequence of random variables* $P_1 > P_2 > \cdots > 0$ *with* $\sum P_i = 1$. *Then* $\tilde{p}_j$ *can be represented as*

$$\tilde{p}_j = v_j \prod_{i=1}^{j-1} (1 - v_i) \tag{3.4.2}$$

for a sequence of iid random variables v_j *if and only if the sequence* $\{P_i\}$ *has PD* (λ) *distribution for some* $\lambda > 0$. *The random variables* v_j *then have a common distribution Be* $(1, \lambda)$.

This kind of representation is further explored in Perman et al. (1992) and forms the basis of the definition of the two-parameter Poisson–Dirichlet distribution. They provide formulae which explain why the size-biased permutation of the normalized jumps of a subordinator can be represented, in certain cases, by a stick-breaking scheme defined by independent beta random variables.

The above theorem implies the following.

Corollary 3.5 *Let a sequence* $\{\overline{p}\}$ *with* $p_1 \geq p_2 \geq \cdots$ *be defined by ranking* $\{p\}$ *having* GEM (λ) *distribution. Then*

(i) $\{\overline{p}\}$ *has* PD (λ) *distribution, and*
(ii) $\{p\}$ *is a size-biased permutation of* $\{\overline{p}\}$.

Connection of these sequences 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 $(1, \lambda), \lambda > 0$) 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.

These sequences when interpreted as weights or atom masses of Ferguson and Sethuraman representation of the DP, the relationship among them may be summarized loosely as

$$\mathbf{p} \overset{\text{rank-perm}}{\rightarrow} \overline{\mathbf{p}}(\text{F-weights} \sim \text{PD})$$
$$\overset{\text{SB-perm}}{\rightarrow} \tilde{\mathbf{p}}(\text{S-weights} \sim \text{GEM}) \overset{\text{rank-perm}}{\rightarrow} (\text{F-weights} \sim \text{PD}),$$

where F- and S-weights are weights in Ferguson and Sethuraman representations.

Recall that as a result of Polya urn characterization representation, the conditional distribution of $X_{n+1}|X_1, \ldots, X_n$ was expressed as in (2.1.26) and it was indicated that $\lim_{n\to\infty} (n_j/n) = p_j$ the weight of j-th atom in Sethuraman representation. This along with the above connection of $\overline{\mathbf{p}}$ and $\tilde{\mathbf{p}}$ was expressed in a different light by Pitman (1996a) as follows.

Theorem 3.6 (Pitman) *Let $F \sim \mathcal{D}(MF_0)$ and F_0 nonatomic. Let $\overline{p}_i$ denote the magnitude of i-th largest atom of F and ξ_i be its location in $\mathfrak{X}$. Locations are a.s. distinct. Let X_j^* denote the j-th distinct value observed in a sample $(X_1, X_2, \ldots)$ from F and $p_j^* = F(\{X_j^*\})$ the size of atom of F at X_j^*. Then a.s.*

$$F = \sum_i \overline{p}_i \delta_{\xi_i} = \sum p_j^* \delta_{X_j^*}, \tag{3.4.3}$$

where

(i) $(\overline{p}_1, \overline{p}_2, \ldots)$ *has* PD (M) *distribution;*
(ii) *the* ξ_i *are iid* (F_0)*, independently of* $(\overline{p}_1, \overline{p}_2, \ldots)$*;*
(iii) $(p_1^*, p_2^*, \ldots)$ *is a size-biased permutation of* $(\overline{p}_1, \overline{p}_2, \ldots)$*;*
(iv) $(p_1^*, p_2^*, \ldots)$ *has* GEM (M) *distribution; and*
(v) *the* X_j^* *are iid* (F_0)*, independently of* $(p_1^*, p_2^*, \ldots)$.

The key observation to be made here is that the permutation of weights carries the corresponding locations along as well. This correspondence was exploited in simulation methods discussed earlier in the last chapter.

The vector $\mathbf{p}$ is said to be *invariant under size-biased permutation* (ISBP) if the sequence $\tilde{\mathbf{p}}$ obtained by size-biased permutation of $\mathbf{p}$ has the same finite dimensional distributions as that of $\mathbf{p}$. Characterization of such random distributions that are ISBP is obtained by Pitman (1996b).

McCloskey's result in this regard is that 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 $v_i \overset{\text{iid}}{\sim} \text{Be}(1, \lambda)$, $\lambda \geq 0, i \geq 1$. The "only if" part was established by Pitman (1996b).

A sample of size n drawn from PD (λ) 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 $\sum_{j=1}^{n} j m_n = n$. Let $\mathbf{m} = (m_1, m_2, \ldots, m_n)$. Then probability of observing such a sample is given by

$$\mathcal{P}(\mathbf{m} = \mathbf{m}) = \frac{n! \Gamma(\lambda)}{\Gamma(n + \lambda)} \prod_{j=1}^{n} \left(\frac{\lambda^{m_j}}{j^{m_j} m_j!} \right), \tag{3.4.4}$$

and is known as Ewens's (1972) sampling formula derived in the context of genetics. This formula was independently discovered by Antoniak (1974) (see Eq. (2.1.20)). This formula is generalized when sampling is from a two-parameter Poisson–Dirichlet distribution and is given in Eq. (3.4.7).

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

3.4.2 Two-Parameter Poisson–Dirichlet Process

As noted in Sect. 2.1 (property 21) 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 frequency 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,b) 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(α, θ). 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 $\mathcal{D}(\theta)$; 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 $\text{Be}(1, \lambda)$ are replaced by independent variables V_i such that $V_i \sim \text{Be}(1 - \alpha, \theta + i\alpha)$, for $i = 1, 2, \ldots$. Consonant with the constructive definition of PD (λ), Pitman and

Yor (1997) give the following definition of PD (α, θ) in terms of independent beta random variables.

Definition 3.7 (Pitman and Yor) For $0 \leq \alpha < 1$ and $\theta > -\alpha$, let a probability $\mathbf{P}_{\theta,\alpha}$ govern independent random variables V_k with $V_k \sim$ beta $(1-\alpha, \theta + k\alpha)$ and set $\tilde{p}_1 = V_1$, $\tilde{p}_n = V_n \prod_{k=1}^{n-1}(1 - V_k)$, $n \geq 2$. Further let $p_1 \geq p_2 \geq \cdots$ be the ranked values of $\tilde{p}_1, \tilde{p}_2, \ldots$. Then the Poisson–Dirichlet distribution with parameters (α, θ), denoted as PD (α, θ), is defined to be the $\mathbf{P}_{\theta,\alpha}$ distribution of the sequence $(p_1, p_2, \ldots)$.

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

As parallel to the case of Poisson–Dirichlet distribution, it is shown that under $\mathbf{P}_{\theta,\alpha}$ governing $(V_1, V_2, \ldots)$, the sequence $(p_1, p_2, \ldots)$ is such that $p_1 > p_2 > \cdots$ and $\sum_{n=1}^{\infty} p_n = 1$ a.s. and $(\tilde{p}_1, \tilde{p}_2, \ldots)$ 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 (α, θ) distribution with $0 \leq \alpha < 1$ and $\theta > -\alpha$ and $\tilde{\mathbf{p}}$ a size-biased permutation of $\mathbf{p}$ and V_n defined as $V_n = \tilde{p}_n/(\tilde{p}_n + \tilde{p}_{n+1} + \cdots)$, then the sequences $\mathbf{p}$, $\tilde{\mathbf{p}}$ and $(V_1, V_2, \ldots)$ have the same distributions as those in the definition. $V_n = \tilde{p}_n/(1 - \tilde{p}_1 - , \ldots, -\tilde{p}_{n-1})$ can also be viewed as residual fractions.

Thus as in the case of one-parameter Poisson–Dirichlet distribution, it is shown here that the joint distribution of ranked values of size-biased picks is the Poisson–Dirichlet distribution; and 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; also the random weights in infinite mixture representation of a RPM have similar interpretation as in one-parameter Poisson–Dirichlet distribution.

As PD (α, θ) 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. Unlike the DP, it is not a normalized completely random measure since the variable V_k depends on k. Various applications discussed in Chaps. 6 and 7 may be reworked with replacing the Dirichlet process there with PD (α, θ). Its application in hierarchical modeling is discussed in Teh and Jordan (2010).

Pitman and Yor also give a Poisson process characterization which is essentially a generalization of the same given for the DP mentioned in Sect. 2.1. However, its complexity makes it unfavorable for practical use as was the case for the DP. Whereas, the SB construction is easy to use and as such it is rightly regarded as a tremendous innovation (Ishwaran and James 2001).

Characterization Suppose we have a random sample from an unknown RPM P having a PD (α, θ) distribution, i.e., $X_i | P \overset{\text{iid}}{\sim} P, i = 1, 2, \ldots, n, P \sim$ PD (α, θ). 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).

$$X_{n+1}|X_1, X_2, \ldots, X_n \sim \sum_{i=1}^{K} \frac{n_i - \alpha}{\theta + n} \delta_{X_i^*} + \frac{\theta + K\alpha}{\theta + n} H, \tag{3.4.5}$$

where $X_i^* \overset{\text{iid}}{\sim} H$, H nonatomic. From this it is clear 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 equal to one of the previous observation X_i^* with probability $(n_i - \alpha)/(\theta + n), i = 1, 2, \ldots, K$, and as a new observation with probability $(\theta + K\alpha)/(\theta + n)$. 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. The distinct values $X_1^*, \ldots, X_K^*$ may be treated as clusters. Since the probability of X_{n+1} belonging to the i-th cluster is proportional to its size n_i, and if n_i is small the effect of α will be significant and will tend to keep small clusters relatively small. On the other hand, if the cluster size is large, α being less than 1 will have negligible effect and large clusters will continue to grow. 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 (α, θ) is derived by Pitman (1995) as

$$p(\mathbf{n}) = p_{n,K}(n_1, \ldots, n_K|K) = \frac{\prod_{i=1}^{K-1}(\theta + i\alpha)}{(\theta+1)^{(n-1)}} \prod_{j=1}^{K} (1-\alpha)^{(n_j-1)}, \tag{3.4.6}$$

where as before, $s^{(n)} = \Gamma(s+n)/\Gamma(s) = s(s+1)\ldots.(s+n-1)$. When $\alpha = 0$, it reduces to the formula presented in the section on the DP (property 18).

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, $i = 1, \ldots, n$, 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, $i = 1, \ldots, n$. Then we get (Pitman 1995)

$$\mathcal{P}(\mathbf{m} = \mathbf{m}) = n! \frac{\prod_{i=1}^{K-1}(\theta + i\alpha)}{(\theta+1)^{(n-1)} \prod_{i=1}^{n} m_i!} \prod_{j=1}^{n} \left(\frac{(1-\alpha)^{(j-1)}}{j!} \right)^{m_j}, \tag{3.4.7}$$

known as *Pitman's sampling formula*. This is a two-parameter extension of Ewens's (1972) formula (3.4.4). When $\alpha \to 0$, we get Ewen's formula, which was also discovered by Antoniak (1974) as given in Sect. 2.1, property 20.

Updating Rule Recall that in the case of the DP, given a sample $X_1, X_2, \ldots, X_n$ from F, and $F \sim \mathcal{D}(\alpha)$, the conditional distribution of F was simply $\mathcal{D}(\alpha_n)$ and the updating rule was to change the parameter α to $\alpha_n = \alpha + \sum_{i=1}^{n} \delta_{X_i}$. Similarly, Pitman provides the following updating rule for PD (α, θ). Given a sample $Y_1, Y_2, \ldots, Y_n$ from P and $P \sim \text{PD}(\alpha, \theta)$, the posterior random measure may be approximated by a finite sum

$$P(\cdot|\mathbf{Y}) = \sum_{j=1}^{K} p_j^* \delta_{Y_j^*}(\cdot) + p_{K+1}^* Q(\cdot), \tag{3.4.8}$$

where $Y_1^*, \ldots, Y_K^*$ are as before K distinct observations with multiplicities $n_1, \ldots, n_K$, respectively,

$$\left(p_1^*, \ldots, p_K^*, p_{K+1}^*\right) \sim D(n_1 - \alpha, \ldots, n_K - \alpha, \theta + K\alpha) \tag{3.4.9}$$

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

An in-depth investigation of PD (α, θ) has been carried out by Pitman and Yor (1997). Also related papers by Perman et al. (1992) and Pitman (1995, 1996b) shed more light on this distribution.

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).

Gibbs Sampler for PD (α, θ) Consider the hierarchical model:

$$X_i|Y_i, \varphi \overset{\text{ind}}{\sim} \pi(X_i|Y_i, \varphi), i = 1, \ldots, n$$
$$Y_i|P \overset{\text{iid}}{\sim} P, P \sim \text{PD}(\alpha, \theta), \varphi \sim \pi(\varphi).$$

Integrating out P, one can create a marginalized model

$$\begin{aligned} X_i|Y_i, \varphi &\overset{\text{ind}}{\sim} \pi(X_i|Y_i, \varphi), i = 1, \ldots, n \\ (Y_1, \ldots, Y_n) &\sim \pi(Y_1, \ldots, Y_n) \\ \varphi &\sim \pi(\varphi), \end{aligned}$$

where $\pi(Y_1, \ldots, Y_n)$ denotes the joint distribution for $Y = (Y_1, \ldots, Y_n)$ defined by the underlying Polya urn scheme. Letting $g(p)$ to be any positive integrable function on $(\Pi, \sigma(\Pi))$, Ishwaran and James proved a generalization of Lo's (1984) result as follows:

$$\int_{\Pi} g(P) \mathcal{P}(dP|\mathbf{X}) = \int_{\mathfrak{X}^n} \int_{\Pi} g(P) \mathcal{P}(dP|\mathbf{Y}) \pi(d\mathbf{Y}|\mathbf{X}), \tag{3.4.10}$$

where

$$\pi\left(d\mathbf{Y}|\mathbf{X}\right)=\frac{\pi\left(d\mathbf{Y}\right)\int\prod_{i=1}^{n}\pi\left(X_i|Y_{i,}\varphi\right)\pi\left(d\varphi\right)}{\int\int\prod_{i=1}^{n}\pi\left(X_i|Y_{i,}\varphi\right)\pi\left(d\varphi\right)\pi\left(d\mathbf{Y}\right)} \tag{3.4.11}$$

and $\pi\left(\mathbf{Y}\right)$ is the joint distribution of $\mathbf{Y}$ determined by the generalized Polya urn scheme given above, and integrations are defined over appropriate domains. The posterior distribution $\mathcal{P}\left(dP|\mathbf{Y}\right)$ was given above (3.4.8).

Then the marginal approach for Gibbs sampling algorithm for PD (α,θ) is as follows: To draw values from the posterior distribution $\pi\left(Y,\varphi|X\right)$ of the above model, we draw iteratively from the conditional distributions of

$$\begin{aligned} &Y_i|Y_{-i},\varphi,X,i=1,\ldots,n,\\ &\varphi|Y,X. \end{aligned} \tag{3.4.12}$$

(a) The conditional distributions are given by

$$\mathcal{P}\left\{Y_i\in\cdot|Y_{-i},\varphi,X\right\}=\sum_{j=1}^{m}q_j^*\delta_{Y_j^*}\left(\cdot\right)+q_0^*P\left\{Y_i\in\cdot|\varphi,X_i\right\},$$

where

$$q_0^*\propto\left(\theta+\alpha m\right)\int\pi\left(X_i|Y,\varphi\right)H\left(dy\right),\ q_j^*\propto\left(n_j^*-\alpha\right)\pi\left(X_i|Y_j^*,\varphi\right),j=1,\ldots,m,$$

and $\left\{Y_1^*,\ldots,Y_m^*\right\}$ are distinct values in $\mathbf{Y}_{-i}$ with Y_j^* having frequency $n_j^*,j=1,\ldots,m$.

(b) draw from the density of $\varphi|Y,X$

$$\pi\left(\varphi|Y,X\right)\propto\pi\left(d\varphi\right)\prod_{i=1}^{n}\pi\left(X_i|Y_i,\varphi\right).$$

For the conditional approach, one can use the Blocked Gibbs sampler described earlier in Sect. 2.4 with some obvious modifications.

3.5 Species Sampling Models

While discussing mixture and hierarchical models earlier, it was pointed out that the discreteness of the RPM P under the DP prior served as an advantage in Bayesian analysis of such models. Another area where discreteness also plays a natural role is in ecology and population genetics dealing with sampling from a large population of individuals of various species. This connection was explored in a series of papers by

Jim Pitman and his colleagues. That development was named as *species sampling models* that we briefly describe in this section.

We have seen earlier that if $X_1, \ldots, X_n$ is sample of size n from P and $P \sim \mathcal{D}(\alpha)$, then Ferguson (1973) proved that the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ is given by

$$\mathcal{P}(X_{n+1} \in \cdot | X_1, \ldots, X_n) = \frac{\alpha(\cdot) + \sum_{i=1}^{n} \delta_{x_i}(\cdot)}{\alpha(\mathfrak{X}) + n} = \frac{\alpha_n(\cdot)}{\alpha_n(\mathfrak{X})}, \tag{3.5.1}$$

where $\alpha_n = \alpha + \sum_{i=1}^{n} \delta_{X_i}$. Its connection to the Polya urn scheme was mentioned in Sect. 2.1. Note that here the above prediction rule (3.5.1) for a new observation emerged as a consequence of iid sampling from P having the DP prior. The other way is also true as the Blackwell and MacQueen theorem (Theorem 2.4) shows: the prediction rule (3.5.1) characterizes the DP process.

Blackwell and MacQueen (1973), start with a sequence $\{X_n : n \geq 1\}$ of random variables taking values in $\mathfrak{X}$, along with the following prediction rule

$$\begin{aligned} &\mathcal{P}(X_1 \in \cdot) = \alpha(\cdot)/\alpha(\mathfrak{X}) \text{ and} \\ &\mathcal{P}(X_{n+1} \in \cdot | X_1, \ldots, X_n) = \frac{\alpha_n(\cdot)}{\alpha_n(\mathfrak{X})}, \end{aligned} \tag{3.5.2}$$

and call the sequence as a *Polya sequence with parameter* α. With this tool, they proved their fundamental theorem (see Sect. 2.1) which says that

(a) The sequence $\alpha_n(\cdot)/\alpha_n(\mathfrak{X})$ converges with probability one as $n \to \infty$ to a limiting discrete measure P;
(b) P is the Dirichlet process with parameter α; and
(c) given P, $X_1, X_2, \ldots$ are independent with distribution P.

The conditional probability (3.5.2) is known as the *predictive rule for the Polya urn sequence*. The key point here is that this predictive rule not only produces an exchangeable sequence but also identifies the underlying RPM, as the DP. This interesting result raises a question: are their other type of schemes and predictive rules which lead to an exchangeable sequence, and do they characterize the underlying RPM? Pitman (1996b) and Hansen and Pitman (2000) investigated this question and developed sufficient and necessary conditions in terms of a symmetric nonnegative function called *exchangeable partition probability function* (EPPF). This question has also been explored by Ishwaran and Zarepour (2003) and based on a modified urn scheme they provide a sufficient condition to ensure exchangeability of the resulting sequence, and as a consequence, the limit of the sequence is a discrete RPM. Thus, in a way they generalize Blackwell and MacQueen theorem. They assume α to be nonatomic, although it was not necessary for the Polya sequence.

Since P is discrete a.s., there will be ties among the sample drawn from it with probability one. Denote the $k(n) = k$ distinct values (distinct species when

sampling from a large population) among the sample $X_1, \ldots, X_n$ by $X_1^*, \ldots, X_{k(n)}^*$, in the order they appear and $n_{1n}, \ldots, n_{k(n)n}$ their respective frequencies. This could be thought of as if the distinct values creating a partition of sample labels $\{1, 2, \ldots, n\}$ into k classes or clusters $C_j = \left\{i : X_i = X_j^*\right\}$, with cluster sizes $n_{1n}, \ldots, n_{k(n)n}$. Thus the conditional distribution can be written as

$$\mathcal{P}\left(X_{n+1} \in \cdot | X_1, X_2, \ldots, X_n\right) = \sum_{j=1}^{k} \frac{n_{jn}}{\alpha\left(\mathfrak{X}\right) + n} \delta_{X_j^*}\left(\cdot\right) + \frac{\alpha\left(\mathfrak{X}\right)}{\alpha\left(\mathfrak{X}\right) + n} \overline{\alpha}\left(\cdot\right), \tag{3.5.3}$$

where $\overline{\alpha}\left(\cdot\right) = \alpha\left(\cdot\right)/\alpha\left(\mathfrak{X}\right)$ as before. Note that the point masses associated with X_j^* depend exclusively on the frequency of X_j^* and not on the value of X_j^*, nor on the frequencies of other distinct values X_i^*.

Thus, the rule (3.5.3) can be written in a more general form in terms of the masses which are functions of the cluster sizes. For some nonatomic distribution G_0, let

$$\mathcal{P}\left(X_1 \in \cdot\right) = G_0\left(\cdot\right), \text{ and}$$

$$\mathcal{P}\left(X_{n+1} \in \cdot | X_1, X_2, \ldots, X_n, k\left(n\right) = k\right) = \sum_{j=1}^{k} p_j\left(\mathbf{n}_n\right) \delta_{X_j^*}\left(\cdot\right) + p_{k+1}\left(\mathbf{n}_n\right) G_0\left(\cdot\right), \tag{3.5.4}$$

where $\mathbf{n}_n = \left(n_{1n}, \ldots, n_{k(n)n}\right)$ and $p_j\left(\mathbf{n}\right) = \mathcal{P}\left(X_{n+1} = X_j^* | \mathbf{n}_n = \mathbf{n}\right)$, $1 \leq j \leq k\left(n\right)$ with $\mathbf{n} = \left(n_1, \ldots, n_k\right)$, a vector of frequencies of X_j^*'s, $p_j\left(\mathbf{n}\right) \geq 0$ and $\sum_{j=1}^{k(n)+1} p_j\left(\mathbf{n}\right) = 1$. Equation (3.5.4) is known as the *prediction rule (PR)*. Given that n observations results in the vector $\mathbf{n}$, $p_j\left(\mathbf{n}\right)$ here is the probability that the next observation is the j-th species already observed for $1 \leq j \leq k\left(n\right)$, and $p_{k+1}\left(\mathbf{n}\right)$ is the probability that it is a new species. $p_j\left(\mathbf{n}\right)$ plays a special role in the PR and is some symmetric function of finite sequences of frequencies $\mathbf{n}$ known as the *predictive probability function* (PPF). It provides a clue that a sequence of functions $p_j\left(\mathbf{n}\right)$ with the above constraints determine the distribution of the sequence $\{X_n\}$ via the above predictive rules. For example, in the case of Polya sequence,

$$p_j\left(\mathbf{n}\right) = \frac{n_j}{\alpha\left(\mathfrak{X}\right) + n} I\left[1 \leq j \leq k\right] + \frac{\alpha\left(\mathfrak{X}\right)}{\alpha\left(\mathfrak{X}\right) + n} I\left[j = k + 1\right], \tag{3.5.5}$$

with $n = \sum_{j=1}^{k} n_j$, and in the case of two-parameter Poisson–Dirichlet distribution PD (α, θ),

$$p_j\left(\mathbf{n}\right) = \frac{n_j - \alpha}{n + \theta} I\left[1 \leq j \leq k\right] + \frac{\theta + k\alpha}{n + \theta} I\left[j = k + 1\right]. \tag{3.5.6}$$

Starting with an exchangeable sequence and a predictive rule in general form of (3.5.4), Pitman (1996b) proved the following theorem which characterizes the

general form of RPM G, essentially generalizing the Blackwell–MacQueen theorem proved for the Polya urn sequence.

Theorem 3.8 (Pitman) *Let $\{X_n\}$ be an exchangeable sequence subject to a prediction rule of the form (3.5.4) and let G_n denote the conditional distribution of X_{n+1} given $X_1, X_2, \ldots, X_n$ defined by (3.5.4). Then (1) $G_n \overset{\text{a.s.}}{\to} G$ (in total variation norm) where*

$$G = \sum_j \overline{P}_j \delta_{X_j^*} + \left(1 - \sum_j \overline{P}_j\right) G_0, \tag{3.5.7}$$

and where $\overline{P}_j$ is the frequency of j-th species to appear, i.e., $\overline{P}_j = \lim_{n\to\infty} (n_j/n)$; (2) the $X_j^* \overset{\text{iid}}{\sim} G_0$ and independent of $\overline{P}_j$, and (3) given G, $(X_1, X_2, \ldots)$ is a sample from G.

The key differences in relation to the Blackwell and MacQueen result are that here exchangeability is assumed to start with unlike in Blackwell and MacQueen theorem where it is a consequence, and further the RPM G is not identified here as the DP as is done in the Blackwell and MacQueen paper.

In the context of species sampling, suppose that a random sample $X_1, X_2, \ldots$ is drawn from a large population of individuals of various species, where X_i represents species of the i-th individual drawn, and the j-th distinct species to appear is denoted by X_j^*. When governed by a prediction rule, Pitman (1996b) calls this sequence as a *species sampling sequence* (SSS).

Definition 3.9 An exchangeable sequence $\{X_n\}$ is called a *species sampling sequence (SSS) if the PR* (3.5.4) *holds true and* G_0 is a nonatomic distribution.

In view of the above theorem, it follows that $\{X_n\}$ is a SSS iff $\{X_n\}$ is a sample from a random distribution function G that admits a representation of the form

$$G = \sum_j P_j \delta_{\xi_j} + \left(1 - \sum_j P_j\right) G_0, \tag{3.5.8}$$

for some sequence of random variables $\{P_j\}$ such that $P_j \geq 0$ for all j and $\sum_j P_j \leq 1$ a.s., and some sequence $\xi_j \overset{\text{iid}}{\sim} G_0$, independent of P_j. This shows that the PR of an SSS characterizes an RPM. Such a setup with a random distribution G of the above form and a sample $(X_1, X_2, \ldots)$ from G is called a *species sampling model (SSM)*. P_j can be interpreted as the relative frequency of the j-th species in a population and ξ_j as tag assigned to it. G has an atom P_j at ξ_j for each j such that $P_j > 0$ and rest of the mass distributed according to G_0. The model is said to be *proper* if $\sum_j P_j = 1$

a.s. so that

$$G = \sum_j P_j \delta_{\xi_j} = \sum_j \overline{P}_j \delta_{X_j^*} \tag{3.5.9}$$

where $\overline{P}_j$ and X_j^* are as in (3.5.7) defined in terms of a sample from G.

In general, the RPM G can be defined directly by specifying the weights subject to the condition that they add to one, and the distribution of point masses. It can alternatively defined by specifying the sequence $\left\{p_j(\mathbf{n}) : 1 \leq j \leq k(n) + 1\right\}$ of its PPF as indicated above in (3.5.5) and (3.5.6).

A third alternative to characterize an SSM is through the prior on the sequence of random partitions (Pitman 1996b; Hansen and Pitman 2000). Suppose $\{X_n\}$ is an exchangeable sequence admitting a prediction rule of type (3.5.4) and let Π_n be the partition of $\{1, 2, \ldots, n\}$ generated by the duplicate values of $(X_1, \ldots, X_n)$. Then for each n and $1 \leq j \leq k(n)$, $p_j(\mathbf{n})$ and $p_{k+1}(\mathbf{n})$ are almost surely some functions of Π_n, the induced random partition of $\{1, 2, \ldots, n\}$, with $\mathbf{n}_n = \left(n_1, n_2, \ldots, n_{k(n)}\right)$, a vector of cluster sizes, is also exchangeable. (Gnedin and Pitman (2007) show that any PPF with weights that are functions of the cluster size only must be essentially of form $p_j \propto n_j$.) Therefore, it suffices to specify the probability of $\mathbf{n}_n$ for all possible values of $\mathbf{n}_n$, defining a prior $p(\mathbf{n}_n)$ over the set $\mathbb{N}^* = \cup_{k=1}^{\infty} \mathbb{N}^k$. This prior is known as *exchangeable partition probability function* (EPPF) and must satisfy Kolmogorov consistency criteria,

$$p(\mathbf{n}) = \sum_{j=1}^{k(n)+1} p\left(\mathbf{n}^{j+}\right) \text{ for all } \mathbf{n} \in \mathbb{N}^* \text{ and } p(1) = 1, \tag{3.5.10}$$

where $\mathbf{n}^{j+}$ denotes $\mathbf{n}$ with the j-th component incremented by 1. Thus $p : \mathbb{N}^* \to [0, 1]$.

Converse is also true. Pitman (his Proposition 13) shows that any function regarded as EPPF satisfying (3.5.10), and ν a diffused probability distribution, there is a unique distribution for an SSS $\{X_n\}$ such that p in (3.5.10) is the EPPF of $\{X_n\}$ and ν is the distribution of X_1. Thus he [see also Hansen and Pitman (2000)] gives the following characterization of an exchangeable sequence.

Theorem 3.10 (Pitman) *Given ν a diffused measure and a sequence of functions $\left\{p_j(\mathbf{n})\right\}$ defined on $\mathbb{N}^*$ satisfying $p_j(\mathbf{n}) \geq 0$ and $\sum_{j=1}^{k(n)+1} p_j(\mathbf{n}) = 1$, let $\{X_n\}$ govern by the prediction rule (3.5.4). The sequence is exchangeable iff there exists $p(\mathbf{n})$ defined on $\mathbb{N}^*$ such that $p_j(\mathbf{n}) = p\left(\mathbf{n}^{j+}\right)/p(\mathbf{n})$ holds. Then $\{X_n\}$ is a sample from G and EPPF of $\{X_n\}$ is the unique nonnegative symmetric function $p(\mathbf{n})$ such that $p_j(\mathbf{n}) = p\left(\mathbf{n}^{j+}\right)/p(\mathbf{n})$ holds and $p(1) = 1$.*

PPF can obviously be defined via EPPF as $p_j(\mathbf{n}) = p\left(\mathbf{n}^{j+}\right)/p(\mathbf{n})$. The PPF for the PD (α, θ) distribution is

$$X_{n+1}|X_1, X_2, \ldots, X_n = \sum_{j=1}^{k(n)} \frac{n_{jn} - \alpha}{\theta + n} \delta_{X_j^*}(x_{n+1}) + \frac{\theta + k(n)\alpha}{\theta + n} G_0(x_{n+1}),$$

and EPPF reduces to

$$p(\mathbf{n}) = \frac{\Gamma(\theta + 1)}{(\theta + k(n)\alpha)\Gamma(\theta + n)} \prod_{j=1}^{k(n)} \{(\theta + k(n)\alpha)\} \frac{\Gamma\left(n_{jn} - \alpha\right)}{\Gamma(1 - \alpha)}. \tag{3.5.11}$$

EPPF for the DP given in Sect. 2.1 can be obtained by setting $\alpha = 0$ in the above expression.

Ishwaran and Zarepour (2003) also provide a sufficient condition to ensure exchangeability of a sequence in terms of the probabilities of selecting new values or choosing a previously sampled value under a modified Polya urn scheme. Consider a Polya urn sequence $X_1, X_2, \ldots$ with parameter ν whose distinct values in order of appearance are $X_1^*, X_2^*, \ldots$ with frequencies $n_1, n_2, \ldots$, respectively, generated by the following predictive rule:

$$\begin{aligned} \mathcal{P}(X_1 \in \cdot) &= \nu(\cdot) \\ \mathcal{P}(X_{n+1} \in \cdot | X_1, X_2, \ldots, X_n) &= \sum_{j=1}^{k(n)} \frac{q_j(\mathbf{n})}{\sum_{i=0}^{k(n)} q_i(\mathbf{n})} \delta_{x_i}(\cdot) + \frac{q_0(\mathbf{n})}{\sum_{i=0}^{k(n)} q_i(\mathbf{n})} \nu(\cdot), \end{aligned} \tag{3.5.12}$$

where $k(n)$ denotes the number of distinct values among the first n members of the sequence, $q_j(\mathbf{n})$, $i = 0, \ldots, k(n)$, are symmetric functions of $\mathbf{n} = \left(n_1, n_2, \ldots, n_{k(n)}\right)$. Their sufficiency condition is:

Condition: For each $n \geq 1$, $q_i(\mathbf{n}) = \psi(n_{n,i})$ and $q_0(\mathbf{n}) = \psi_0(n_{n,0})$, where ψ and ψ_0 are some fixed nonnegative real valued functions; and for each partition $\mathbf{n}$ of $\{1, \ldots, n\}$,

$$\sum_{i=0}^{k(n)} q_i(\mathbf{n}) = \xi(n) > 0$$

for ξ some fixed real valued function.

Then it is shown that if this condition holds, then the sequence $X_1, X_2, \ldots$ is exchangeable. Among their examples are (1) sequence of iid random variables are obviously exchangeable with $q_0(\mathbf{n}) = 1$ and $q_i(\mathbf{n}) = 0$, $i = 1, \ldots, k(n)$; (2) Blackwell–MacQueen sequence correspond to $q_0(\mathbf{n}) = \nu(\mathfrak{X})$, $q_i(\mathbf{n}) = n_i$ and $\sum_{i=0}^{k(n)} q_i(\mathbf{n}) = \nu(\mathfrak{X}) + n$; (3) the two-parameter Poisson–Dirichlet process with parameters θ and α corresponds to $q_0(\mathbf{n}) = \theta + k(n)\alpha$ and $q_i(\mathbf{n}) = n_i - \alpha$,

where $0 \leq \alpha < 1$ and $\theta > -\alpha$. In this case $\sum_{i=0}^{k(n)} q_i(\mathbf{n}) = \theta + n$; and (4) finite dimensional Dirichlet priors. Let $N > 1$ be a positive integer and let $q_0(\mathbf{n}) = \theta(1 - n/N) I\{n < N\}$ and $q_i(\mathbf{n}) = n_i + \theta/N$, where $\theta > 0$. Then $\sum_{i=0}^{k(n)} q_i(\mathbf{n}) = \theta + n$. This leads to the generalization of Blackwell–MacQueen theorem: If we have a sequence defined by the prediction rule in (3.5.12), then there exists a discrete RPM P^* such that the conclusion of Blackwell–MacQueen theorem holds true, but P^* need not be a DP.

Chapter 4
Priors Based on Levy Processes

4.1 Introduction

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, by defining a class of priors called *neutral to the right (NTR)* processes. It is defined in a distinctly different manner. 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 has close connection with the processes with independent nonnegative (will be dealing with only nonnegative) increments and can be defined in terms of an independent increment process. In fact it is shown that F is neutral to the right if and only if the process $Y_t = -\log(1 - F(t))$ is a nondecreasing process with independent increments. This allows us to specify a wide choice of family of NTR prior processes, one for each independent increment process. Several processes discussed in this chapter belong to this family of Levy processes.

By writing $H(t) = -\log(1 - F(t))$, Kalbfleisch (1978) considered a class $\mathcal{H}$ of cumulative hazard functions (CHFs) and defined an independent increment process as a prior over this class (and consequently on $\mathcal{F}$) in which the increments were taken to be distributed as gamma distribution, resulting in a *gamma process* prior. To allow the case when F may not have a density, Hjort (1990) defines the CHF as $A(t) = \int_{[0,t]} \frac{dF(s)}{F[s,\infty)}$ and considered the class of CHFs. He defines a prior over this class via the independent increment process (IIP) and assumes independent increments to be distributed approximately as beta distribution. The resulting prior process was named as *beta process*. Walker and Mallick (1997a) define a class of priors known as *beta-Stacy processes* directly over $\mathcal{F}$ by taking a particular independent increment process which they call as *log-beta process,*

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1_4

whose increments are distributed as beta-Stacy distribution. If F is an NTR, then irrespective of the definition of CHF, both H and A turn out to be independent increment processes. In the beta process, the focus was on the cumulative integral of the sample realization of the process. Thibaux and Jordan (2007) found it useful to take the realizations themselves in defining a *beta-Bernoulli* process which was then followed by *stable-beta* and *kernel beta* processes. It is interesting to note that some of these processes can be constructed by the SB construction as will be noted.

It is well known that if the fixed points of discontinuities and the deterministic part of an independent increment process are removed then it can be characterized by its Levy measure. This makes it convenient to study the NTR processes via their Levy measures. A further generalization is possible and useful. Instead of viewing priors as processes on the real line, they can be studied on abstract spaces via the *completely random measures* (CRMs) introduced by Kingman (1967), and the fact that such measures can be constructed using the *Poisson processes*. This approach was espoused in Lijoi and Prünster (2010). The advantage is that it provides a unified theory. However, it does not provide any insight into their origination. For this reason, original derivations will be described here.

Since the neutral to the right processes as well as other processes discussed here is intimately connected to independent increment processes it is instructive to review briefly the independent increment processes first. For the same reason, the CRM, Poisson process, and related theorems are useful to review briefly before using them in describing various processes. Following these short digression, several prior processes along with their properties and posterior distributions, will be presented in more details. Since the neutral to the right process plays a central role, it is described first and more space is devoted to it. Thereafter, gamma, extended gamma, beta, beta-Stacy, and Indian buffet processes are presented.

4.1.1 Nondecreasing Independent Increment Processes[1]

A nondecreasing process with independent increments (also called *additive process*, *positive Lévy process*, and *subordinator*) is a continuous time version of the sum of independent random variables. It is defined as $\{Y_t\}_{t\in T}$ where T is an interval, mostly taken as $(0, \infty)$, and the increments of $\{Y_t\}$ are nonnegative and independent. It plays 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 prior processes also emerge from

[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.

nondecreasing processes with independent increments (and their associated 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-Stacy, and gamma processes. 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 Y_t with $t \in R$ has independent increments if for every positive integer n and for every real numbers $t_0 < t_1 < \cdots < t_n$ the increments $Y_{t_1} - Y_{t_0}, Y_{t_2} - Y_{t_1}, \ldots, Y_{t_n} - Y_{t_{n-1}}$ are stochastically independent. We consider below only processes with independent increments for which the sample paths Y_t are nondecreasing with probability one. Thus each increment $Y_t - Y_s$ for $t > s$ is nonnegative 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 nonnegative 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 nondecreasing 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

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

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 nonnegative. In such a case, the one-sided Laplace transform of the distribution exists and has a very simple form. (See, for example, Feller 1966, Chap. XIII.7.) Its logarithm form for $\theta \geq 0$ is

$$\psi(\theta) = \log\, \mathcal{E}\, e^{-\theta X} = -\theta b + \int_0^\infty (e^{-\theta z} - 1) dN(z), \tag{4.1.1}$$

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

$$\int_0^1 z dN(z) < \infty \text{ and } \int_1^\infty dN(z) < \infty. \tag{4.1.2}$$

Some simple well-known examples of a nonnegative 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 (4.1.1) 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

$$f(x) = \Gamma(\alpha)^{-1}\beta^{-\alpha}e^{-x/\beta}x^{\alpha-1}I_{(0,\infty)}(x), \quad \alpha > 0 \text{ and } \beta > 0,$$

has the log Laplace transform $\psi(\theta) = -\alpha \log(1+\beta\theta)$ which may be represented in the form of Eq. (4.1.1) as

$$\psi(\theta) = \alpha \int_0^\infty (e^{-\theta z} - 1)e^{-z/\beta}z^{-1}dz$$

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 (4.1.2).

3. The random variable $X = -\log Y$, where $Y \sim \text{Be}(\alpha, \beta)$, has the *Log-Beta distribution* with density

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}e^{\alpha x}(1-e^x)^{\beta-1}I_{(0,\infty)}(x),$$

and the log Laplace transform of X,

$$\psi(\theta) = \log\Gamma(\alpha+\beta) + \log\Gamma(\alpha+\theta) - \log\Gamma(\alpha+\beta+\theta) - \log\Gamma(u),$$

which may be represented in the form of Eq. (4.1.1) as

$$\psi(\theta) = \int_0^\infty (e^{-\theta z} - 1)\frac{e^{-\alpha z}(1-e^{-\beta z})}{(1-e^{-z})z}dz$$

with $b = 0$ (see Ferguson 1974, Lemma 1). Again the corresponding Lévy measure

$$dN(z) = \frac{e^{-\alpha z}(1-e^{-\beta z})}{(1-e^{-z})z}dz$$

is not finite yet satisfies conditions (4.1.2).

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

Sampling an ID Distribution

In applications, we need to sample an ID distribution. Bondesson (1982) seems to be the first one to develop an algorithm to generate random variables from an ID distribution given the Levy measure in its Laplace transform form. His procedure is as follows. Suppose we want to sample X from an ID distribution F. In that case F has Laplace transform without the deterministic part as

$$\phi(s) = \exp\left\{\int_0^\infty (e^{-sz} - 1) dN(z)\right\},$$

where the Levy measure satisfies $N((1,\infty)) < \infty$ and $\int_0^1 z dN(z) < \infty$. If $\lambda = N((0,\infty)) < \infty$, define a d.f. $G(x) = \lambda^{-1} N((0,x])$. Let $\gamma(s)$ be the Laplace transform of G and $\nu \sim \text{Poisson}(\lambda)$. Then $\phi(s) = \exp\{\lambda(\gamma(s) - 1)\}$ is the Laplace transform of random variable $X = \sum_{i=1}^{\nu} Y_i$, where Y_i are iid with d.f. G and independent of ν. This suggests a method of simulating F. However, when λ is not finite, this method does not work and Bondesson develops a more general and flexible method. Let $\{Z(u) : u > 0\}$ be a family of nonnegative independent random variables and $T_i, i = 1, 2, \ldots$ be the points in increasing order in an independent Poisson point process with rate λ on R^+, and set $X = \sum_{i=1}^{\infty} Z(T_i)$. It is then shown that the Laplace transform of X is given by

$$\phi(s) = \mathcal{E}\left[\exp\{-sX\}\right] = \exp\left\{\lambda \int_0^\infty (\zeta(s;u) - 1) du\right\}, \tag{4.1.3}$$

where $\zeta(s;u) = \mathcal{E}\left[\exp\{-sZ(u)\}\right]$ with $Z(u) \sim H(z;u)$. Changing the order of integration, the above expression can be written as

$$\phi(s) = \exp\left\{\lambda \int_0^\infty (e^{-sz} - 1)\left(\int_0^\infty H(dz;u)\, du\right)\right\}. \tag{4.1.4}$$

Now suppose that for every $u > 0$ we can find a suitable distribution $H(z;u)$ on R^+ and a λ such that

$$\lambda \int_0^\infty H(dz;u)\, du = N(dz). \tag{4.1.5}$$

Then one can simulate points T_i of a Poisson (λ) process and then values $Z(T_i)$ from the distribution functions $H(x;T_i)$ and set $X = \sum_{i=1}^{\infty} Z(T_i)$. Then X will have the desired ID distribution. In particular, we may define $H(z;u) = H(z/g(u))$ where

$g(u)$ is a nonnegative function. By taking H to be degenerate at 1, he shows that it is possible to find the function g so that $\lambda \int_0^\infty H(dz; u)\, du = N(dz)$ is satisfied. An obstacle is that in practice it may be difficult to identify the distribution H.

Damien et al. (1995) have also developed an algorithm to generate approximate random variables from an ID distribution based on the Levy measure, $N(\cdot)$ in its Laplace transform, and random variables approximating stable distributions. Their motivation is derived from the fact that the characteristic function of any ID distribution can be expressed as the limit of Poisson-type characteristic functions. Thus any ID random variable may be expressed as a weighted sum of Poisson random variables. The weights are derived from the Levy measure. The method is illustrated by applying it to the gamma process, the Dirichlet process, and simple homogeneous process (Ferguson and Phadia 1979). This method is used by Damien et al. (1995) in simulating ID random variables for the extended gamma process Dykstra and Laud (1981), discussed later. We state here their algorithm in the general case and apply to specific processes later during their discussions.

1. Generate $x_1, \ldots, x_n \overset{\text{iid}}{\sim} (1/k)\, d\theta(x)$, where $k = \int_0^\infty d\theta(x)$ and $\theta(x)$ is a finite measure.
2. Generate $y_i \sim \text{Poisson}\left(k(1 + x_i)/nx_i\right), i = 1, \ldots, n$,
3. Define $z = \sum_{i=1}^n x_i y_i$.

They prove that the characteristic function of z converges as $n \to \infty$ to the characteristic function of an ID distribution and thus z serves as a random variable whose distribution approximates an ID distribution.

Wolpart and Ickstadt (1998) suggest an exact method called Inverse Levy measure to simulate an IID on the interval $[0, T]$.

Nondecreasing Infinitely Divisible Processes

Let Y_t be an almost surely nondecreasing 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

$$\psi_t(\theta) = -b(t)\theta + \int_0^\infty (e^{-\theta z} - 1) dN_t(z), \tag{4.1.6}$$

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 nondecreasing.

4.1.2 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 $\gamma\,(t)$, a right continuous function on R^+ and intensity parameter $\tau > 0$, 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

$$dN_t(s) = \frac{ds}{(1 - e^{-s})} \int_0^t \exp\left(-s\left(\beta\,(u) + \alpha\,\{u\}\right)\right) d\alpha_c\,(u)\,, \tag{4.1.7}$$

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

$$dN_t(s) = \int_0^t c(u)s^{-1}(1 - s)^{c(u)-1} dA_{0c}(u)ds, \tag{4.1.8}$$

where $A_{0c}(z)$ is the continuous part of $A_0(t)$ and $c\,(\cdot)$ is a piecewise 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

$$dN_t(s) = \frac{ds}{(1 - e^{-s})} \int_0^t c(u) \exp\left(-sc(u)G(u, \infty)\right) dG_c(u) \tag{4.1.9}$$

defines a beta-Stacy process (Walker and Muliere 1997a) with parameters $G\,(\cdot)$ and $c\,(\cdot)$, where $G_c(\cdot)$ 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

$$dN_t(s) = \frac{ds}{(1 - e^{-s})} \int_0^t c \exp\left(-sc(1 - G(u))\right) dG(u), \tag{4.1.10}$$

which upon simplification reduces to

$$dN_t(s) = \frac{e^{-sc}(e^{(1-G(t))} - 1)}{s(1 - e^{-s})} ds, \tag{4.1.11}$$

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 taken without cumulative and modified for continuous A_0 as

$$\upsilon(du, ds) = c(u) s^{-1} (1 - s)^{c(u)-1} A_0(du) ds$$

on $\Omega \times [0, 1]$, and 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.

The NTR and other processes can be described on an abstract space in terms of the CRMs introduced by Kingman (1967). The advantage is that it allows us to treat these processes in a unified manner. Since CRMs are constructed using the Poisson process on abstract spaces, this approach offers another method of constructing prior processes as will be seen later in connection with the dependent processes. A major disadvantage is that it does not provide any insight into how such processes arise. Nevertheless, it is instructive to view this approach as well. For this purpose, the following brief introduction to the CRM and Poisson process is taken from Kingman (1967, 1993).

4.1.3 Completely Random Measures

All of the above mentioned prior processes were developed by taking different forms of independent increment processes and their Levy measures. The notion of independent increment process has been generalized by Kingman (1967) in defining a CRM thereby making it possible to consider abstract spaces and not just the real line. He studied certain properties of the same and proved that such measures are almost surely discrete. Lijoi and Prünster (2010) recast the processes with independent increments in a more general and elegant manner in terms of CRMs tying together a number of neutral to the right type processes.

Definition 4.1 (Kingman) 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 CRM if, $\mu(A)$ is a random variable defined on some probability space $(\Omega, \mathcal{F}, Q)$ into $(\Pi^*, \sigma(\Pi^*))$ for all $A \in \mathcal{A}^*$ such that for any pair-wise 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.

A CRM μ is a stochastic process with index set $\mathcal{A}^*$, also known as the parameter set. The whole distribution over μ is determined once the distributions of random variables $\mu(A)$ are given for all $A \in \mathcal{A}^*$. The distribution of μ can be derived in terms of its characteristic function. The CRMs are discrete with probability one. It is a generalization of the independent increment process on the real line, since $\mu((0,t])$ on the real line may be viewed as a stochastic process with independent increments.

Kingman described a way to construct a CRM using the nonhomogeneous Poisson process (Kingman 1993). The Poisson process is defined on an abstract space S as a random set of points such that the number of points in disjoint sets is independent Poisson variates. Formally,

Definition 4.2 (Kingman) Let $(S, \sigma(S), \mu)$ be a measure space, let Π be a random subset of S and let $N(A) = \text{card}(A \cap \Pi), A \in S$. Then Π is said to be a Poisson process with measure μ if

(i) whenever for pair-wise disjoint members $A_1, A_2, \ldots, A_k, k > 1$ of $\sigma(S), N(A_1), N(A_2), \ldots, N(A_k)$ are independent.
(ii) if $\mu(A) < \infty$, then $N(A) \sim$ Poisson $(\mu(A))$; and if $\mu(A) = \infty$, then $N(A) = \infty$ a.s.

Note that Π is a Poisson process with mean measure μ if and only if N is a CRM. In conformity with the definition of a stochastic process, Π is a mapping from some probability space $(\Phi, \sigma(\Omega), P)$ into S^∞ of all countable subsets of S. Thus $N(A)$ is a function from $\Omega \to \{0, 1, \ldots, \infty\}$. For a Poisson process to exist, it is necessary that the measure μ must be nonatomic. N is finitely additive. To ensure it to be σ-additive, it is necessary that μ can be expressed as the sum of a countable number of σ -finite measures. If so, then there exists a Poisson process Π on S with measure μ and Π is almost certainly countable.

Theorem 4.3 (Superposition Theorem, Kingman) *Let $\Pi_1, \Pi_2, \ldots$ be a countable collection of independent Poisson processes on S and let Π_n have mean measure μ_n for each n. Then their superposition $\Pi = \cup_{n=1}^{\infty} \Pi_n$ is a Poisson process with mean measure $\mu = \sum_{n=1}^{\infty} \mu_n$.*

Suppose that with each point $\omega \in \Pi$, we associate a random variable X_ω, called *mark* of ω, taking value in some space Φ. The distribution of X_ω may depend on ω but otherwise independent of other points of Π. The pair (ω, X_ω) is then a random point ω^* of the product space $S \times \Phi$. The totality of such points forms a random countable subset $\Pi^* = \{(\omega, X_\omega) : \omega \in \Pi\}$ of $S \times \Phi$. Then a fundamental result is the following marking theorem:

Theorem 4.4 (Marking Theorem, Kingman) *The random subset Π^* is a Poisson process on $S \times \Phi$ with mean measure μ^* given by*

$$\mu^*(D) = \int \int_{(\omega,x)\in D} \mu(d\omega)\, p(\omega, dx), \tag{4.1.12}$$

where $p(\omega, \cdot)$ is a probability distribution depending upon ω and defined on Φ such that for any set $C \subseteq \Phi, p(\cdot, C)$ is a measurable function on S.

Kingman's excellent book *Poisson Processes* is a good source for further details.

If the measure of the Poisson process has infinite mass, it generates an infinite number of points. Thus a CRM can be expressed as $\mu = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}$, where ω_i are atoms of measure μ and p_i are the weights attached to the atoms.

As in processes with independent increments, the measure μ may be described as having (apart from the deterministic component) two components: μ_1 representing the part with fixed points of discontinuity and μ_2 representing random points of discontinuities. Write

$$\mu = \mu_1 + \mu_2 = \sum_{i \geq 1} S_i \delta_{c_i} + \sum_{i \geq 1} J_i \delta_{\xi_i}, \tag{4.1.13}$$

where S_i are random masses at fixed points c_i of discontinuities, are nonnegative and mutually independent, and are also independent of μ_2. The jumps in μ_1 will occur at sample observations, whether the prior process may or may not have them. With the fixed discontinuity part removed, μ will admit Levy representation

$$\log \mathcal{E}\, e^{-\theta \mu_2(0,t]} = \int_0^{\infty} (e^{-\theta z} - 1) d N_t(z), \tag{4.1.14}$$

where $d N_t(z)$ is the Levy measure, at times written as $d N_t(z) = \nu(dz, dt) = \rho_t(dz)\, \alpha(dt)$, 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 is independent of their location and both ν and μ_2 are termed *homogeneous;* otherwise they are *nonhomogeneous.* Then μ_2 and subsequently μ can be constructed by using the Poisson process with mean measure υ. μ_2 can be characterized in terms of the distribution of random points set $\{(J_i, \xi_i)\}_{i \geq 1}$ as a Poisson process with intensity measure given by the Levy measure υ. It is often convenient to specify the Levy measure of μ_2 and the distribution of mass at fixed points of discontinuities separately rather than specifying the full Levy measure of μ. Some examples of Levy measures of μ_2 leading to well-known processes are: the mean measure

$$\upsilon(dz, ds) = \gamma(z) s^{-1} e^{-\tau s} ds dz, s > 0 \tag{4.1.15}$$

produces the gamma process with shape parameter $\gamma(t)$ and intensity parameter τ;

$$\upsilon(dz, ds) = c(z)\, s^{-1} (1 - s)^{c(z)-1} A_0(dz)\, ds, 0 < s < 1 \tag{4.1.16}$$

results in the beta process with parameters $(c(\cdot), A_0)$ where $c(\cdot)$ is a positive function known as concentration function, and A_0 is a fixed base measure on Ω.

Random probability measures can also be obtained by normalizing completely random measures. For example, consider the points $\{(q_i, \xi_i)\}$ obtained from a Poisson process with mean measure υ used in the gamma process. Now define

a random probability measure $P = \sum_{i\geq 1} p_i \delta_{\xi_i}$, where $p_i = q_i / \sum_{i\geq 1} q_i$. Then P is a Dirichlet process with Levy measure given by

$$dN_t(s) = s^{-1}\left(1 - e^{-s}\right)^{-1} e^{-sc}(e^{(1-G(t))} - 1)ds, \tag{4.1.17}$$

which can also be written as

$$\upsilon\left(dt, ds\right) = s^{-1}\left(1 - e^{-s}\right)^{-1} e^{-sc}(e^{(1-G(t))} - 1)G(dt))ds. \tag{4.1.18}$$

where $c > 0$ is a constant and $G(t)$ is a continuous base distribution function. Note that P is not a CRM, since $P(A_1)$ and $P(A_2)$ for disjoint sets A_1 and A_2 are not independent but are negatively correlated.

For this line of development, see Lijoi and Prünster (2010). It is clear from the above that different forms of Levy measure ν define different processes as observed in their paper. A good account of defining discrete nonparametric priors by normalizing CRMs is given in a recent article (Favaro and Teh 2013). In view of the similarity of the CRM approach to the one presented below, we prefer to stick with the our 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 CRMs approach. This is evident, for example, in the development of the beta and beta-Stacy processes.

4.2 Processes Neutral to the Right

The Dirichlet process has a single parameter, α, and with its specification the prior is completely identified. As noted earlier, Doksum (1974) saw it as a limitation and introduced a general method of defining a broad class of priors called *neutral to the right* processes. They are constructed on the basis of independence of successive normalized increments of the distribution function F. They are closely connected to the processes with independent increments and can be defined in terms of an independent increment process. If the fixed points of discontinuities and the deterministic part of an independent increment process are removed then it can be characterized by its Levy measure. Thus the NTR processes may be studied in terms of their Levy measures. This approach was found to be critical in developing various other prior processes as will be seen later on.

The neutral to the right process can be specified by four quantities instead of a single parameter α of the DP: fixed points of discontinuities, distributions of jumps, continuous deterministic part, and continuous Levy measure. Thus the process allows a large amount of freedom to the statistician in choosing a prior. On the other hand, in contrast to α, it is difficult to interpret these parameters in terms of prior belief. Like the DP, it also select a discrete distribution function with probability one. But unlike the DP, it is conjugate with respect to the data that may include right

censored observations. However, the posterior distribution is not in a close form and usually difficult to compute. It can be characterized in terms of where the number of observations fall relative to the point of evaluation of F. It has clear advantage over the tailfree processes (to be formally defined later on) in that it may be defined independently of any partition points. This class of priors is sufficiently broad and as will be seen later that besides the Dirichlet process, it includes processes such as gamma, simple homogeneous, and beta-Stacy processes which are found to be useful in specific applications. It can be extended to spatial neutral to the right processes in the same way the DP was extended to form the class of dependent DPs (see Sect. 3.3). 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. Doksum's paper is the main source of the following material.

4.2.1 Definition

A random probability measure P is defined in the same way as in the case of the DP. However, Doksum assumes P to be σ-finite, i.e., the distribution of $P(A_n)$ tends to a degenerate distribution at 0 as $n \to \infty$. If $\{A_n\}$ is a decreasing sequence of sets A_n, then it is equivalent to $\lim_{n\to\infty} P(A_n) = 0$ a.s. 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. Essentially, P and F (as well as their distributions) are said to be neutral to the right if for all partitions $-\infty = t_0 < t_1 < \cdots < t_k < t_{k+1} = \infty$ of R, and k a positive integer, the normalized increments of F,

$$F(t_1), [F(t_2) - F(t_1)]/[1 - F(t_1)], \ldots, [F(t_k) - F(t_{k-1})]/[1 - F(t_{k-1})] \qquad (4.2.1)$$

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 $t_1 < t_2$,

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

That is 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 (4.2.1) are the hazard contributions of F of respective intervals and are known as *residual fractions*. See Sect. (3.2.4).

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

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

are independent. Because in (4.2.1) 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 $t_1 < t_2 < \cdots < t_m$ of real numbers, there exists independent nonnegative 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 nonnegative 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 $(V_1, V_1V_2, \ldots, \prod_1^m V_i)$.

By solving the set of equations

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

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

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

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 (4.2.1) are stochastically independent.

Doksum defines a sample from a random distribution function F in the same way as Ferguson (1973) did (see Sect. 2.1).

Viewing it differently and writing $1 - F(t) = e^{-H(t)}$, a prior can be specified on the space $\{H(t)\}$ of cumulative hazard functions, 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,

$$q_j = \big(F(t_j) - F(t_{j-1})\big) \,/\, \big(1 - F(t_{j-1})\big) = 1 - V_j, \tag{4.2.2}$$

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. Hjort (1990) develops a prior process called the beta process (see Sect. 4.5), for $H(t)$ even in the case when F may not have a density. Walker and Muliere (1997a) have shown its usefulness in models related to survival event history data analysis.

The definition may also be stated in terms of random probability measures as was done with the Dirichlet process. For any measurable partition $B_1, B_2, \ldots, B_m$, $m \geq 1$ of the real line, with $B_i = (t_{i-1}, t_i], i = 1, \ldots, m-1, t_0 = -\infty$ and $B_m = (t_{m-1}, \infty)$, neutral to the right and tailfree processes are defined in terms of independence properties of $P(B_1), \ldots, P(B_m)$. P is said to be neutral to the right if there exist independent nonnegative random variables $V_1, V_2, \ldots, V_m$ such that 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 $V_1, V_2, \ldots$ are taken to be independent with $V_i \sim \text{Be}(\alpha_i, \beta_i)$ 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 is still defined on the real line and therefore difficult to extend to higher dimensions.

The NTR process is defined on the real line. However, as pointed out by Doksum, the concept of neutrality can also be defined for random probabilities on abstract measurable spaces. This concept of a neutral random probability is an extension of Connor and Mosimann (1969) concept of neutrality for k-dimensional random vectors to a process as is the DP an extension of k-dimensional Dirichlet distribution to a process.

Let $\{\pi_m : m = 0, 1, \ldots\}$ denote a sequence of nested, measurable partition of $\mathfrak{X}$ with $\pi_0 = \{\mathfrak{X}\}$ and $\pi_m = \{A_{m1}, \ldots, A_{mk_m}\}$. We further assume that the partition is *ordered* from left to right by which we mean the set A_{mi} is left of (comes before) A_{mj} for all $i < j$. Now the definition may be stated as follows:

Definition 4.5 P is neutral with respect to the sequence $\{\pi_m\}$ of nested measurable, ordered partitions of $\mathfrak{X}$ if for each $m \geq 1$, there exist nonnegative independent random variables $V_{m1}, \ldots, V_{mk_m}$ with $V_{mk_m} = 1$ and

$$(P(A_{m1}), \ldots, P(A_{mk_m})) \stackrel{d}{=} (V_{m1}, V_{m2}(1-V_{m1}), \ldots, V_{mk_m} \prod_{j=1}^{k_m-1} (1-V_{mj})).$$

For the DP, we had only one partition and the random probability vector had the finite dimensional Dirichlet distribution. The term neutral refers to independence properties of sets within partitions, namely for each $m \geq 1$, the random variables

$$P(A_{m1}), P\left(A_{m2}|A_{m1}^c\right), P\left(A_{m3}|(A_{m1} \cup A_{m2})^c\right), \ldots, P(A_{mk_m}|A_{mk_m}) = 1$$

are independent. The interpretation is as one moves to the right, the relative random probability assigned to the next set is independent of the corresponding relative random probability assigned to the other sets in the partition. P is neutral since we can define random variables $V_{mi} = P(A_{mi}|A_{mi} \cup \ldots \cup A_{mk_m})$. We see a clear parallel to the definition of F being neutral on the real line.

4.2.1.1 Alternate Representation of NTR

The random distribution function F, neutral to the right, 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

$$F(t) = 1 - e^{-Y_t}. \tag{4.2.3}$$

Here the equality is to be understood as in distribution. Then the process Y_t has independent increments, since for any partition $t_1 < t_2 < \cdots < t_m$ 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. It can be interpreted as a random integrated hazard function. Furthermore, since $F(t)$ is assumed to be a distribution function a.s., it is nondecreasing 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 4.6 (Doksum) Let Y_t be a process with independent increments, nondecreasing 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 (4.2.3) 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 nonnegative, 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

$$Z_t = Y_t - \sum_j S_j I_{[t_j,\infty)}(t). \tag{4.2.4}$$

This process has independent increments, is nondecreasing a.s., and has no fixed points of discontinuity. Therefore, Z_t has infinitely divisible increments and its moment generating function (MGF) has Lévy formula

$$\log \mathcal{E}\, e^{-\theta Z_t} = -\theta b(t) + \int_0^\infty (e^{-\theta z} - 1) d N_t(z), \tag{4.2.5}$$

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

1. for every Borel set B, $N_t(B)$ is continuous and nondecreasing,
2. for every real t, $N_t(\cdot)$ is a measure on the Borel subsets of $(0, \infty)$,
3. $\int_0^\infty (z/(1+z)) d N_t(z) \to 0$ as $t \to -\infty$.

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

(a) $M = \{t_1, t_2, \ldots, \}$ the set of fixed points of discontinuity
(b) $\mathbf{f} = \{f_1, f_2, \ldots, \}$ 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

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

On the other hand, if $b \equiv 0$, 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 measures with probability one. The above four quantities enable us to describe the posterior distribution as will be seen next. We take $b \equiv 0$ and thus a neutral to the right process is characterized by $M, \mathbf{f}$, and N_t. We will denote in symbols, $F \sim \text{NTR}(M, \mathbf{f}, N_t)$.

It is difficult to interpret these parameters as was possible in the case of the DP. Walker and Damien (1998) offer a partial solution and provide a general method for specifying the prior mean and variance of an NTR, and as a consequence, provide interpretation of the parameters involved. They do so by relating the mean and variance of the random distribution function in terms of the Levy measure of the process. Let $S(t) = 1 - F(t)$. Using the Levy representation without the fixed points of discontinuity, it can be seen that

$$\mu(t) = -\log \mathcal{E}[S(t)] = \int_0^\infty (1 - e^{-z})\, dN_t(z) \quad \text{and}$$

$$\lambda(t) = -\log \mathcal{E}[S^2(t)] = \int_0^\infty \left(1 - e^{-2z}\right) dN_t(z).$$

Now the task is to find a Levy measure $N_t(\cdot)$ which satisfy these two equations with prior specification of μ and λ. They consider the Levy measures of type

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

where α is a finite measure and $\beta(\cdot)$ is a nonnegative function. It will be seen in a later section that this type of Levy measure characterizes a new process, the beta-Stacy process, which covers many of the known NTR processes. For example, the DP arises when $\beta(t) = \alpha(t, \infty)$ and the simple homogeneous process (Ferguson

and Phadia 1979) arises when β is constant. Walker and Damien (1998) prove the existence of $\alpha(\cdot)$ and $\beta(\cdot)$ such that

$$\mu(t) = \int_0^\infty \int_0^t e^{-z\beta(s)} d\alpha(s)\, dz \text{ and}$$
$$\lambda(t) = \int_0^\infty \int_0^t \left\{ \frac{1-\exp(-2z)}{1-\exp(-z)} \right\} e^{-z\beta(s)} d\alpha(s)\, dz. \tag{4.2.6}$$

If β is constant, then we have $\mu(t) = \alpha(t)/\beta$ and $\lambda(t) = [(1+2\beta)/(1+\beta)]$ $[\alpha(t)/\beta]$ allowing us to interpret the parameters α and β in terms of μ and λ.

The above alternate representation (4.2.3) shows a close connection between an NTR process and a nondecreasing IIP through the cumulative hazard function. The CHF is defined as $H(t) = -\log(1-F(t))$ and if the F does not have a Density, then $A(t) = \int_0^t \frac{dF(s)}{\overline{F}(s^-)}$, where $\overline{F} = 1-F$. If F has an NTR prior, then the prior distribution of A or H turns out to be an IIP, albeit with different Levy measures, λ_A and λ_H, respectively. In exploring this connection, it is sufficient to restrict attention to independent increment processes when $b \equiv 0$ and there are no fixed points of discontinuities. These are the NTR priors which give mass one to all discrete distribution and such priors can be described fully in terms of their Levy measures. This line is pursued in Dey et al. (2003) and the authors prove some interesting properties on the way. In particular,

Proposition 4.7 (Dey et al.) *The following are equivalent for a prior* Π *on* $\mathcal{F}$.

(1) Π is neutral to the right
(2) the process $\{H(t) : t > 0\}$ induced by the mapping $F \to H(t)$ has independent increments.
(3) the process $\{A(t) : t > 0\}$ induced by the mapping $F \to A(t)$ has independent increments.

The Levy measure $dN_t(z)$ can alternatively expressed as $\nu(dt, dz)$. Thus many NTR processes can be defined by taking different forms of ν. One convenient way is to express it in terms of Levy measures λ_A and λ_H of cumulative hazard processes arising via A and H in which case λ_A is defined on the space $(0,\infty)\times(0,1)$, since the increment in A is restricted to the interval $(0,1)$ as will be seen later on while discussing the beta process, and λ_H on $(0,\infty)\times(0,\infty)$. One general form of λ_A may be considered as $\lambda_A(ds,du) = a(s,u)A_0(ds)\,du, 0 < s < \infty, 0 < u < 1$, where A_0 is an increasing right continuous function such that $A_0(t) - A_0(t^-) \le 1$ and $a(s,u)$ is such that $\int_0^1 ua(s,u)\,du < \infty$ for all s. By taking a specific form of $\lambda_A(dz,ds) = c(z)s^{-1}(1-s)^{c(z)-1}A_0(dz)\,ds$, where $c(\cdot)$ is a positive function known as concentration function, and A_0 is a fixed base measure on Ω, Hjort (1990) introduced the beta process prior (see Sect. 4.5) and showed its usefulness in a variety of applications. Similarly, by defining

$$\lambda_H(dz,ds) = (1-\exp(-s))^{-1} c(z)\exp(-sc(z)\overline{F}(z))dzH_0(ds), 0 < z, s < \infty,$$

where H_0 is a fixed measure, Walker and Muliere (1997a) introduced the beta-Stacy process prior on $\mathcal{F}$, which generalizes the DP and is defined as an NTR process with λ_H as above. The corresponding independent increment process was called a log-beta process. Beta and log-beta give rise to similar NTR priors via A^{-1} and H^{-1}. The relationship between the two priors is stated explicitly in Dey et al. (2003) as

Proposition 4.8 *λ_A is the distribution of $(s,u) \rightarrow (s, -\ln(1-u))$ under λ_H. Conversely, λ_H is the distribution of $(s,u) \rightarrow (s, 1-e^{-u})$ under λ_A.*

The representation (4.2.3) of a CDF in terms of an IIP has facilitated the development of several new processes. Kalbfleisch (1978) used a gamma process to place a prior on survival function S which he considered as a nuisance parameter in dealing with a regression problem. In order to construct a prior for hazard rates $\lambda(t)$, Dykstra and Laud (1981) define the extended gamma process, generalizing the gamma process, where the independent increments of the Y_t process were taken to be gamma distributed with scale parameters taken to be the corresponding increments of a continuous function $\alpha(t)$ treated as parameter of the extended gamma process. Then the process was defined as a convolution of another right continuous function $\beta(t)$, which was treated as a second parameter

4.2.2 Properties

Various properties of the neutral to the right process were highlighted in Doksum (1974).

1. The following three conditions are equivalent:

(i) F is neutral to the right.

(ii) For all $t_1 < \cdots < t_k, k \geq 1$ there exist 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.

(iii) F is tailfree with respect to every sequence of nested ordered partitions of the form $\pi_m = \{A_{m,1}, \ldots, A_{m,k_m}\}$ with $A_{m,i} = (t_{m,i}, t_{m,i+1}]$, $i = 1, \ldots, k_m; -\infty = t_{m,1} < \cdots < t_{m,k_m+1} = \infty, m = 1, 2, \ldots$

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 4.9 *$F(t)$ 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. nondecreasing, a.s. right continuous, independent increments process with $\lim_{t \rightarrow -\infty} Y(t) = 0$ a.s. and $\lim_{t \rightarrow \infty} Y(t) = \infty$ 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 property, one can compute $\mathcal{E}(F(t))$ by using $\mathcal{E}(F(t)) = 1 - \mathcal{E}\left(e^{-Y(t)}\right)$.

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. This follows from theorem in property 2 and a property of independent increment processes.

4. If F is neutral to the right and neutral to the left, then F is a Dirichlet process (on R) or a limit of Dirichlet processes or processes concentrated on two nonrandom points.

5. In the above definition of the neutral to the right process, if V_i's are chosen to be $\text{Be}(\alpha_i, \beta_i)$ such that $\beta_i = \sum_{j \geq i+1} \alpha_i$, then it reduces to the Dirichlet process on R as noted before.

6. If $P \notin \mathcal{C}$ (see characterization of the DP) is neutral with respect to all sequences of nested, measurable, ordered partitions, then P is a Dirichlet process.

7. Dirichlet process is neutral to the right process with respect to every sequence of nested, measurable ordered partitions of R. 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 of $(V_{m1}, V_{m2}(1 - V_{m1}), \ldots, V_{mk_m} \prod_{j=1}^{k_m-1} (1 - V_{mj}))$, 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 $\text{Be}(\alpha(A_{mi}), \alpha(R) - \sum_{j=1}^{i} \alpha(A_{mj}))$, 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 parameter

$$(\alpha(A_{m1}), \alpha(R) - \alpha(A_{m1})), \left(\alpha(A_{m2}), \alpha(R) - \sum_{j=1}^{2} \alpha(A_{mj})\right),$$

$$(\alpha(A_{m3}), \alpha(R) - \sum_{j=1}^{3} \alpha(A_{mj})), \text{etc.},$$

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

$$V_{mi}\prod_{j=1}^{i-1}(1 - V_{mj}) \sim \text{Be}\left(\alpha(A_{mi}), \alpha(R) - \sum_{j=1}^{i}\alpha(A_{mj})\right).$$

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

Theorem 4.10 (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.*

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.

Description of the posterior distribution is given next.

4.2.3 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 $\alpha + \sum_{i=1}^{n}\delta_{X_i}$. Similarly for the NTR process, the posterior distribution of F is NTR with updated parameters. However, it is not that simple to update parameters in this case. 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. As noted earlier, 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 $t_1 < t_2 < \cdots < t_k$ 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 of Y_t process, 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

$$\begin{aligned} f(z_1, \ldots, z_k, x) &= \left(\prod_{i=1}^{k} f_i(z_i) \right) P(X \in (t_{j-1}, t_j] \mid \mathbf{Z} = \mathbf{z}) \\ &= \left(\prod_{i=1}^{k} f_i(z_i) \right) (F(t_j) - F(t_{j-1})) \\ &= \left(\prod_{i=1}^{k} f_i(z_i) \right) e^{-\sum_{i=1}^{j-1} z_i} (1 - e^{-z_j}) \\ &= \left(\prod_{i=1}^{j-1} f_i(z_i) e^{-z_i} \right) f_j(z_j)(1 - e^{-z_j}) \left(\prod_{i=j+1}^{k} f_i(z_i) \right). \end{aligned} \tag{4.2.7}$$

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 $(1 - e^{-z})$ and normalizing. This can be extended for a sample of size n, in part as follows.

We have to be mindful of three cases: what happens to the increments of intervals in which no observation fall, intervals before and after an observation; what happens to the distribution of jumps at fixed points of discontinuities; and what happens to the jumps at points other than the points of discontinuities. The first one is easy to handle and was answered in Doksum's paper.

Theorem 4.11 (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.*

Thus the posterior distribution of an increment of an interval depends only upon the number of observations beyond the interval and not where they fall. This is similar to the case of the DP where the posterior distribution depended on only the number of observations belonging to the interval and not where they were located.

For the second case, 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

$$g_x^*(z) = \frac{e^{-rz}(1-e^{-z})^m g_x(z)}{\int e^{-rs}(1-e^{-s})^m g_x(s)\,ds}. \tag{4.2.8}$$

There is a 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 4.10.

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

$$\mu_B((-\infty,t]) = b(t)I_B(0) + \int_B (1-e^{-z})dN_t(z). \tag{4.2.9}$$

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

$$\nu_B(t) = \frac{d\mu_B}{d\mu_{[0,\infty)}}(t) \tag{4.2.10}$$

exists for all B.

Theorem 4.12 (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*

$$H_x(s) = \nu_{[0,s]}(x). \tag{4.2.11}$$

Thus we summarize the above observations as a theorem for sample size one. Following Hjort (1990) and Walker and Muliere (1997a), we write $dN_t(z) = dz\int_{(0,t]} a(z,s)\,ds$, where $a(z,s)$ is some nonnegative function.

Theorem 4.13 *Let F be a random distribution function neutral to the right with parameters, $M, \mathbf{f} = \{f_j\}_{j=1}^k$ and $N_t(z)$, and let X be a sample of size 1 from F. Then the posterior distribution of F given X is again neutral to the right with updated parameters $M^*, \mathbf{f}^* = \{f_j^*\}_{j=1}^k$ and $N_t^*(z)$ as described below. In symbols; If $F \sim NTR(M,\mathbf{f},N_t)$, then $F|X \sim NTR(M^*,\mathbf{f}^*,N_t^*)$, where $dN_t^*(z) = dz\int_{(0,t]} a^*(z,s)\,ds$. Let x be a real number.*

(i) *Given $X = x$, and $x = t_i$ for some i, $M^* = M, f_j(s)$ changes to*

$$f_j^*(s) = \begin{cases} \kappa e^{-s} f_j(s) & \text{if } t_j < x \\ \kappa(1 - e^{-s}) f_i(s) & \text{if } t_j = x \\ f_j(s) & \text{if } t_j > x, \end{cases} \tag{4.2.12}$$

$$a^*(s,z) = \begin{cases} a(s,z)e^{-z} & \text{if } z \le x \\ a(s,z) & \text{if } z > x, \end{cases} \tag{4.2.13}$$

where κ is the normalizing constant.

(ii) *Given $X = x$, and $x \neq t_i$ for any i, an additional point of discontinuity is added to the set $M, M^* = M \cup \{x\}$, with $f_j(s)$ changes to*

$$f_j^*(s) = \begin{cases} \kappa e^{-z} f_j(s) & \text{if } t_j < x, \\ f_j(s) & \text{if } t_j > x, \end{cases}$$

$$f_x^*(s) = \kappa(1 - e^{-z})\, a(s,z), 0 < s < 1, \tag{4.2.14}$$

$a^*(s,z)$ as in (*i*) *and* $f_x^*(s)$ is the *density of new jump S* at *x*.

Ferguson and Phadia (1979) extended the above result for right censored data since the NTR is more amenable to deal with censored data than the DP. The posterior distribution of F given a censored observation is as follows:

Theorem 4.14 (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 $X \ge x$ 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.

The proof is straightforward. Consider an arbitrary partition of the real line $-\infty = t_0 < t_1 < \cdots < t_{j-1} < t_j = x < t_{j+1} < \cdots < t_n < \infty$, and define $W_i = Y_{t_{i+1}} - Y_{t_i}$ for $i = 0, \ldots, j-2$ (with $Y_{t_0} \equiv 0$), $W_{j-1} = Y_{t_j}^- - Y_{t_{j-1}}$, $W_j = Y_{t_j} - Y_{t_j}^-$, and for $i = j+1, \ldots, n, W_i = Y_{t_i} - Y_{t_{i-1}}$. Under the prior distribution, $W's$ are independent with their joint density, with respect to some convenient product measure, can be written as

$$f_{\mathbf{W}}(w_0, w_1, \ldots, w_n) = \prod_{i=0}^{n} f_{W_i}(w_i).$$

Given F, the probability that $X > x$ is

$$1 - F(x) = e^{-Y_x} = e^{-\sum_{i=0}^{j} w_i},$$

and therefore

$$f_{\mathbf{W}}(w_0, w_1, \ldots, w_n | X > x) \propto \prod_{i=0}^{j} e^{-w_i} f_{W_i}(w_i) \prod_{i=j+1}^{n} f_{W_i}(w_i).$$

This shows that the W's are independent a posteriori and hence the posterior distribution of F is neutral to the right. Similarly, for the case $X \geq x$ it can be seen that

$$f_{\mathbf{W}}(w_0, w_1, \ldots, w_n | X > x) \propto \prod_{i=0}^{j-1} e^{-w_i} f_{W_i}(w_i) \prod_{i=j}^{n} f_{W_i}(w_i),$$

from which the theorem follows.

For the general case of sample size n which may include censored data, Ferguson and Phadia derive the posterior distribution of F and give compact formulas in terms of the posterior MGF of the process Y_t. The idea was to be able to compute posterior moments—obviously this was before the advent of simulations. The formulas are given in Sect. 3.3.2 and illustrated with the calculation involved in two specific cases where the Y_t process is assumed to be homogeneous so that the Levy measure is of type $N_t(\cdot) = \gamma(t) N(\cdot)$, where $\gamma(t)$ is continuous nondecreasing function and N is any measure, both on $(0, \infty)$.

For the uncensored data case, the result is as follows. Let $M_t(\theta) = \mathcal{E} e^{-\theta Y_t}$ denote the MGF; $u_1 < u_2 < \cdots < u_k$ be the distinct ordered values among the sample of n observations $x_1, \ldots, x_n$ with $\delta_1, \ldots, \delta_k$ denoting the number of uncensored observations at $u_1, \ldots, u_k$, respectively, so that $\sum_1^k \delta_i = n; h_j = \sum_{i=j+1}^{k} \delta_i$ 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; $M_t^-(\theta)$ denotes the MGF of Y_{t-}, $M_t^-(\theta) = \lim_{s \to t} M_s(\theta)$, for $s < t$; and finally, let $G_u(s)$ 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 we have

Theorem 4.15 (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. Then the posterior distribution of F given the data is neutral to the right, and Y_t has posterior MGF*

$$M_t(\theta \mid data) = \frac{M_t(\theta + h_j(t))}{M_t(h_j(t))} \cdot \prod_{i=1}^{j(t)} \left[\frac{M_{u_i}^-(\theta + h_{i-1})}{M_{u_i}^-(h_{i-1})} \cdot \frac{C_{u_i}(\theta + h_i, \delta_i)}{C_{u_i}(h_i, \delta_i)} \cdot \frac{M_{u_i}(h_i)}{M_{u_i}(\theta + h_i)} \right], \tag{4.2.15}$$

where, if u is not a prior fixed point of discontinuity of Y_t,

$$C_u(\alpha, \beta) = \begin{cases} [c]c \int_0^\infty e^{-\alpha z}(1 - e^{-z})^{\beta-1} d\, H_u(z) & \text{if } \beta \geq 1 \\ 1 & \text{if } \beta = 0, \end{cases} \tag{4.2.16}$$

while, if u is a prior fixed point of discontinuity of Y_t, *then* $d\, H_u(z) = (1 - e^{-z}) d\, G_u(z)$.

Now it is easy to evaluate posterior moments of F. For example, the posterior expectation of the survival function $S = 1 - F$ is obtained by plugging $\theta = 1$ in the above expression, $\mathcal{E}(S(t)|\text{data}) = M_t(1|\text{data})$. However, the difficulty is still 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.

For the case of homogeneous processes, there are no prior fixed points of discontinuities, $b(t) \equiv 0$ and the Lévy measure of Y_t has the simple form, $dN_t(z) = \gamma(t) d\, N(z)$. Thus the above formula simplifies. $Y_t = -\log(1 - F(t))$ has the MGF

$$M_t(\theta) = \exp\left\{ \gamma(t) \int_0^\infty (e^{-\theta\, z} - 1) d\, N(z) \right\}. \tag{4.2.17}$$

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

$$H_x(s) = \int_0^s (1 - e^{-z}) d\, N(z) / \int_0^\infty (1 - e^{-z}) d\, N(z), \tag{4.2.18}$$

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

For the sample of n observations, the posterior MGF turns out to be

$$M_t(\theta \mid \text{data}) = \prod_{i=1}^{j(t)} \left[\frac{M_{u_i}^-(\theta + h_{i-1})}{M_{u_i}^-(h_{i-1})} \cdot \frac{C_{u_i}(\theta + h_i, \delta_i)}{C_{u_i}(h_i, \delta_i)} \cdot \frac{M_{u_i}(h_i)}{M_{u_i}(\theta + h_i)} \right] \cdot \frac{M_t(\theta + h_{j(t)})}{M_t(h_{j(t)})}, \tag{4.2.19}$$

where

$$C_u(\alpha, \beta) = \int_0^\infty e^{-\alpha z}(1 - e^{-z})^{\beta-1} d\, H_u(z). \tag{4.2.20}$$

For the gamma process prior, substitute $d\, N(z) = e^{-\tau z} z^{-1} dz$ and for a simple homogeneous process, $d\, N(z) = e^{-\tau z}(1 - e^{-z})^{-1} dz$, in the above formula (see Ferguson and Phadia 1979).

In view of the complication of evaluating the posterior distribution in general, and the need to simulate the same, the above description is not helpful and alternate descriptions are given in Hjort (1990) and Walker and Damien (1998) in terms of

the set of fixed points of discontinuities and the Levy measure. This will allow simulation of increments of intervals in which no observation fall, the simulation of jump sizes at fixed points of discontinuities, and the posterior form of the Levy measure. To carry out the simulation, procedures are developed in Damien et al. (1995), Damien et al. (1996), Bondesson (1982), and Wolpart and Ickstadt (1998), which are described in the sections dealing with these processes.

In practical applications, we need to simulate the jumps at fixed points of discontinuities and the increments which have ID distribution. The simulation procedures are discussed next.

4.2.3.1 Simulation of the Posterior Process

As has been recognized so far that except for the DP, the posterior distribution for the NTR process, and processes that follow, is difficult to handle. For this reason, most of the papers published early in 1970s and 1980s concentrated on getting estimates, mainly the posterior mean, in closed form and because of this limitation, the priors developed were limited in their applications to carry out the full Bayesian analysis. This changed with the pioneering paper of Damien et al. (1995) in which they developed simulation procedures for generating the posterior distribution thus obviating the need for applications limited to estimation alone. Since then it has become routine in application of nonparametric Bayesian methodology in analyzing data in various fields including the so called big data. Bayesian analysis in covariate and spatial data modeling was presented in the previous chapter. Here we describe briefly how the simulation of posterior distribution can be carried out for the case when the prior is an NTR process. This would be instructive since the other priors included in subsequent sections belong to the NTR family. For details the original papers can be referred.

If F is NTR, it can be written as $F(t) = 1 - e^{-Y_t}$ and Y_t is a nondecreasing process with independent increments. Y_t corresponds to a CRM and hence it can be constructed using the Poisson process methodology (Kingman 1967, 1993). This approach is suggested in Wolpart and Ickstadt (1998). Alternate approach is to recognize that an increment consists of sum of two parts: the jump part and the continuous part. They are independent by virtue of the process defined. Thus they can be generated separately. The jump components are independent random variables and hence they can be simulated by standard rejection algorithms once their prior densities are specified. The random variates corresponding to the continuous part is known to have an ID distribution and their simulation is achieved via the procedure of sampling an ID distribution described earlier.

Damien et al. (1995) use the theory developed in Ferguson and Phadia (1979) in terms of the posterior MGF to simulate this part of the process. As indicated earlier, their motivation arises from the fact that the characteristic function of any ID distribution can be expressed as the limit of Poisson-type characteristic functions. Thus in principle any ID random variable can be expressed as the infinite weighted sum of Poisson-type random variables. However, how to identify the weights is not

clear. Therefore, they replace the weights by suitable functions of random variables derived from the Levy measure of the process. Their algorithm to generate random variables approximating ID distributions described earlier is as follows:

1. Generate $x_1, \ldots, x_n \overset{\text{iid}}{\sim} (1/k)\, d\theta\,(x)$, where $k = \int_0^\infty d\theta\,(x)$ and $\theta\,(x)$ is a finite measure. ($d\theta\,(x) = x\,(1+x)^{-1}\, dN\,(x)$ if the Levy form of the MGF is available.)
2. Generate $y_i \sim \text{Poisson}\,(k\,(1+x_i)\,/nx_i)\,, i = 1, \ldots, n$,
3. Define $z = \sum_{i=1}^n x_i y_i$.

Their detail steps for generating a random variate from the distribution of an increment of the process are as follows:

(a) Given $g(w)$ the distribution of jump at a fixed point of discontinuity, generate $w \sim g(w)$ via Gibbs sampler or rejection algorithm.
(b) Let $d\theta\,(x) = f(x)dx$. Generate a sample $x_1, \ldots, x_n$ of size n from the density $f(x)$.
(c) Evaluate the normalizing constant $k = \int_0^\infty f(x)dx$.
(d) Simulate n Poisson variates $y_1, \ldots, y_n$ with parameter $\lambda_i = k\,(1+x_i)\,/nx_i$.
(e) Set $z = \sum_{i=1}^n x_i y_i$.
(f) Set $v = w + z \cdot v$ is a random variate from the desired distribution of the increment in question.

For example, in the case of the simple homogeneous process the steps (a)–(c) are as follows. Steps (d)–(f) are self-evident.

(a) the distribution of jump W is given by

$$g\,(w) \propto \exp\{-(\tau + s)\, w\}\,.$$

(b) $f\,(z) \propto \exp\{-(\tau+s)\, z\}\, z/\,(1+z)\,(1-\exp(-z))\,$, where s is the number of observations to the right of interval whose increment is under consideration.
(c) the normalizing constant k can be evaluated by numerical calculations since the numerator of $f\,(z)$ is an exponential density.

See their paper for a numerical example.

Simulation of the posterior process is also given in Walker and Damien (1998) paper and based on the description given above of the same (see Theorem 12). Ignoring the deterministic part, an NTR process is described by three quantities: The set $M = \{t_1, t_2, \ldots\}$ of fixed points of discontinuities, $f_1, f_2, \ldots$, the densities (or distributions) of the jumps there, and $N_t(\cdot)$ the continuous Lévy measure.

The Levy measure is usually taken as $dN_t(z) = dz \int_{(0,t]} a\,(z,s)\, ds$. For example, beta-Stacy process with parameters $\alpha\,(\cdot)$ and $\beta\,(\cdot)$ arises when

$$a\,(z,s)\, ds = (1-\exp(-s))^{-1}\,[\exp[-z\beta\,(s)]d\alpha\,(s)\,. \tag{4.2.21}$$

In view of the properties of NTR, the jumps at t_i's and the continuous increments are all independent and need to be simulated. To simulate the posterior distribution

given a single observation X (can be extended easily to any sample size), we note that if X is not one of the fixed points of discontinuity and $X > x$, the set M does not change and the jumps at fixed points of discontinuities can be sampled from the posterior densities of the form $\propto (1-\exp(-s))^{\lambda}(e^{-s})^{\mu}$, λ and μ nonnegative integers, given in the description of the posterior distribution above. This can be done via the Gibbs sampler or Metropolis–Hastings procedure.

If $X = x$ is not a prior fixed point of discontinuity, then an additional point of discontinuity is added. Since the increments are infinitely divisible, it needs to be sampled separately for which a general algorithm was described earlier. Similarly for the increment where no observation fall will have an ID distribution and can be handled as mentioned before. Walker and Damien (1998) give details of necessary steps. Their approach is based on the fact that the distribution of an ID random variable is the same as the distribution of $\int_{(0,\infty)} zdP(z)$, where P is a Poisson process with mean measure $dz\int_I a^*(z,s)\,ds$, I being the interval of the increment and $a^*(\cdot)$ being the updated form of $a(\cdot)$. Procedures pertaining to specific processes are given in respective sections dealing with those processes.

4.2.3.2 Characterization

As mentioned in Sect. 2.1 that if $X_1, X_2, \ldots$ is an exchangeable sequence of random variables defined on $(0,\infty)$, then from a theorem of De Finetti, there exists a random distribution function F conditional on which $X_1, X_2, \ldots \overset{\text{iid}}{\sim} F$, 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

$$\mathcal{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int \prod_{i=1}^{n} F(A_i)\,\mu(dF). \tag{4.2.22}$$

Lo's (1991) characterization of the Dirichlet process is: 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. When expressed in terms of the predictive probability based on a sequence of exchangeable random variables $X_1, \ldots, X_n$,

$$\mathcal{P}(X_{n+1} \in A | X_1, \ldots, X_n) = \varphi_n\left(\sum_{i=1}^{n}\delta_{x_i}(A)\right), \tag{4.2.23}$$

a function of the number of previous observations falling in the set A. In a similar spirit, Walker and Muliere (1999) proved the following characterization of the process neutral to the right. They drew their inspiration from W.E. Johnson's (see Zabel 1982) characterization of the Dirichlet distribution, namely the conditional probability of an outcome belonging to cell k is a function of number of previous

observations that fell in cell k. Let $X_1, X_2, \ldots$ be a sequence of random variables with each X_i defined on $(0, \infty)$, $X_i \sim F$, such that

$$\mathcal{P}\left(X_{n+1} > t | X_1, \ldots, X_n\right) = \prod_0^t \left[1 - dH\left(s, N\{s\}, N^+(s)\right)\right], \tag{4.2.24}$$

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, $i = 1, \ldots, n$, and

$$\begin{aligned} &dH\left(s, N\{s\}, N^+(s)\right)\left[1 - dH\left(s, N\{s\} + 1, N^+(s)\right)\right] \\ &\quad = dH\left(s, N\{s\}, N^+(s) + 1\right)\left[1 - dH\left(s, N\{s\}, N^+(s)\right)\right] \end{aligned} \tag{4.2.25}$$

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, Ericson, and Ramamoorthi (2003). For a discrete distribution function F with $S = 1 - F$, the hazard rate is defined as $F\{s\}/S(s^-)$, and the corresponding cumulative hazard function as $H(\cdot) = \sum_{s \le (\cdot)}[F\{s\}/S(s^-)]$. Now let $\mathcal{F}_0$ be the set of all distribution functions defined on the set $\{1, 2, \ldots\}$. Then for $F \in \mathcal{F}_0$ and Π an NTR on $\mathcal{F}_0$, we have

$$\mathcal{E}(S(j)) = \mathcal{E}\left(S(j-1) \cdot \frac{S(j)}{S(j-1)}\right) = \mathcal{E}(S(j-1))\,\mathcal{E}\left(\frac{S(j)}{S(j-1)}\right).$$

Therefore,

$$\mathcal{E}\left(\frac{S(j)}{S(j-1)}\right) = \frac{\mathcal{E}(S(j))}{\mathcal{E}(S(j-1))}.$$

That is $\mathcal{E}(H_F) = H_{\mathcal{E}(F)}$. This relationship leads to the following characterization of NTR on $\mathcal{F}_0$. Let Π be a prior on $\mathcal{F}_0$ such that for all $i \ge 1$, $S(i)/S(i-1) \in (0, 1)$ a.s. Then it is NTR if and only if for all n,

$$\mathcal{E}_{\Pi|\mathbf{X}}(H_F) = H_{\mathcal{E}_{\Pi|\mathbf{X}}(F)} \text{ and for any } k, \mathcal{E}_{\Pi|X>k}(H_F) = H_{\mathcal{E}_{\Pi|X>k}(F)},$$

where the expectation is with respect to the posterior distribution. It is conjectured by them that a similar characterization holds in general for $\mathcal{F}$.

Doksum had noted that for the NTR process prior, the posterior distribution of $S(t)$ depends upon the number of observations less than t and the number of observations greater than t and not where they fall, a property shared by the DP. Dey et al. show that this property essentially characterizes the NTR process.

4.2.4 Spatial Neutral to the Right Process

For accommodating a covariate in nonparametric Bayesian analysis, MacEachern (1999) defined a family $\{F_x : x \in \chi\}$ of Dirichlet processes, where χ is a subset of R^d, and called them Dependent Dirichlet processes (covered in Sect. 3.3). Note that marginally, each one of them is a DP. If we want to extend this concept to the neutral to the right processes, it is not clear how this can be done directly. James (2006) approaches this problem through the cumulative hazard function of F. Suppose we have random elements (T, X) defined on $[0.\infty] \times \chi$ having a distribution $F(dt, dx)$. The alternative definition $F(t) = 1 - e^{-Y(t)}$ of NTR does not provide any help. However, in dealing with event history data, Kalbfleisch (1978) and Hjort (1990) work directly with cumulative hazard function Λ defined as $\Lambda(dt) = F(dt)/S(t^-)$, where $S(t) = 1 - F(t)$ is the corresponding survival function. When F is a NTR process, Λ is a Levy process or a CRM, and when fixed points of discontinuities are removed, admits a representation (see section on CRMs) with Levy measure $dN_t(s) = \nu(ds, dt) = \rho_t(ds)\Lambda_0(dt)$ say, where $\rho_t(ds)$ is the distribution of jump in Levy process at location t and Λ_0 can be obtained from the initial guess F_0 of F. Now by extending Λ to a completely random hazard measure $\Lambda(dt, dx)$ on the space $[0.\infty] \times \chi$, James (2006) defines a *spatial neutral to the right process* on $[0.\infty] \times \chi$ as

$$F(dt, dx) = S(t^-)\Lambda(dt, dx). \tag{4.2.26}$$

Its Levy measure

$$\nu(ds, dt, dx) = \rho_t(ds)\Lambda_0(dt, dx), \tag{4.2.27}$$

where $\Lambda_0(dt, dx) = F_0(dt, dx)/S_0(t^-)$ is chosen based on the initial guess specification of $F_0(dt, dx)$. Clearly, the marginal yields an NTR process. This definition can be extended to include prior fixed points of discontinuities.

If

$$\rho_t(ds)\Lambda_0(dt, dx) = c(s)s^{-1}(1-s)^{c(s)-1}ds\Lambda_0(dt, dx), 0 < s < 1,$$

for $c(s)$ a positive function on $[0, 1]$ yields a natural extension of Hjort's (1990) beta cumulative hazard process to beta processes on $[0.\infty] \times \chi$, and thus defines what James calls *beta-neutral* distributions on $[0.\infty] \times \chi$. This extension facilitates the implementation of the neutral to the right mixture models on the same line as was done for Dirichlet mixture models in Sect. 2.4. He also derives the posterior distribution of spatial NTR process given the data $(t_1, x_1), \ldots, (t_n, x_n)$ from F. See his paper for details.

4.3 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) was the first one to explore this possibility in the context of Bayesian analysis of survival data. Writing a survival function $S(t) = e^{-H(t)}$, it is clear that $H(t)$ can be modeled as an independent increment process. This is done by choosing a gamma process prior, which has increments distributed independently as gamma distribution. However, his interest was in analyzing a regression model, the Cox model. It is expressed as $S(t) = 1 - F(t) = \exp\left\{-H(t)e^{\boldsymbol{\beta}\mathbf{W}}\right\}$, 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. Since his objective was to estimate the regression parameter $\boldsymbol{\beta}$, H was considered to be a nuisance parameter. Estimation of $\boldsymbol{\beta}$ proceeded by determining the marginal posterior distribution of data, having H eliminated. Later Wild and Kalbfleisch (1981) extended the work of Ferguson and Phadia (1979) to incorporate covariates (see Sect. 7.7).

4.3.1 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 4.16 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 log MGF given by

$$\log \mathcal{E}[\exp\{-\theta Z_t\}] = -\theta b(t) + \int_0^\infty (e^{-\theta s} - 1) d N_t(s), \tag{4.3.1}$$

where the Lévy measure has the form $d N_t(s) = \alpha(t) s^{-1} e^{-cs} ds$ and $b(t) \equiv 0$. Other properties of the gamma process are well known. Recently, Thibaux (2008) gives two interesting size-biased constructions of the gamma process.

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

4.3.2 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, $-\infty = t_0 < t_1 < t_2 < \cdots < t_m = \infty$, 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 nondecreasing and by assigning independent distributions to q_j's, a neutral to the right prior 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 identifying the posterior distributions 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 $x \in [t_{i-1}, t_i)$, the hazard rate 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 distribution of V and subsequent increments in H remain unchanged. While the distribution of U and all increments prior to x have gamma distribution with scale parameter changed from c to $c + 1$ (or by $c + e^{\beta W}$ relative to the observation if the regression model is considered). Thus $r_j \sim G\left(c\left(H_0(t_j) - H_0(t_{j-1})\right), c + 1\right), j = 1, \ldots, i - 1$, 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 function

$$f_J(s) = \frac{e^{-sc} - e^{-s(c+1)}}{s\left(\log\left(\frac{c+1}{c}\right)\right)} \tag{4.3.2}$$

and MGF

$$M_J(\theta) = \log((c + 1 - \theta)/(c - \theta))/\log((c + 1)/(c)). \tag{4.3.3}$$

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.

The conjugacy of Poisson and gamma distributions extend to processes as well in the same way as the conjugacy of multinomial to the DP and Bernoulli process to the beta process. Let the measure $B \sim \mathcal{G}(c, B_0)$, where c is a constant concentration parameter and B_0 the base measure. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(B)$, then the posterior distribution of $B|X_1, \ldots, X_n$ is again a $\mathcal{G}(c + n, B_n)$, where $B_n = \frac{c}{c+n} B_0 + \frac{1}{c+n} \sum_{i=1}^{n} X_i$.

Damien et al. (1995) give a procedure to simulate the posterior distribution of the gamma process with parameter $\gamma(t)$ and scale parameter c, viewed as a process neutral to the right. Recall that the MGF of gamma process is (Ferguson and Phadia 1979)

$$M_t(\theta) = \exp\left\{\gamma(t) \int_0^\infty \left(e^{-\theta z} - 1\right) e^{-cz} z^{-1} dz\right\}.$$

To use the steps described earlier, we need the following to generate an ID increment:

(a) the distribution of jump W is given by

$$g(w) \propto w^{-1} \exp\{-(c + s) w\} \{1 - \exp(-w)\},$$

where s is the number of observations to the right of the interval of the increment under consideration.

(b) $f(z) \propto \exp\{-(c + s) z\} / (1 + z)$.

(c) the normalizing constant k can be evaluated by numerical calculations since the numerator of $f(z)$ is an exponential density.

4.4 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 its variants were constructed on an arbitrary space of probability measures, and tailfree (to be presented in Sect. 5.1) 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.

4.4.1 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 the shape parameter $\alpha(t)$. It is assumed WLOG that this process has nondecreasing, left continuous sample paths.

Definition 4.17 (Dykstra and Laud) Let $Z(t) \in \mathcal{G}(\alpha(t), 1)$. Then a stochastic process defined by

$$r(t) = \int_{[0,t)} \beta(s)dZ(s) \tag{4.4.1}$$

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. Its sample paths are nondecreasing/nonincreasing with probability one and can be used as random hazard rates. Thus the extended gamma process serves directly as a prior on the class of increasing/decreasing hazard rate (IHR/DHR) functions. Its Levy formula for the Laplace transform is

$$M_{r(t)}(\theta) = \exp\left\{\int_0^\infty \left(e^{-\theta z} - 1\right) dN_t(z)\right\} \tag{4.4.2}$$

and

$$dN_t(z) = \left[\int_0^t z^{-1} e^{-\beta(s)z} d\alpha(s)\right] dz. \tag{4.4.3}$$

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.

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.

4.4.2 Properties

1. The MGF of hazard rate is $M_{r(t)}(\theta) = \exp\left\{-\int_{[0,t)} \log(1-\beta(s)\theta)d\alpha(s)\right\}$.
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 as follows:

Theorem 4.18 (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*

$$S(t) = \mathcal{P}(X \geq t) = \exp\left\{-\int_{[0,t)} \log(1+\beta(s)(t-s))d\alpha(s)\right\} \tag{4.4.4}$$

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

$$\begin{aligned} S(t_1, \ldots, t_n) &= \mathcal{P}(X_1 \geq t_1, \ldots, X_n \geq t_n) \\ &= \exp\left\{-\int_{[0,t)} \log\left(1+\beta(s)\sum_{i=1}^{n}(t_i-s)^+\right)d\alpha(s)\right\}, \end{aligned} \tag{4.4.5}$$

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, parameters

$\alpha(\cdot)$ and $\beta(\cdot)$ may be specified suitably in terms of $\mu(\cdot)$ and $\sigma^2(\cdot)$ as follows:

$$\beta(t) = \left(\frac{d\sigma^2(t)}{dt}\right) \Big/ \left(\frac{d\mu(t)}{dt}\right) \quad \text{and} \quad \frac{d\alpha(t)}{dt} = \left[\frac{d\mu(t)}{dt}\right]^2 \Big/ \left(\frac{d\sigma^2(t)}{dt}\right). \tag{4.4.6}$$

4.4.3 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 4.19 (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*

$$\beta^*(t) = \frac{\beta(t)}{1 + \beta(t) \cdot \sum_{i=1}^{m} (x_i - t)^+}; \tag{4.4.7}$$

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

$$\mathcal{P}(r(t) \in B | X_1 = x_1, \ldots, X_m = x_m) = \frac{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \prod_{i=1}^{m} \beta^*(z_i)\, \Psi(B; Q) \prod_{i=1}^{m} d[\alpha + \sum_{j=i+1}^{m} I_{(x_j,\infty)}](z_i)}{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \prod_{i=1}^{m} \beta^*(z_i) \prod_{i=1}^{m} [d\alpha + \sum_{j=i+1}^{m} I_{(x_j,\infty)}](z_i)}, \tag{4.4.8}$$

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.

Damien, Laud, and Smith (1996) give a Monte Carlo method that approximate random increments of the posterior process. However, they conclude that an alternative method given by Bondesson (1982), discussed earlier, is more efficient. Let

d_i be the number of exact observations in the i-th interval and f_{δ_i} be the prior density of the increment δ_i whose Laplace transform is given in the Levy form above. For the purpose of simulating the posterior density $f^*_{\delta_i}$, they choose $g(u) = e^{-u}$ and show that $\lambda = \alpha(s_i) - \alpha(s_{i-1})$, and $H(x)$ is given by $1 - \lambda^{-1} \int_{s_{i-1}}^{s_i} e^{-\beta(s)x} d\alpha(s)$. Simulation of H creates no problem since it is a mixture of exponentials. To simulate δ_i, their algorithm is

(a) Simulate n exponential variates with parameter $\lambda = \alpha(s_i) - \alpha(s_{i-1})$.
(b) Define n-vector $T = (T_1, \ldots, T_n)$, whose elements are the cumulative sums of exponential variates of step (a). These two steps correspond to simulation from the Poisson process with intensity parameter λ.
(c) Sample n independent variates $w_1, \ldots, w_n$ with distribution proportional to $\alpha(s)$ restricted to the interval $(s_{i-1}, s_i]$.
(d) Sample n exponential variates $Z_1, \ldots, Z_n$ with parameter $\beta(w_i) e^{T_i}$, where T_i is the i-th element of T, corresponds to simulation from $H(z; T_i)$.
(e) Define $X = \sum_{j=1}^{n} Z_j \cdot X$ will have the required ID distribution f_{δ_i}.
(f) To sample δ_i use the rejection algorithm to determine whether X can be retained as a realization from the density $f^*_{\delta_i}$.
(g) Repeat the above steps until X is accepted.

4.4.3.1 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 related to 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 4.20 *(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 $\Gamma(\alpha + \sum_{j=1}^{n} N_j, \beta/(1 + n\beta))$.*

4.4.3.2 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

$$F(dx|G) = \frac{w(x)\, G(dx)}{\int w(x)\, G(dx)}, \tag{4.4.9}$$

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 $\gamma(\cdot) = r(t)/r(+\infty)$ 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 $X_1, \ldots, X_n | G \overset{\text{iid}}{\sim} F(dx|G)$, 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 $\Gamma^*\left(\alpha + \sum_1^n \delta_{x_i}, 1/w\right)$.

4.5 Beta Process I

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 unwieldy in practical applications. Nevertheless, as discussed in Sect. 4.3, Kalbfleisch (1978) used a specific independent increment process—the gamma process, as prior for the cumulative hazard function (leading to a neutral to the right process on $\mathcal{F}$). 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, $G = -\ln(1-F)$ is the cumulative hazard function. However, to accommodate the case when F does not have a density, he 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.

In his development the focus was on cumulative realization of the samples. However, by considering the sample realizations itself, Thibaux and Jordan (2007) found its usefulness in modeling data consisting of unlimited number of features where the beta process is used as a parameter for the Bernoulli process. Seizing on this useful application Paisley et al. (2010) provided a stick-breaking construction parallel to that of the Dirichlet process. Later Paisley et al. (2012) connected this construction to the Poisson process and proved some additional properties.

Hjort's construction is based on viewing the cumulative hazard rate as a stochastic process. Clearly, by construction 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 nondecreasing process with independent increments. This representation facilitated in developing expressions for the posterior distribution of F as well as for the posterior MGF 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 and regression models as well. His focus, however, is on the cumulative hazard rates and for reasons described below, the formulas of Ferguson and Phadia do not translate directly to his case. He derives his motivation from the discrete time case and treats the continuous time as a limiting case. We will, however, consider only the continuous time 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 at t of F. To permit the definition of cumulative hazard function when F has no density, a more general form of the definition of H, which is valid when F is absolutely continuous as well, is used.

$$H(t) = \int_{[0,t]} \frac{dF(s)}{F[s,\infty)}, \qquad F(t) = 1 - \prod_{[0,t]} \{1 - dH(t)\}, \tag{4.5.1}$$

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. With all these considerations in mind, the space of all cumulative hazard rates restricted to a subspace $\mathcal{H}$, which results F to be a proper distribution function. To place a prior on $\mathcal{H}$ is to assign probability to every finite set of increments $H[t_j, t_{j-1})$ in Kolmogorov consistent way. This is done, and the existence (Hjort 1990, Theorem 3.1) is proved 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 $\mathcal{H}$. If F has a neutral to the right prior, then the distribution of both G and H turn out to be an independent increment process. A formal discussion on this relationship 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, it is designated as $H(t)$ process to reflect the role of the cumulative hazard rate and it highlights the distinction. Recall that Kalbfleisch (1978) also assigned a prior to the cumulative hazard function, but he used a gamma process since his focus was not on the distribution function per se. 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, the distribution of the jumps at fixed points of discontinuities are taken as independent beta distributions and the increments (infinitesimally small) of the process as independently but approximately beta distributed.

It is shown that the beta process so defined has the usual desirable properties: it has broad support, it is flexible, it has the conjugacy property with respect to the data which may include 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 leads to a special case of the beta process. In addition, its applications to nonhomogeneous models such as Markov Chain, competing risks, and covariate models are pointed out. Damien et al. (1996) provide detailed steps for implementation of the beta process in practice. Hjort's very detailed and comprehensive paper contains many useful results and applications and may be worthwhile to review. In the next section we present an extension of the beta process on arbitrary spaces in which the cumulative hazard function H is replaced by a finite measure in the same way that the distribution function F was replaced by a random probability measure P while developing prior distributions in earlier sections.

4.5.1 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

$$h(x) \sim \text{Be}\{c(x)h_0(x), c(x)(1 - h_0(x))\}. \tag{4.5.2}$$

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

$$dH(x) \sim \text{Be}\{c(x)dH_0(x), c(x)(1 - dH_0(x))\}. \tag{4.5.3}$$

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 $Var(h(x)) = h_0(x)(1 - h_0(x))/[c(x) + 1]$ as the prior "uncertainty." A formal definition is as follows:

Definition 4.21 (Hjort) An independent nonnegative increment process H, also known as (positive) Levy process, is a *beta process* with parameters $c(\cdot)$ and $H_0(\cdot)$, symbolically,

$$H \sim \mathcal{B}e\{c(\cdot),\ H_0(\cdot)\}, \tag{4.5.4}$$

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

$$M_t(t) = \log \mathcal{E}\left(e^{-\theta H(t)}\right) = \sum_{t_j \leq t} \log \mathcal{E}\left(e^{-\theta S_j}\right) - \int_o^1 \left(1 - e^{-s\theta}\right) dL_t(s), \tag{4.5.5}$$

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

$$dL_t(s) = \int_0^t c(z)s^{-1}(1-s)^{c(z)-1}dH_{0,c}(z)ds \quad \text{for } t \geq 0, \text{ and } 0 < s < 1, \tag{4.5.6}$$

and $H_{0,c}(t) = H_0(t) - \sum_{t_j \leq t} H_0\{t_j\}$ is H_0 with its jumps removed. (Note that the Levy measure can also be alternatively written as $\upsilon_{H_{0,c}}(dt, ds) = c(t)s^{-1}(1 - s)^{c(t)-1}dH_{0,c}(t)ds$.)

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 $[0, \infty)$.

The existence of beta process is guaranteed by proving first the existence of such a Levy process.

Theorem 4.22 (Hjort) *Let $H_0 \in \mathcal{H}$ be continuous and $c(\cdot)$ be a piecewise continuous, nonnegative function. then there exists a Levy process H, whose paths*

a.s. fall in $\mathcal{H}$ and whose Levy representation is

$$\log \mathcal{E}\left(e^{-\theta H(t)}\right) = -\int_o^1 \left(1 - e^{-s\theta}\right) dL_t(s), \tag{4.5.7}$$

where

$$dL_t(s) = \int_0^t c(z)s^{-1}(1-s)^{c(z)-1} dH_0(z)ds \quad \text{for } t \geq 0, \text{ and } 0 < s < 1, \tag{4.5.8}$$

Proof Here are the main steps of his proof.

For each n, consider the interval $(\frac{i-1}{n}, \frac{i}{n}]$ and define independent random variables $X_{ni} \sim \text{Be}(a_{ni}, b_{ni})$, where $a_{ni} = c_{ni}H_0(\frac{i-1}{n}, \frac{i}{n}]$, $b_{ni} = c_{ni} - a_{ni}$, and $c_{ni} = c\left(\frac{i-\frac{1}{2}}{n}\right)$, the value of function $c(\cdot)$ at the midpoint of the interval. Further let $H_0(0) = 0$, and define $H_n(t) = \sum_{\frac{i}{n} \leq t} X_{ni}, t \geq 0$. Thus H_n represents the sum of independent random variables. Then it is easy to see that as $n \to \infty$,

$$\begin{aligned} E(H_n(t)) &= \sum_{\frac{i}{n} \leq t} H_0\left(\frac{i-1}{n}, \frac{i}{n}\right] \to H_0(t); \text{ and} \\ \text{Var}(H_n(t)) &= \sum_{\frac{i}{n} \leq t} H_0\left(\frac{i-1}{n}, \frac{i}{n}\right]\left(1 - H_0\left(\frac{i-1}{n}, \frac{i}{n}\right]\right) / (c_{ni}+1) \\ &\to \int_0^t \frac{dH_0(s)\left(1 - dH_0(s)\right)}{c(s)+1}. \end{aligned} \tag{4.5.9}$$

Next we need to show that the sequence $\{H_n(t)\}$ converges in distribution to a Levy process H with the aforementioned properties. The standard technique is to show that the finite dimensional distributions of the sequence $\{H_n(t)\}$ converges properly and that the sequence is tight in the space D of all right continuous functions on $[0, \infty)$ with left-limits, and equipped with the Skorohod topology. This would imply that the sequence $\{H_n(t)\}$ converges in distribution to a random element of D. Hjort first restricts the space to $D[0, R], R > 0$ and shows that the sequence converges to an element of $D[0, R]$, say $H_{[R]}$, for each R. Then takes $H_{[R]}$ to be the $[0, R]$ restriction of the Levy process H on $[0, \infty)$ with the Levy measure as given in the definition. As a final step he shows that H indeed lies in $\mathcal{H}$ by noting that its restriction $\mathcal{H}_{[R]}$ on $D[0, R]$ is closed with respect to the Skorohod topology and that H_n certainly lies in $\mathcal{H}_{[R]}$ for all large n. Thus it follows from Billingsley (1968)'s Theorem 2.1 that H lies in $\mathcal{H}$ with probability 1.

These steps may seem routine, but the work involved is far from simple. A crucial step involved is in showing the convergence of finite dimensional distributions. With

this in mind, we first show

$$\log E\left(\exp\left(-\theta H_n(t)\right)\right) \to \left(\int_0^1 \left(e^{-\theta s}-1\right) dL_t(s)\right)$$
$$= \sum_{m=1}^{\infty} (-1)^m \frac{\theta^m}{m!} \int_0^t \int_o^1 c(z)\, s^{m-1} (1-s)^{c(z)-1}\, ds dH_0(z). \tag{4.5.10}$$

For fix t, the left-hand side is equal to

$$\log E\left(\prod_{\frac{i}{n}\le t} \exp\left(-\theta X_{ni}\right)\right) = \sum_{\frac{i}{n}\le t} \log E\left(\exp\left(-\theta X_{ni}\right)\right). \tag{4.5.11}$$

Expanding the exponential on the right-hand side and evaluating the expectations term by term with respect to the beta distribution, we get

$$\text{LHS} = \sum_{\frac{i}{n}\le t} \log\left[1 + \sum_{m=1}^{\infty} (-1)^m \frac{\theta^m}{m!} \frac{\Gamma(c_{ni})\,\Gamma(a_{ni}+m)}{\Gamma(a_{ni})\,\Gamma(c_{ni}+m)}\right], \tag{4.5.12}$$

where $\Gamma(\cdot)$ denotes the usual gamma function.

Now expanding the log function and ignoring the higher order terms as they can be shown to go to 0 as $n \to \infty$, and noting that $\frac{a_{ni}}{c_{ni}} = H_0(\frac{i-1}{n}, \frac{i}{n}]$, we have

$$\text{LHS} = \sum_{\frac{i}{n}\le t} \sum_{m=1}^{\infty} (-1)^m \frac{\theta^m}{m!} \frac{(a_{ni}+m-1)\ldots(a_{ni}+1)}{(c_{ni}+m-1)\ldots(c_{ni}+1)} H_0\left(\frac{i-1}{n}, \frac{i}{n}\right]$$
$$= \sum_{m=1}^{\infty} (-1)^m \frac{\theta^m}{m!} \left[\sum_{\frac{i}{n}\le t} \frac{(a_{ni}+m-1)\ldots(a_{ni}+1)}{(c_{ni}+m-1)\ldots(c_{ni}+1)} H_0\left(\frac{i-1}{n}, \frac{i}{n}\right]\right]$$
$$\to \sum_{m=1}^{\infty} (-1)^m \frac{\theta^m}{m!} \int_0^t \frac{\Gamma(m)\,\Gamma(c(z)+1)}{\Gamma(c(z)+m)} dH_0(z), \quad \text{as } n \to \infty \tag{4.5.13}$$

which is the right side of (4.5.10).

Likewise it can be shown that for any interval $(a_{j-1}, a_j]$,

$$\log E\left(\exp\left(-\sum_{j=1}^{k} \theta_j H_n(a_{j-1}, a_j]\right)\right) \to \sum_{j=1}^{k} \left(\int_0^1 \left(e^{-\theta_j s}-1\right) dL_{(a_{j-1},a_j]}(s)\right) \tag{4.5.14}$$

implying that the finite dimensional distributions of $\{H_n\}$ converges properly.

4.5.2 Properties

Some of the properties of the beta process are as follows:

1. $\mathcal{E}[H(t)] = \sum_{t_j \le t} \mathcal{E}S_j + H_{0,c}(t) = H_0(t)$.

2. $\text{var}[H(t)] = \sum_{t_j \le t} \text{Var}S_j + \int_0^t \frac{dH_{0,c}(s)}{c(s)+1} = \int_0^t \frac{dH_0(s)(1-dH_0(s))}{c(s)+1}$.
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 we have $H(t) = -\log(1 - F(t))$ and 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 4.23 (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*

$$H|\ data \sim \mathcal{B}e\left\{c(\cdot) + R(\cdot),\ \int_0^{(\cdot)} \frac{c(s)dH_0(s) + dN(s)}{c(s) + R(s)}\right\} \tag{4.5.15}$$

where $R(t) = \sum_{i=1}^n I[X_i \ge t]$, 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

$$H\{t\} \mid \text{ data} \sim \text{Be}\{c(t)H_0\{t\} + dN(t),\ c(t)(1 - H_0\{t\}) + R(t) - dN(t)\}. \tag{4.5.16}$$

Therefore, in describing the posterior distribution care must be taken. Hjort gives the description in his theorem (Hjort 1990, Theorem 4.1) which is reproduced in the next subsection, and is similar to Ferguson (1974) and Ferguson and Phadia's (1979) theorems.

5. Let $H \sim \mathcal{B}e\{c(\cdot), H_0(\cdot)\}$ and let $F(t) = 1 - \prod_{[0,t]}\{1 - dH(t)\}$. Then $\mathcal{E}[F(t)] = F_0(t) = 1 - \prod_{[0,t]}\{1 - dH_0(t)\}$, independent of $c(\cdot)$. Then for k any given positive constant and choosing $c(s) = kF_0[s, \infty)$, it is shown that F is a Dirichlet process with parameter $kF_0(\cdot)$, In fact, F may be considered as a *generalized Dirichlet process* with two parameters, $c(\cdot)$ and $F_0(\cdot)$.

6. Muliere and Walker (1997) have shown that the beta process may also be viewed as a Polya tree process (see Sect. 5.2).

7. MacEachern (1999) introduced Dependent Dirichlet processes by considering a family $\{F_x : x \in \chi\}$ of random distributions, where x was labeled as a covariate and thus it made possible to extend nonparametric Bayesian analysis to models involving covariates and auxiliary variables. Similar extension can be considered for hazard function. James (2006) follows this line and defines what he calls *spatial neutral to the right processes* described earlier.

4.5.3 Posterior Distribution

The beta process being an independent increment process, the theory developed for Y_t in the case of NTR can be adopted here. This is what precisely is done here, but formulas turn out to be different. In stating the description of the posterior distribution (needed for simulation), Hjort considers a slightly more general case. He assumes as prior a process with independent nonnegative increments with fixed points of discontinuities in order, $M = \{t_1, \ldots, t_k\}$, $\mathbf{f} = \{f_j\}_{j=1}^{k}$, where f_j is the prior density of jump S_j at $t_j, j = 1, \ldots, k$ and Lévy measure $dL_t(s) = ds \int_0^t a(s,z)dz$ on $0 < s < 1, t \geq 0$ where $a(s,z)$ is some continuous nonnegative function such that $\int_0^1 sdL_t(s) < \infty$. In the case of beta process priors, f_j's are beta densities, $a(s,z) = c(z)s^{-1}(1-s)^{c(z)-1}dG(s)$, where G is a continuous nondecreasing function with $G(0) = 0$, and $G = H_0$. (See Theorem 4.12 in Sect. 4.2.)

Theorem 4.24 (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 $M, \mathbf{f}$, $a(s,z)$ and H_0. Let $x \in R$. Then the posterior distribution of H is again a nonnegative independent increment process with parameters updated as follows. Here κ is the normalizing constant.*

(i) *Given $X > x$, $f_j(s)$ changes to*

$$f_j^*(s) = \begin{cases} \kappa(1-s)f_j(s) & \text{if } t_j \leq x, \\ f_j(s) & \text{if } t_j > x \end{cases}$$

and $a(s,z)$ gets multiplied by $(1-s)$ for only $z \leq x$. $M^* = M$.

(ii) *Given $X = x$, and $x = t_i$ for some i, $f_j(s)$ changes to*

$$f_j^*(s) = \begin{cases} \kappa(1-s)f_j(s) & \text{if } t_j < x \\ \kappa s f_i(s) & \text{if } t_j = x \\ f_j(s) & \text{if } t_j > x \end{cases}$$

and $a(s,z)$ gets multiplied by $(1-s)$ for only $z \leq x$. $M^* = M$.

(iii) *Given* $X = x$, *and* $x \neq t_i$ *for any* i, $f_j(s)$ *changes to*

$$f_j^*(s) = \begin{cases} \kappa(1-s)f_j(s) & \text{if } t_j < x, \\ f_j(s) & \text{if } t_j > x \end{cases}$$

and $a(s,z)$ *gets multiplied by* $(1-s)$ *for only* $z \leq x$, *and an additional point of discontinuity is added to the set* M *at* $x, M^* = M \cup \{x\}$, *with density of the jump* S at $x, f_x^*(s) = ksa(s,z), 0 < s < 1$.

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

$$f_r^*(s) = \kappa s^{n_r}(1-s)^{m_r} a(s, u_r), \tag{4.5.17}$$

where n_r is the number of uncensored observations at u_r and m_r is the number of observations greater than u_r. The density of the jump at t_j, a prior fixed point is given by

$$f_j^*(s) = \kappa s^{n_j}(1-s)^{m_j} f_j(s), \tag{4.5.18}$$

where n_j is the number of uncensored observations at t_j and m_j is the number of observations greater than t_j. 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 Sect. 7.3).

Damien et al. (1996) tailor the general method developed in Damien et al. (1995) for NTR processes to the present case and provide a technique to simulate the posterior distribution, which can then be used to carry out the analysis. Their approach is to discretize the time axis and differs from Hjort's method in that the distribution of the increments is not approximated. The distribution of the jump components at fixed points of discontinuities has beta posterior distribution given in (4.5.18) and can be generated by the standard rejection algorithm. To generate the posterior distribution of the continuous component of an increment of the process, they develop an algorithm to generate approximately random variates from the posterior process using the Levy formula for its MGF and proceed as follows. When the fixed points of discontinuities are removed, the distribution of the process has Levy measure

$$dL_t^*(s) = \left[\int_0^t \{c(z) + R(z)\} s^{-1}(1-s)^{c(z)+R(z)-1} \frac{c(z)dH_{0,c}(z)}{c(z)+R(z)} \right] ds, \tag{4.5.19}$$

where $R(z)$ is the number of observations greater than or equal to z. Now divide the time axis into intervals and let Δ_j denote the j-th interval and define

$$sdL_{\Delta_j}(s) = s\int_{t_{j-1}}^{t_j} c(z)s^{-1}(1-s)^{c(z)-1}dH_{0c}(z)ds \tag{4.5.20}$$

and let z^* denote a random variate from the normalizable measure dH_{0c} restricted to Δ_j. Then it is shown that $s^* \sim \text{Be}(1, c(z^*))$ is a random variate from $sdL_{\Delta_j}(s)$. Now the algorithm to generate random variates approximating the beta process increments can be stated as follows:

1. Generate $\mathbf{s}=\left(s_1^*, \ldots, s_n^*\right)$ from sdL_{Δ_j} using the above result;
2. Generate $y_i \sim \text{Poisson}\left(k/ns_i^*\right), i = 1, \ldots, n$, where $k = \int_0^1 sdL_{\Delta_j}(s)$;
3. Now the sample for the j-th increment is $H_j^* = \sum_{i=1}^n s_i^* y_i$.

They prove that as $n \to \infty, H_j^*$ tends in distribution to a variate having an infinitely divisible distribution. Details are given in their paper. They also illustrate their technique by reworking the example of Ferguson and Phadia (1979) and comparing the results with the results of Hjort (1990).

Various applications of the beta process in statistical inference problems, such as estimation of survival function, semiparametric regression models, Markov Chain, and dynamic Bayesian estimation discussed in his paper, are dealt with in the Applications Chaps. 6 and 7.

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

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 $R^+ \times \mathcal{X}$, discussed earlier.

4.6 Beta Process II

4.6.1 Beta Processes on General Spaces

While Hjort's primary interest was in placing a prior over a class of cumulative hazard functions, it was natural for him to define the beta process based on the independent increment process and taking the parameters c and H_0 as nondecreasing functions. The focus was on the cumulative integral of the sample realizations of the process, whereas in certain applications (such as word occurrence in document corpora), the attention is on the realizations themselves. This makes it necessary to define the beta process on more general spaces and need not be restricted to the real line only.

In Sect. 2.1 we alluded to the Chinese restaurant process (CRP) (see Sect. 4.8.1 for details) which produces an exchangeable sequence of random variables and the DP serves as its underlying mixing De Finetti measure. Similarly, the IBP (to be discussed in Sect. 4.8.2) generates an exchangeable sequence of binary matrices and the question would be, what is the underlying De Finetti measure for this sequence? This served as a motivation for Thibaux and Jordan (2007) to propose the beta process presented in this section and show that it is that measure.

As in the case of the Dirichlet process, greater flexibility is achieved by replacing Hjort's cumulative process with a random measure B, with parameters c, a positive function and B_0, a finite base measure, and the Levy measure without the integral. This is exactly what is done in Thibaux and Jordan (2007). The resulting random measure is also discrete and enjoys some similarities with the Dirichlet process in its construction. Thus many of the extensions of the Dirichlet process can also be formulated for the beta process. The authors also define a hierarchical set up for factorial modeling which is an analog of the hierarchical Dirichlet process. This makes it possible to consider models in which features are shared among a number of groups. For this general case of the beta process, we use slightly different notations to distinguish from the earlier section. Thereafter we will present a three-parameter generalization of the beta process developed by Teh and Gorur (2009) to take into account the power-law behavior. Similar to the stick-breaking construction of the Dirichlet process, we will also give stick-breaking construction of the beta processes. Thibaux (2008) defines the geometric process and shows that the beta process is also conjugate with respect to the geometric process.

Let $(\Omega, \sigma(\Omega))$ be a probability space. Recall that a positive random measure Q defined on Ω is said to be an independent increment process (or a completely random measure Sect. 4.1) if for any pair-wise disjoint sets $A_1, \ldots, A_k, k > 1$ of Ω, the measures $Q(A_1), \ldots, Q(A_k)$ are independent random variables. As stated there, this is a generalization of independent increment processes on abstract spaces. Now the beta process is redefined as follows:

Definition 4.25 (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 $B \sim \mathrm{BP}(c, B_0)$ if its Lévy measure for continuous B_0 is given by

$$\nu(d\omega, dp) = B_0(d\omega)\, c(\omega)\, p^{-1}(1-p)^{c(\omega)-1}\, dp \tag{4.6.1}$$

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. When B_0 is discrete of the form $B_0 = \sum_{i=1}^{\infty} q_i \delta_{\omega_i}$, where δ_{ω_i} is a unit point mass at $\omega_i \in \Omega$, then B has atoms at the same locations $B = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}$, but $p_i \sim \mathrm{Be}(c(\omega_i) q_i, c(\omega_i)(1-q_i))$, which

necessarily imply $0 \leq q_i \leq 1$. B_0 can also be expressed as αG_0 with α the mass parameter and G_0 a smooth base distribution function. α controls the overall mass of the process and G_0 controls the random locations of atoms.

For the case B_0 is continuous (nonatomic) such that $B_0(\Omega) = \gamma$, and c a positive scalar (can be extended to c, a positive function with extra work), Paisley et al. (2010) have shown that the beta process B can be derived as a limit of B_K, as $K \to \infty$, where $B_K = \sum_{k=1}^{K} p_k \delta_{\omega_k}$, with

$$p_k \overset{\text{iid}}{\sim} \text{Be}\left(\frac{c\gamma}{K}, c\left(1 - \frac{\gamma}{K}\right)\right), \quad \text{and } \omega_k \overset{\text{iid}}{\sim} \frac{1}{\gamma} B_0.$$

This can be seen as follows. When the K-vector of p's is integrated out and taking the limit $K \to \infty$, the marginal distribution produces the two-parameter extension of the IBP (Ghahramani et al. 2007) which can be shown to have the beta process as its underlying De Finetti mixing measure. Recall that the Dirichlet process was shown to be the underlying De Finetti measure for the CRP. Teh et al. (2007) (see Sect. 4.8.1) proposed a slightly different construction for the IBP in which they take $\text{Be}\left(\frac{c}{K}, 1\right)$ instead, which simplify the derivation. This construction, however, does not extend to the two-parameter generalization of the IBP).

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 mean measure ν. Since ν has infinite mass, there are infinite number of points. Thus B may be represented as an infinite sum $B = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}$. 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. This raises a question. Can B be constructed by the stick-breaking construction that was possible in the case of the Dirichlet process? This has been answered affirmatively in Paisley et al. (2010).

4.6.1.1 Stick-Breaking Construction

This construction is similar to the one obtained by Sethuraman (1994) for the Dirichlet process discussed earlier. In their construction, Paisley et al. (2010) use a special stick-breaking representation of the beta distribution (credited to Jayaram Sethuraman), namely, $X \sim \text{Be}(a, b)$ can be sampled according to the following stick-breaking scheme:

$$X = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1} (1 - V_j) I[Y_i = 1] \quad \text{with } V_i \overset{\text{iid}}{\sim} \text{Be}(1, a + b) \quad \text{and } Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{a}{a + b}\right), \tag{4.6.2}$$

and where $I[Y_i = 1]$ is the indicator function. Steps for the construction of the beta process B are

$$\begin{aligned}
&\text{For } i = 1, 2, \ldots \text{ select } C_i \overset{\text{iid}}{\sim} \text{Poisson}(\gamma);\\
&\text{for } k = 1, 2, \ldots, C_i \text{ draw } V_{ij}^{(l)} \overset{\text{iid}}{\sim} \text{Be}(1, c) \text{ and } \omega_{ik} \overset{iid}{\sim} \frac{1}{\gamma} B_0;\\
&\text{set } B = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{ij}^{(i)} \prod_{l=1}^{i-1} \left(1 - V_{ij}^{(l)}\right) \delta_{\omega_{ij}}. \qquad (4.6.3)
\end{aligned}$$

Here as stated before, $c > 0$ scalar and B_0 is nonatomic finite base measure. B can also be expressed as

$$\sum_{j=1}^{C_1} V_{1j} \delta_{\omega_{1j}} + \sum_{i=2}^{\infty} \sum_{j=1}^{C_i} V_{ij} e^{-T_{ij}} \delta_{\omega_{ij}} \text{ where } T_{ij} \overset{\text{iid}}{\sim} \text{Be}(i-1, c).$$

Then it has been shown that $B \sim \text{BP}(c, B_0)$.

Unlike in the case of the Dirichlet process, here each summand consists of a sum of random number C_i (drawn according to a Poisson distribution) of atoms, each atom drawn according to $\frac{1}{\gamma} B_0$ and weighted according to the stick-breaking scheme. That is, at each round i, C_i is drawn from Poisson(γ). For this round, round specific stick i is used. The mass associated with each atom in the i-th round is equal to the i-th break (discarding the previous $i-1$ breaks) from this stick, where the stick-breaking weights follow a Be$(1, c)$ stick-breaking process. In other words, all atoms ω_{ij} are chosen according to $\frac{1}{\gamma} B_0$. Identify the C_i atoms in the i-th round by $\omega_{i1}, \omega_{i2}, \ldots, \omega_{iC_i}$. Now corresponding to each of these atoms, take a different stick and use a Be$(1, c)$ stick-breaking process. Discard the first $i-1$ cuts and use the i-th cut as weight for the atom ω_{ij}. Superscript denotes the cuts up to the i-th round.

Since the number of prior cuts increases with each round, the attached weights decrease stochastically as was the case in the stick-breaking construction of the Dirichlet process. The weights are controlled by the parameter c. As c decreases, they get smaller more rapidly since smaller and smaller fractions of each stick remains to yield the weights; if c increases, the weights decay more gradually. The expected weight on an atom in round i is shown to be equal to $c^{(i-1)}/(1+c)^i$. The number of atoms at each round is controlled by γ.

In a later publication, Paisley et al. (2012) have shown that the above stick-breaking construction can be derived from the characterization of the beta process as a Poisson process with the mean measure ν. Let $\pi_{1j} = V_{1j}$ and $\pi_{ij} = V_{ij} \exp(-T_{ij})$ for $i > 1$, where $T_{ij} \sim G(i-1, c)$. Let $B_i = \sum_{j=1}^{C_i} \pi_{ij} \delta_{\omega_{ij}}$, and finally, $B = \sum_{i=1}^{\infty} B_i$. Noting that $C_i \overset{\text{iid}}{\sim} \text{Poisson}(\gamma)$, for each B_i the set of atoms $\{\omega_{ij}\}$ form a Poisson process with mean measure B_0. Each ω_{ij} is marked with a $\pi_{ij} \in [0, 1]$ that has

probability measure λ_i (to be indicated below). It is shown that each B_i has an underlying Poisson process $\Pi_i = \{\omega_{ij}, \pi_{ij}\}$, on $\Omega \times [0,1]$ with mean measure $B_0 \times \lambda_i$. Then by the superposition theorem (Kingman 1993) (see Sect. 4.1), $\Pi = \cup_{i=1}^{\infty} \Pi_i$ is a Poisson process with mean measure $\nu = B_0 \times \sum_{i=1}^{\infty} \lambda_i$. Thus B has an underlying Poisson process with mean measure $\nu \cdot \lambda_i$ determines the probability distribution of π_{ij} which has density $\lambda_1(d\pi) = c(1-\pi)^{c-1} d\pi$ for $i = 1$. The probability distribution of π_{ij} for $i > 1$ does not have a closed form. However, the authors have shown that the sum turn out to be a simple expression:

$$\sum_{i=1}^{\infty} \lambda_i = c\pi^{-1}(1-\pi)^{c-1}. \tag{4.6.4}$$

Thus,

$$\nu(d\omega, d\pi) = B_0(d\omega)\, c\pi^{-1}(1-\pi)^{c-1} d\pi. \tag{4.6.5}$$

In the above treatment c was taken to be a scalar. It can be extended to the general case of function $c(\cdot)$ considered by Hjort (1990) as long as B_0 is σ-finite. The idea is to partition the set Ω as the union of sets D_k with $\Omega = \cup_k D_k$ such that $B_0(D_k) < \infty$, for each $k = 1, 2, \ldots$. Now for each D_k construct the Poisson process as above with mean measure

$$\nu_{D_k}(d\omega, d\pi) = B_0(d\omega)\, c(\omega)\, \pi^{-1}(1-\pi)^{c(\omega)-1} d\pi \quad \text{for } \omega \in D_k \tag{4.6.6}$$

and apply the superposition theorem of the Poisson process. Each round of construction adds Poisson $(B_0(D_k))$ new atoms $\omega_{ij}^{(k)} \in D_k$ drawn iid from $B_0/B_0(D_k)$. For each of these atoms, the weight $\pi_{ij}^{(k)}$ is the i-th cut of a Be $\left(1, c\left(\omega_{ij}^{(k)}\right)\right)$ stick-breaking process. Now the union of these processes yield the beta process in the general case.

4.6.1.2 Hierarchical Distribution

Just as the Dirichlet process has been found to be useful in hierarchical modeling, so has been the beta process. Thibaux and Jordan (2007) give an example in which they model a document by the set of words it contains. They assume that document $X_{i,j}, i = 1, \ldots, n_j$ and $j = 1, \ldots, n$, is generated by including each word θ independently with a probability p_θ^j specific to category j. These probabilities form a discrete measure P_j over the space of words and a beta process prior $\text{BP}(c_j, B)$ is placed on P_j 's. In turn, B itself is generated as a realization of a beta process with parameters c_0 and B_0. Thus we have an hierarchical model as follows: $X_{i,j} \sim P_j, P_j \sim \text{BP}(c_j, B)$ and $B \sim \text{BP}(c_0, B_0)$.

The beta process is shown to be conjugate with respect to the *Bernoulli process* (encountered in connection with feature data modeling and the IBP discussed below) defined as follows:

Definition 4.26 (Thibaux and Jordan) Let B be a measure on Ω. An independent increment process Z on $(\Omega, \sigma(\Omega))$ is said to be a Bernoulli process with parameter (hazard measure) B, denoted by $Z \sim \text{BeP}(B)$ if its Lévy measure is given by

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

If B is continuous, then Z is a Poisson process, $Z = \sum_{i=1}^{N} \delta_{\omega_i}$ where $N \sim \text{Poisson}(B(\Omega))$ and ω_i's are independently and identically 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 or simply a Poisson process in which weights are either 0 or 1. 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 the atoms are as in B and q_i are independent Bernoulli random variables taking value 1 with probability p_i.

4.6.1.3 Posterior Distribution of B

The conjugacy property of the beta process with respect to independent Bernoulli processes may be used to derive the posterior distribution of B. Let $Z_1, \ldots, Z_n$ be a set of independent Bernoulli processes with hazard measure B, $Z_i|B \sim \text{BeP}(B), i = 1, \ldots, n$, and $B|c, B_0 \sim \text{BP}(c, B_0)$, then the posterior is $B|\mathbf{Z}, c, B_0 \sim \text{BP}\left(c+n, \frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} Z_i\right)$ which resembles expression (4.5.15). To generate posterior beta processes, we note the following. The posterior distribution of $B|\mathbf{Z}$ reveals that it is the sum of two independent beta processes. Say, for $n = 1, F_1 \sim \text{BP}\left(c+1, \frac{1}{c+1}Z_1\right)$ and $G_1 \sim \text{BP}\left(c+1, \frac{c}{c+1}B_0\right)$. This makes it possible to sample B by first sampling Z_1 which is a Poisson process with mean measure B_0, then sampling F_1 and G_1. Let $Z_1 = \sum_{i=1}^{K_1} \delta_{\omega_i}$. Since the base measure for F_1 is discrete, using the earlier discussion of discrete base measure, we get $F_1 = \sum_{i=1}^{K_1} p_i \delta_{\omega_i}$, where $p_i \sim \text{Be}(1, c)$. G_1 can further be decomposed into F_2 and G_2. Proceeding inductively, we get for any $n, B \stackrel{d}{=} \widehat{B}_n + G_n$ where $\widehat{B}_n = \sum_{i=1}^{n} F_i$. At each stage the expected mass of F_n gets accumulated where as the remaining expected mass, $E(G_n(\Omega)) = \frac{c\gamma}{c+n}$ decreases to 0. Thus $\lim_{n\to\infty} \widehat{B}_n \stackrel{d}{=} B$ and $\widehat{B}_n$ can be used as an approximation to B. This leads to the following iterative algorithm at any stage n given by the authors:

Sample $K_n \sim \text{Poisson}\left(\frac{c\gamma}{c+n-1}\right)$,
Sample K_n new locations ω_j from $\frac{1}{\gamma}B_0$ independently,
Sample weight of $\omega_j, p_j \sim \text{Be}(1, c+n-1)$ independently,
Set $\widehat{B}_n = \widehat{B}_{n-1} + \sum_{i=1}^{K_n} p_i \delta_{\omega_i}$.

Conditional Distribution

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

$$Z_{n+1}|\mathbf{Z}_n, c, B_0 \sim \text{BeP}\left(\frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} Z_i\right) = \text{BeP}\left(\frac{c}{c+n}B_0 + \sum_{k=1}^{K}\frac{m_k}{c+n}\delta_{\omega_k^*}\right), \tag{4.6.7}$$

where K is the number of distinct atoms among $Z_1, \ldots, Z_n$ with atom ω_k^* having m_k frequency. Therefore to sample $Z_{n+1}|\mathbf{Z}_n$ an atom is added at ω_k^* with probability $m_k/(c+n)$and a Poisson $(c\gamma/(c+n))$ number of new atoms are added at new locations drawn independently from B_0/γ. When c is a constant and B_0 continuous, this can be seen as the two-parameter version of the IBP establishing that the beta process is the de Finetti mixture measure of the IBP.

4.6.1.4 Geometric Process

Thibaux (2008) defines a *geometric process* with *hazard measure* B, denoted by $X \sim \text{GeoP}(B)$, as the Levy process with Levy measure given by

$$\nu(d\omega, d\pi) = (1 - B(d\omega))\sum_{k=1}^{\infty}\delta_k(dp)B(d\omega)^k.$$

This means that each atom $p_i\delta_{\omega_i}$ of B, X has an atom $N_i\delta_{\omega_i}$, where $N_i \sim \text{Geometric}(1-p_i)$. It is noted that the beta process is also conjugate to the geometric process, and the posterior distribution of B is given by

$$B|X_1, \ldots, X_n \sim \text{BP}\left(c_n, \frac{c}{c_n}B_0 + \frac{1}{c_n}\sum_{i=1}^{n} X_i\right), \quad \text{where } c_n = c + n + \sum_{i=1}^{n} X_i.$$

4.6.2 Stable-Beta Process

Parallel to the three-parameter generalization of the Dirichlet process, namely the Pitman–Yor process, Teh and Gorur (2009) generalize the beta process by introducing a stability parameter thereby incorporating the power-law behavior. Their motivation is to define the underlying de Finetti measure of the IBP with the power-law behavior extending the fact of beta process being the underlying de Finetti measure of the two-parameter IBP. Thus the new measure is a generalization of the beta process. This three-parameter generalization was named as a *stable-beta process* in view of the fact that it generalizes both the stable (Perman et al. 1992)

process and the beta process. It has no fixed point atoms and its Levy measure is given by

$$\nu\left(d\omega, dp\right) = \frac{\Gamma\left(1+c\right)}{\Gamma\left(1-\sigma\right)\Gamma\left(c+\sigma\right)} p^{-1-\sigma}\left(1-p\right)^{c+\sigma-1} dpB_0\left(d\omega\right), 0 < p < 1,$$

where $c > -\sigma$ is a concentration parameter, $0 \leq \sigma < 1$ is the stability parameter and B_0 is the base measure. The stability parameter governs the power-law behavior of the stable-beta process. When $\sigma = 0$, the process reduces to a standard two-parameter beta process. It can be shown that $E\left(B\left(A\right)\right) = B_0\left(A\right)$, and $\text{Var}\left(B\left(A\right)\right) = \left(\left(1-\sigma\right)/\left(1+c\right)\right) B_0\left(A\right)$ for $A \in \Omega$. See Broderick et al. (2012) for further elucidation and characterization of the beta process as a CRM.

Stick-Breaking Construction

The size-biased ordering of atoms leads to the following stick-breaking construction procedure for the stable beta process, where $\gamma = B_0\left(\Omega\right)$:

$$\text{For } i = 1, 2, \ldots \text{ select } C_i \sim \text{Poisson}\left(\gamma \frac{\Gamma\left(1+c\right)\Gamma\left(i-1+c+\sigma\right)}{\Gamma\left(i+c\right)\Gamma\left(c+\sigma\right)}\right);$$

$$\text{For } k = 1, 2, \ldots, C_i \text{ draw } \pi_{ik} \sim \text{Be}\left(1-\sigma, i-1+c+\sigma\right) \text{ and } \omega_{ik} \overset{\text{iid}}{\sim} \frac{1}{\gamma} B_0;$$

$$\text{Set } B = \sum_{i=1}^{\infty} \sum_{k=1}^{C_i} \pi_{ik} \delta_{\omega_{ik}}.$$

Posterior Distribution

Since the beta process is a conjugate prior of Bernoulli process, one can easily derive the posterior distribution which is still a beta process but not a stable-beta process, since it will include atoms generated by the observed data points. Let $Z_1, \ldots, Z_n$ be iid Bernoulli processes with parameter B, a stable-beta process. Since we expect some repetitions among the atoms, this can further be expressed in terms of distinct atoms with their respective distributions. Let $\omega_1^*, \ldots, \omega_K^*$ be K distinct atoms with ω_k^* occurring m_k times among $Z_1, \ldots, Z_n$, with the distribution $F_{nk}, k = 1, 2, \ldots, K$. Then the posterior distribution of $B|Z_1, \ldots, Z_n$ is a beta process with the updated Levy measure given by

$$\nu_n\left(d\omega, dp\right) = \frac{\Gamma\left(1+c\right)}{\Gamma\left(1-\sigma\right)\Gamma\left(c+\sigma\right)} p^{-1-\sigma}\left(1-p\right)^{c+n+\sigma-1} dpB_0\left(d\omega\right), \qquad (4.6.8)$$

and the distribution of mass at $\omega_k^*, k = 1, 2, \ldots, K$ given by

$$F_{nk}(dp) = \frac{\Gamma(n+c)}{\Gamma(m_k-\sigma)\Gamma(n-m_k+c+\sigma)} p^{m_k-1-\sigma}(1-p)^{c+n-m_k+\sigma-1}\, dp, \tag{4.6.9}$$

which is simply a beta distribution, $\text{Be}(m_k-\sigma, n-m_k+c+\sigma)$.

Conditional and Joint Distribution

The conditional distribution of $Z_{n+1}|Z_1, \ldots, Z_n$ can also be obtained easily by noting that

$$\mathcal{P}(Z_{n+1}(d\omega) = 1|Z_1, \ldots, Z_n) = E[B(d\omega)|Z_1, \ldots, Z_n] = \int_0^1 p\nu_n(d\omega, dp) \tag{4.6.10}$$

$$= \frac{\Gamma(1+c)\Gamma(n+c+\sigma)}{\Gamma(n+1+c)\Gamma(c+\sigma)} B_0(d\omega). \tag{4.6.11}$$

This shows that since Z_{n+1} is a random Bernoulli process, it can be sampled from a Poisson process defined on the space $\Omega \setminus \{\omega_1^*, \ldots, \omega_K^*\}$ with mean measure $\frac{\Gamma(1+c)\Gamma(n+c+\sigma)}{\Gamma(n+1+c)\Gamma(c+\sigma)} B_0$. It will have Poisson $\left(\gamma \frac{\Gamma(1+c)\Gamma(n+c+\sigma)}{\Gamma(n+1+c)\Gamma(c+\sigma)}\right)$ number of new atoms, each independently and identically distributed according to $\frac{1}{\gamma} B_0$. Multiplying successively the conditional probabilities of Z_i given the previous ones, we obtain the joint probability distribution of $Z_1, \ldots, Z_n$ eliminating B as

$$f(Z_1, \ldots, Z_n) = \exp\left(-\gamma \sum_{i=1}^{n} \frac{\Gamma(1+c)\Gamma(i-1+c+\sigma)}{\Gamma(i+1+c)\Gamma(c+\sigma)}\right) \cdot \prod_{i=1}^{K} \frac{\Gamma(m_k-\sigma)\Gamma(n-m_k+c+\sigma)\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)\Gamma(n+c)} B_0(\omega_k^*). \tag{4.6.12}$$

4.6.3 Kernel Beta Process

In the last chapter covariant dependent Dirichlet processes were constructed so that nonparametric Bayesian analysis of covariate data can be carried out. There an RPM having a DP prior was attached to each covariate. In the same way, Ren, Wang, Dunson, and Carin (2011) construct a family of dependent beta processes attached to each covariate in a covariate space. They call the resulting process as the *kernel*

beta process (KBP). This may be considered as an example of the general approach developed by Foti et al. (2012) for constructing dependent processes, presented in the last chapter, Sect. 3.3.

In the IBP, we have a vector Z_n of ones and zeros representing whether a particular feature (patron partakes the dish) is present or not in the n-th individual. Z_n is a Bernoulli process with parameter $B, Z_n \sim \text{BeP}(B)$, where B is a beta process with parameters c and base measure $B_0, B \sim \text{BP}(c, B_0)$. Like in the case of DP where an RPM F was associated with a covariate $x \in \chi, \chi$ a covariate space, here also a beta process is attached to the covariate x, say B_x and let the set $\mathcal{B} = \{B_x : x \in \chi\}$. In the DDP, the RPM F_x was constructed in terms of the Sethuraman representation and the covariate was associated with the location atom ξ or weight p. Parallel to that here the infinite sum representation of B_x is constructed by appealing to the Poisson process construction of a CRM.

Since B is a beta process with Levy measure

$$\nu(d\pi, d\omega) = c(\omega)\pi^{-1}(1-\pi)^{c(\omega)-1} d\pi B_0(d\omega), \tag{4.6.13}$$

it can be constructed using the PP with mean measure ν and has representation

$$B = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}, \tag{4.6.14}$$

where $(\omega_i, \pi_i) \in \Omega \times [0,1]$ are points of the PP. Since the points are generated via Poisson(λ) distribution and $\lambda = \int_{\Omega}\int_{[0,1]} \nu(d\pi, d\omega) = \infty$, there are countably infinite number of points. If Z_n is the n-th vector drawn from a BeP(B), then

$$Z_n = \sum_{i=1}^{\infty} b_{ni}\delta_{\omega_i}, \quad b_{ni} \sim \text{Bernoulli}(\pi_i).$$

The data corresponds to the set $\{Z_n\}_{n=1}^N$ where the indicator b_{ni} is utilized to represent whether the feature $\omega_i \in \Omega$ is present ($b_{ni} = 1$) or not ($b_{ni} = 0$) in Z_n. When B is integrated out, it yields conditional probability for $\{Z_n\}$ that corresponds to the IBP.

In the above beta-Bernoulli construction, the same B is used for generation of all Z_n's implying that the probability π_i of presence of feature ω_i remains the same for each n. Now to incorporate covariates, we associate covariate $x_n \in \chi$ with Z_n, and denote the set of covariates as $\{x_n\}$. Now a condition is imposed that if samples n and n' have similar covariates x_n and $x_{n'}$, then it is more likely that they will possess a similar subset of the features $\{\omega_i\}$; if the covariates are distinct, it is less probable that features will be shared.

Thus the set $\mathcal{B}$ will be generalized as

$$\mathcal{B} = \{B_x : x \in \chi\} = \sum_{i=1}^{\infty} \gamma_i \delta_{\omega_i}, \quad \omega_i \sim B_0, \tag{4.6.15}$$

where $\gamma_i = \{\gamma_i(x) : x \in \chi\}$ is a stochastic process from $\chi \to [0,1]$ independent of $\{\omega_i\}$, i.e. $\mathcal{B}$ is a set of dependent beta processes $B_x = \sum_{i=1}^{\infty} \gamma_i(x)\,\delta_{\omega_i}$. This is a general set up and governs various special cases. For example $\gamma_i(x)$ can be used to model predictor-dependent functions in probit, logit, or kernel SB processes. One such special case is the KBP.

The standard practice is to adjust the weights $\gamma_i(x)$ according to the distance between the value of the covariate x and a set of nearby fixed locations x' drawn according to some measure H on χ and the distance reflected in the kernel $K(\cdot,\cdot) : \chi \times \chi \to [0,1]$. In the kernel SB process discussed earlier we had, $\omega_i \overset{\text{iid}}{\sim} B_0$, $\gamma_i(x) = V_i(x) \prod_{k<i}(1 - V_k(x))$, $V_j(x) = \lambda_j K(x, x')$, where $\lambda_j \in \Psi$ is a feature dependent measure of proximity between x and x'. Another example is the Gaussian kernel where, $\chi = R$, and

$$K(x, x') = \exp\left\{-\lambda \left\|x - x'\right\|^2\right\}, \tag{4.6.16}$$

λ a positive real valued random variable distributed according to a measure Q defined on R^+. In this case the revised Levy measure is defined as

$$\nu_x = H(dx')\, Q(d\lambda)\, \nu(d\pi, d\omega), \tag{4.6.17}$$

where $\nu(d\pi, d\omega)$ is the Levy measure associated with the beta process.

Thus the dependent beta process can be expressed as $B_x = \sum_{i=1}^{\infty} \lambda_i K(x, x_i')\, \pi_i \delta_{\omega_i}$. Here $\lambda_i K(x, x_i')$ serves as a Foti et al. (2012) identifier $r_i^{x'}$ indicating whether the atom i in the global measure is included in the local measure B_x at covariate value x or not. We supplemented the CRM and PP with covariance space and dependent measures are now defined on an extended product space $\chi \times \Psi \times R^+ \times \Omega$, and the mean measure of the PP is given by ν_x. Therefore draws from the augmented PP would yield measures B_x. If $\chi = \{x_1, \ldots, x_k\}$, then $\mathcal{B}$ will be a k-dimensional vector. This covariance dependent beta process can be used in the feature model to define the covariance dependent feature model. With Z_{nx} distributed as $\text{BeP}(B_x)$, it generalizes the beta-Bernoulli model, where $Z_{nx} = \sum_{i=1}^{\infty} b_{nxi}\delta_{\omega_i}$, with $b_{nxi} \sim \text{Bernoulli}\left(\pi_i \lambda_i K(x, x_i')\right)$. This can be expressed equivalently as $b_{nxi} = z_{xi}^{(1)} z_{xi}^{(2)}$, with $z_{xi}^{(1)} \sim \text{Bernoulli}\left(\lambda_i K(x, x_i')\right)$ and $z_{xi}^{(2)} \sim \text{Bernoulli}(\pi_i)$. Thus the process evolves as a 2-step generalization of the beta-Bernoulli process.

For any Borel set $A \in \mathfrak{B}$, when $\mathcal{B}$ is drawn from the KBP and for covariates x and x', the authors derive

$$E(B_x(A)) = B_0(A)\, E(K_x)$$

$$\text{Cov}(B_x(A), B_{x'}(A)) = E(K_x K_{x'}) \int_A \frac{B_0(d\omega)(1 - B_0(d\omega))}{c(\omega) + 1} - \text{cov}(K_x K_{x'}) \int_A B_0^2(d\omega)$$

where $E(K_x) = \int_{\chi\times\Psi}\lambda K(x,x')H(dx')Q(d\lambda)$. If $\lambda K(x,x') = 1$ for $x \in \chi, E(K_x) = E(K_xK_{x'}) = 1, \text{cov}(K_xK_{x'}) = 0$ and the above results reduces to the original hierarchical beta process and IBP.

4.7 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 a particular independent non-negative increment process Z , namely the *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 in the beta process they take the increments to be distributed as *beta-Stacy distribution* with parameters, α and β and with density given by

$$f(y) = \frac{1}{B(\alpha,\beta)}y^{\alpha}\frac{(x-y)^{\beta-1}}{x^{\alpha+\beta-1}}I_{(0,x)}(y) \qquad \text{for } 0 < x \le 1, \tag{4.7.1}$$

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 special 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. Thus the beta-Stacy process belong to the class of neutral to the right processes.

4.7.1 Definition

First we define the log-beta process. 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 $\alpha_c(t) = \alpha(t) - \sum_{t_j \le t} \alpha\left\{t_j\right\}$. The log-beta process is defined as follows:

Definition 4.27 (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

$$\log \mathcal{E}\left(e^{-\theta Z(t)}\right) = \sum_{t_j \le t} \log \mathcal{E}\left(e^{-\theta S_j}\right) - \int_o^\infty \left(1 - e^{-v\theta}\right) dN_t(v), \tag{4.7.2}$$

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

$$dN_t(v) = \frac{1}{(1 - e^{-v})} \int_o^t \exp(-v(\beta(s) + \alpha\{s\}))\, d\alpha_c(s)\, dv, \quad v > 0. \tag{4.7.3}$$

In contrast to the beta process, here the Levy measure is not restricted to $(0, 1)$ interval and instead of jumps themselves, their functions are distributed as beta distribution.

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 $G_c(t) = G(t) - \sum_{t_j \le t} G\left\{t_j\right\}$ so that $G_c(t)$ is continuous. Let Z be an independent nonnegative increments process with log Laplace transform

$$M_t(\theta) = \log \mathcal{E}\left(e^{-\theta Z(t)}\right) = \sum_{t_j \le t} \log \mathcal{E}\left(e^{-\theta S_j}\right) - \int_o^\infty \left(1 - e^{-v\theta}\right) dN_t(v), \tag{4.7.4}$$

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

$$dN_t(v) = \frac{1}{(1 - e^{-v})} \int_o^t \exp(-vc(s)\, G(s, \infty))\, c(s)\, dG_c(s)\, dv. \tag{4.7.5}$$

Definition 4.28 (Walker and Muliere) Let Z 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 $t \geq 0, F(t) = 1 - e^{-Z(t)}$.

Thus we desire a process which has increments distributed (infinitesimally speaking) as beta-Stacy distribution and also that F is a distribution function with probability 1. The existence of such a process relies on two steps. First to show the existence of the log-beta process and then the existence of F such that $F \in \mathcal{F}$ with probability 1. The proof is on the line of Hjort's (1990) proof for the existence of beta process.

Theorem 4.29 (Walker and Muliere) *Let $G \in \mathcal{F}$ be continuous and let $c(\cdot)$ be a piecewise continuous positive function. Then*

(i) *There exists an independent nonnegative increment process Z with Levy representation given by*

$$\log \mathcal{E}\left(e^{-\theta Z(t)}\right) = -\int_o^\infty \left(1 - e^{-v\theta}\right) dN_t(v), \tag{4.7.6}$$

where

$$dN_t(v) = \frac{1}{(1-e^{-v})} \int_o^t \exp(-vc(s)\,G[s,\infty))c(s)\,dG(s)\,dv. \tag{4.7.7}$$

(ii) *If $F(t) = 1 - \exp(Z(t))$, then $F \in \mathcal{F}$ with probability 1.*

Proof The following are the main steps of their proof:

(i) For $i = 1,\ldots,n$ define $a_{ni} = c_{ni}G[\frac{i-1}{n}, \frac{i}{n})$, $b_{ni} = c_{ni}G[\frac{i}{n},\infty)$, and $c_{ni} = c\left(\frac{i-\frac{1}{2}}{n}\right)$. Let $X_{ni} \sim \text{Be}(a_{ni}, b_{ni})$, $W_{ni} = -\log(1-X_{ni})$, and let $Z_n(0) = 0$ and $Z_n(t) = \sum_{\frac{i}{n}\leq t} W_{ni}, t \geq 0$. We wish to show that the sequence $\{Z_n(t)\}$ converges in distribution properly to a Levy process Z with the aforementioned properties. We have, upon evaluating the expectation,

$$\begin{aligned}\log E(\exp(-\theta Z_n(t))) &= \sum_{\frac{i}{n}\leq t} \log E(\exp(-\theta W_{ni})) \\ &= \sum_{\frac{i}{n}\leq t} \log \frac{\Gamma(a_{ni}+b_{ni})\,\Gamma(b_{ni}+\theta)}{\Gamma(b_{ni})\,\Gamma(a_{ni}+b_{ni}+\theta)}.\end{aligned} \tag{4.7.8}$$

Ferguson (1974) has shown that using the formula $\Gamma(x) = x^{-1}\Gamma(x+1)$, the above term involving gamma functions can be written as

$$\prod_{m=0}^{k-1} \frac{(a_{ni}+b_{ni}+\theta+m)(b_{ni}+m)}{(a_{ni}+b_{ni}+m)(b_{ni}+\theta+m)} \cdot \frac{\Gamma(a_{ni}+b_{ni}+k)\Gamma(b_{ni}+\theta+k)}{\Gamma(b_{ni}+k)\Gamma(a_{ni}+b_{ni}+\theta+k)}, \tag{4.7.9}$$

and further using Sterling's formula $\Gamma(x) \sim (2\pi x)^{1/2}(x/e)^x$ for large x, it can be shown that the term involving gamma functions tends to 1 as $k \to \infty$. Thus

$$\begin{aligned} \log E(\exp(-\theta Z_n(t))) &= \sum_{\frac{i}{n} \le t} \sum_{m=0}^{\infty} \log \frac{(a_{ni}+b_{ni}+\theta+m)(b_{ni}+m)}{(a_{ni}+b_{ni}+m)(b_{ni}+\theta+m)} \\ &= \sum_{\frac{i}{n} \le t} \int_0^{\infty} \left(e^{-\theta v}-1\right) \frac{\exp(-b_{ni}v)(1-\exp(-a_{ni}v))}{v(1-\exp(-v))} dv, \end{aligned} \tag{4.7.10}$$

where the second equality follows from the Levy representation

$$\log \frac{\lambda}{\lambda-\phi} = \int_0^{\infty} \left(e^{\phi v}-1\right) \frac{e^{-\lambda v}}{v} dv, \tag{4.7.11}$$

for the moment generating function of the negative exponential distribution with parameter λ [see Ferguson (1974)'s Lemma 1]. Noting that

$$\sum_{\frac{i}{n} \le t} \exp(-b_{ni}v)(1-\exp(-a_{ni}v)) \to v \int_0^t c(s) \exp(-vc(s) G[s,\infty)) dG(s), \tag{4.7.12}$$

as $n \to \infty$, it follows that

$$\begin{aligned} \log E(\exp(-\theta Z_n(t))) &\to \int_0^{\infty} \frac{\left(e^{-\theta v}-1\right)}{(1-\exp(-v))} \int_0^t c(s) \exp(-vc(s) G[s,\infty)) dG(s) dv \\ &= \int_0^{\infty} \left(e^{-\theta v}-1\right) dN_t(v). \end{aligned} \tag{4.7.13}$$

Likewise it can be shown that for any interval $(a_{j-1}, a_j]$,

$$\log E\left(\exp\left(-\sum_{j=1}^{k} \theta_j Z_n(a_{j-1}, a_j]\right)\right) \to \sum_{j=1}^{k} \left(\int_0^1 \left(e^{-\theta_j v}-1\right) dN_{(a_{j-1},a_j]}(v)\right), \tag{4.7.14}$$

implying that the finite dimensional distributions of $\{Z_n\}$ converge properly. Now following the arguments used for the beta process, the existence of the process Z can be concluded.

(ii) To show that such a process F exists, we proceed as follows. Let $Y_{n1} = X_{n1}$ and $Y_{nk} = X_{nk}\prod_{j=1}^{k-1}\left(1 - X_{nj}\right)$, for $k > 1$. Then the joint distribution of $(Y_{n1}, \ldots, Y_{nm})$ is $\mathcal{G}\left(a_{n1}, b_{n1}, \ldots, a_{nm}, b_{nm}\right)$. Letting $F_n(t) = \sum_{\frac{k}{n} \le t} Y_{nk}$ and $F_n(0) = 0$, we have

$$-\log\left(1 - F_n(t)\right) = -\sum_{\frac{k}{n} \le t} \log\left(1 - X_{nk}\right) = \sum_{\frac{k}{n} \le t} W_{nk} = Z_n(t),$$

implying that $\{F_n\}$ is a discrete time beta-Stacy process and $F_n = 1 - \exp\left(-Z_n(t)\right)$. The sequence $\{F_n\}$ converges in distribution to F. $F \in \mathcal{F}$ with probability 1 since $\int_0^\infty dG(s)/G[s, \infty) = \infty$. Further details may be found in their paper.

Note that in defining the neutral to the right process, F was reparametized in terms of Y_t, a non-negative independent increment process. Here, essentially, Y_t is taken to be the log-beta process. Thus the beta-Stacy process is an NTR 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:

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

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

$$\alpha(t) = \int_o^t c(s)\, dG_c(s) + \sum_{t_j \le t} c\left(t_j\right) G\left\{t_j\right\}, \quad \beta(t) = c(t)\, G[t, \infty).$$

As noted above, there is a connection between the beta process and beta-Stacy process. This connection was explicitly stated in Dey et al. (2003) as follows. The prior Π is a beta-Stacy process with parameters (D, F_0) if and only if it is a beta process prior with parameters (C, H_0), where $C = D(1 - F_0)$ and $H_0 = H(F_0)$.

4.7.2 Properties

Here are some of the properties of the beta-Stacy process.

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

$$dN_t(v) = \frac{dv}{(1 - e^{-v})}\int_o^t \exp(-vcG(s, \infty))cdG_c(s) = \frac{e^{-v\alpha(0,\infty)}}{v(1 - e^{-v})}\left(e^{v\alpha(t)} - 1\right)dv,$$

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

$$\log \mathcal{E}\left(e^{-\theta Z(t)}\right) = -\int_o^t \log(1 + \theta/\beta(s))d\alpha(s),$$

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,

$$dL_t(s) = \frac{1}{1-s} dN_t(-\log(1-s)). \tag{4.7.15}$$

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 (4.7.4), the prior mean can be seen to be

$$\begin{aligned}\mathcal{E}(F(t)) &= 1 - \prod_{t_k \le t}\left(1 - \frac{G\{t_k\}}{G(t_k, \infty)}\right)\exp\left(-\int_o^t dG_c(s)/G(s, \infty)\right)\\ &= 1 - \prod_{[0,t]}\left(1 - \frac{dG(s)}{G(s, \infty)}\right) = G(t).\end{aligned} \tag{4.7.16}$$

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 censored observations, is stated in the following theorem:

Theorem 4.30 (Walker and Muliere) *Let $X_1, \ldots, X_n$ 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*

$$G^*(t) = 1 - \prod_{[0,t]}\left\{1 - \frac{c(s)\,dG(s) + dN(s)}{c(s)\,G[s, \infty) + R(s)}\right\}, \tag{4.7.17}$$

$$c^*(t) = \frac{c(t)\,G[t, \infty) + R(t) - N\{t\}}{G^*[t, \infty)}, \tag{4.7.18}$$

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

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 4.31 (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 also 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. 6 and is the same estimator as obtained by Hjort (1990) for the beta process.

4.7.3 Posterior Distribution

In view of the fact that the beta-Stacy process is an NTR process, the description given earlier for the posterior distribution of an NTR is equally valid here and is similar to the one given for the beta process. Again we have sets M of fixed points of discontinuities, $\mathbf{f} = \{f_j\}_{j\geq 1}$ of associated densities of jump at t_j, and Levy measure of the form $dN_t(z) = (\int_0^t a(z,s)ds)dz$. The beta-Stacy process with parameters $\alpha(\cdot)$ and $\beta(\cdot)$ arises when $a(z,s)ds = e^{-z\beta(s)}d\alpha(s)/(1-e^{-z})$. The updated parameters are $M^*, \mathbf{f}^*$ and $a^*(z,s)$ as described in Theorem 4.12. If, however, $X > x$ (as would be in the case of censored observation) the parameters are $M^* = M$,

$$f_j^*(s) = \begin{cases} \kappa e^{-s} f_j(s) & \text{if } t_j \leq x \\ f_j(s) & \text{if } t_j > x, \end{cases} \tag{4.7.19}$$

$$\text{and } a^*(z,s) = \begin{cases} a(z,s)e^{-s} & \text{if } s \leq x \\ a(z,s) & \text{if } s > x, \end{cases} \tag{4.7.20}$$

where κ is the normalizing constant.

This approach would allow one to carry out full Bayesian analysis via simulation.

1. The densities to be sampled corresponding to the jumps points of M^* are of form

$$f_j^*(z) \quad \propto \quad (1-e^{-z})^{\lambda} e^{-\mu z} \text{ with integers } \lambda, \mu \geq 0 \tag{4.7.21}$$

and

$$f_j^*(z) \quad \propto \quad (1-e^{-z})^{\lambda}\, a(z,s), \quad \text{with integer } \lambda > 0. \tag{4.7.22}$$

These can be sampled via Gibbs sampler algorithm.

2. The random variable Z corresponding to the continuous increment of the interval $[a,b)$ is infinitely divisible and therefore the algorithm developed by Damien et al. (1995) can be used to sample its density as described earlier in connection with ID distributions. However, Walker and Damien (1998) introduce an alternative approach based on the fact $Z \stackrel{d}{=} \int_0^{\infty} z dP(\cdot)$, where $P(\cdot)$ is a Poisson process with mean measure $dz \int_{[a,b)} a(z,s)ds$. It is similar to the method of Bondesson (1982) described earlier. Details may be found in their paper along with a numerical example in which they rework Kaplan and Meier (1958) data and compare the results with those of Ferguson and Phadia (1979).

4.7.3.1 Characterization

Recall that for the Dirichlet process we had $\mu(t) = \mathcal{E}[F(t)] = F_0(t)$ and $\text{Var}[F(t)] = F_0(t)(1-F_0(t))/(M+1)$ and therefore we may set the parameter of the DP as $\alpha(\cdot) = MF_0(\cdot)$. The neutral to the right process was described in terms of the stochastic process with Lévy measure $N_t(\cdot)$. As $\mu(t)$ and $\text{Var}[F(t)]$ characterizes the Dirichlet process, Walker and Damien (1998) define two functions $\mu(t)$ and $\text{Var}[F(t)]$ in terms of the Lévy measure that characterizes 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, consider functions

$$\mu(t) = -\log \mathcal{E}[S(t)] = -\log \mathcal{E}[e^{-Z(t)}] = \int_0^{\infty} (1-e^{-z}) dN_t(z) \tag{4.7.23}$$

and

$$\lambda(t) = -\log \mathcal{E}[S^2(t)] = \int_0^{\infty} (1-e^{-2z}) dN_t(z), \tag{4.7.24}$$

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

$$dN_t(z) = (1-e^{-z})^{-1} \int_0^t e^{-z\beta(s)} d\alpha(s) dz, \tag{4.7.25}$$

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

$$\mu(t) = \int_0^\infty \int_0^t e^{-z\beta(s)} d\alpha(s) dz \quad \text{and} \quad \lambda(t) = \int_0^\infty \int_0^t \frac{(1 - e^{-2z})}{(1 - e^{-z})} e^{-z\beta(s)} d\alpha(s) dz. \tag{4.7.26}$$

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.

4.8 NB Models for Machine Learning

The aim of machine learning is to develop computational programs which would improve performance given the observed data. To facilitate this task, the search is for generative probabilistic models for discrete data, such as text corpora, which would allow the practitioner to detect hidden structures, patterns, or clusters. Nonparametric Bayesian methods provide a rich tools kit for such tasks and have been increasingly becoming popular in such exploration, see, for example, Shahbaba and Neal (2009), Hannah et al. (2011), Wade et al. (2014), Blei and Frazier (2011) and Blei et al. (2003). These authors have shown the benefit of using Dirichlet process mixtures. However, it is clear that other processes presented in this book and illustrated their use at various places, may also be used.

Generally, the use of mixture models assumes that one latent cause is associated with each observation or data point. This assumption can be quite restrictive (Titsias 2008). As an alternative to multinomial representation underlying mixture models, factorial models which associates each data point a set of latent variables seems to be more preferable in featural modeling. An added advantage is that we do not need to know the cardinality of features a priori. Besides the CRP mentioned earlier in Sect. 2.1, Indian buffet and infinite gamma-Poisson processes in particular have garnered a lot of interest from outside the statistical community, and found useful in a wide range of interesting applications. These processes, along with the Bayesian nonparametric hierarchical modeling (as is related to machine learning) have proved to be indispensable tools in solving problems in the areas such as information retrieval and word segmentation.

The CRP is a marginal distribution in the case of DP prior and it is characterized as a prior distribution over partitions of N objects distributed in K classes. This can be conveniently represented by an $N \times K$ matrix, each row having a single entry of one and rest of the entries of zeros. In featural models, the columns may serve as features. To permit the possibility of each object or subject (data point) may have multiple features, Griffiths and Ghahramani (2006) proposed a process called the IBP, which is essentially a distribution over a class of equivalent binary matrices. Titsias (2008) saw this as a limitation. To deal with features having multiple occurrences, a natural generalization would be to replace each non-zero entry with an entry of positive integer. Titsias (2008) followed this extension and developed a process called *infinite gamma-Poisson process.* It is interesting to note that the conjugacy of the DP to multinomial sampling results in the CRP, conjugacy of beta process to Bernoulli process yields the IBP, and conjugacy of gamma process to Poisson process sampling yields the infinite gamma-Poisson process.

These processes are found to be useful in representing featural models where potentially there are unlimited number of features, i.e. each data point is associated with a set of possibly unlimited number of features. Thibaux (2008) discusses these models in the context of machine learning applications, and develops efficient sampling procedures which are shown to be faster than Gibbs sampling methods. Our objective in this section is limited to presenting the developmental aspects of these processes.

4.8.1 Chinese Restaurant Process

The CRP is a process of generating a sample from the Dirichlet process and is equivalent to the extended Polya urn scheme introduced by Blackwell–MacQueen and discussed in Sect. 2.1.1. In the restaurant analogy, it is a sequential process of assigning N customers to K tables, K potentially unbounded, each assignment being independent. This is same thing as allocating N objects to K classes or cells independently, or grouping N objects into K clusters. In machine learning and other applications it is generally assumed that K is unbounded. A formal derivation of the process given by Griffiths and Ghahramani (2011) is as follows. The strategy is to consider first the allocation when K is finite and then take the limit $K \to \infty$.

Let the vector $\mathbf{c} = (c_1, \ldots, c_N)$ represent allocations of these N objects, with $c_i = k \in \{1, \ldots, K\}$ indicating the assignment of object i to class k with probability, say $p_k \geq 0$, such that $p_1 + \cdots + p_K = 1$. That is $\mathcal{P}(c_i = k) = p_k, k = 1, \ldots, K$. Let $p = (p_1, \ldots, p_K)$. The joint distribution of $\mathbf{c}$ is

$$p(\mathbf{c}|\mathbf{p}) = p(c_1, \ldots, c_N|\mathbf{p}) = \prod_{i=1}^{N} p(c_i|p_i) = \prod_{k=1}^{K} p_k^{n_k}, \tag{4.8.1}$$

where $n_k = \sum_{i=1}^{N} I[c_i = k]$ is the number of objects assigned to class $k, k = 1, \ldots, K$.

In Bayesian analysis, p may be treated as a parameter vector and it is customary to assign a conjugate prior, namely the Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_K, D(\alpha_1, \ldots, \alpha_K)$. Since we are going to allow K to be potentially unbounded, it is the usual practice to take the Dirichlet distribution to be symmetric with parameter vector $(\alpha/K, \ldots, \alpha/K)$, where $\alpha > 0$ is a positive measure. By integrating out the $\mathbf{p}$ vector, we get the distribution of $\mathbf{c}$ as

$$p(\mathbf{c}|\alpha) = \frac{\prod_{k=1}^{K} \Gamma\left(n_k + \frac{\alpha}{K}\right)}{\Gamma\left(\frac{\alpha}{K}\right)^K} \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}.$$

Expanding the gamma function and using the recursion relation $\Gamma(x) = (x-1)\Gamma(x-1)$, we can write $\Gamma(n_k + \alpha/K) = \prod_{j=1}^{n_k - 1} (j + \alpha/K)(\alpha/K)\Gamma(\alpha/K)$. Cancelling the like terms, the above expression simplifies to

$$p(\mathbf{c}|\alpha) = \left(\frac{\alpha}{K}\right)^{K_+} \prod_{k=1}^{K_+} \prod_{j=1}^{n_k - 1} (j + \alpha/K) \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}, \tag{4.8.2}$$

where K_+ is the number of classes with at least one object, and the indices are reordered so that $n_k > 0$ for all $k \leq K_+$. There K^N possible assignment values of $\mathbf{c}$. Assuming that the assignment is made randomly, the probability of any one particular set of assignment would be $1/K^N$ which goes to zero as K tends to ∞. Since N is finite and $K_+ \leq N, p(\mathbf{c}|\alpha) \to 0$ as $K \to \infty$. For this reason, the authors define a distribution over equivalence classes of assignment vectors instead of the vectors themselves. This defines a joint distribution for all class assignments $\mathbf{c}$ in which individual class assignments are not independent, but are exchangeable with the probability of an assignment vector remains the same when the indices of the objects are permuted. For example, for $N = 4$, assignments $(1, 2, 1, 3)$ and $(2, 1, 2, 3)$ produce the same 3 classes except for the labelling of classes. Therefore consider the distribution of $\mathbf{c}$ over the equivalence class of class assignments. Now suppose that we partition N objects into K_+ classes but have $K = K_0 + K_+$ labels to assign to those subsets. Then there are $K!/K_0!$ assignment vectors $\mathbf{c}$ that belong to the same equivalence class, $[\mathbf{c}]$, and probability of each of them will be the same. Therefore, adding over them results in the distribution of assignment vectors over the class of equivalence as

$$p([\mathbf{c}]|\alpha) = \frac{K!}{K_0!} \left(\frac{\alpha}{K}\right)^{K_+} \prod_{k=1}^{K_+} \prod_{j=1}^{n_k - 1} \left(j + \frac{\alpha}{K}\right) \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}.$$

Now taking the limit as $K \to \infty$ and noting that $K!/(K_0!K^{K+}) \to 1$, we get

$$\lim_{K\to\infty} p([\mathbf{c}]|\alpha) = \alpha^{K+} \prod_{k=1}^{K^+} (n_k - 1)! \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}, \tag{4.8.3}$$

which serves as a prior over class assignments for an infinite mixture model.

The above formula can also be derived using the distribution of Polya urn sequence (see Sect. 2.1.1) $c_1, c_2, \ldots,$ with parameter α. Recall that the conditional distribution of c_i given $c_1, \ldots, c_{i-1}$ is given by

$$c_i|c_1, \ldots, c_{i-1} \sim \frac{\alpha(\cdot) + \sum_{i=1}^{n} \delta_{c_i}(\cdot)}{\alpha + i - 1}.$$

Therefore the joint distribution of $c_1, \ldots, c_N$ is

$$\begin{aligned} p(c_1, \ldots, c_N) &= \frac{\alpha(c_1)}{\alpha} \prod_{i=2}^{N} \frac{(\alpha + \sum_{j=1}^{i-1} \delta_{c_j})(c_i)}{(\alpha + i - 1)} \\ &= \left[\alpha(c_1) \prod_{i=2}^{N} \left(\alpha + \sum_{j=1}^{i-1} \delta_{c_j} \right)(c_i) \right] \cdot \left[\prod_{i=1}^{N} \frac{1}{\alpha + i - 1} \right]. \end{aligned} \tag{4.8.4}$$

The second product can be seen to be $\Gamma(\alpha)/\Gamma(N+\alpha)$. We expect ties among the Nc's. Let the K distinct c's be denoted by $c_1^*, \ldots, c_K^*$ with c_k^* having n_k ties, $k = 1, \ldots, K$ as before. Noting that $\alpha(c_i^*) = \alpha/K$ for all $i = 1, \ldots, K$, the first term in the second line of the above equation becomes

$$\alpha(c_1) \prod_{i=2}^{N} \left(\alpha + \sum_{j=1}^{i-1} \delta_{c_j} \right)(c_i) = \prod_{k=1}^{K+} \frac{\alpha}{K} \prod_{j=1}^{n_k - 1} \left(\frac{\alpha}{K} + j \right). \tag{4.8.5}$$

Putting two terms together, we get the expression (4.8.2) for the joint distribution of $c_1, \ldots, c_N$ derived earlier.

The Chinese restaurant analogy described before is as follows. Customers or patrons enter the restaurant one after another and each choose a table at random. The first customer chooses the first table with probability $\alpha/\alpha = 1$. The next customer joins the occupied table with probability $1/(\alpha + 1)$ and chooses a second table with probability $\alpha/(\alpha + 1)$. The $(n+1)$-th customer chooses to join previous customers with probability $n/(\alpha + n)$ or chooses a new table with probability $\alpha/(\alpha + n), n = 1, \ldots, N$. If he joins previous customers and there are already m tables occupied, then he joins the k-th table, $k = 1, 2, \ldots, m$, with probability proportional to the number of customers already occupying the table, that is, with probability $n_k/(\alpha + n)$, where n_k is the number of customers sitting at that table.

This defines the conditional probability of the $(n+1)$-th customer occupying the k-th table. The process continues until all N customers are seated occupying K tables. This is same as allocating N objects (patrons) to K cells (tables) in a sequential manner. The joint distribution of this allocation results in the same 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 and not on the order in which they arrive and sit. As n increases, there is a greater probability that the next patron will choose an existing table rather than a new table. After N steps, the output of the CRP is a partition of N customers across K tables, or partitioning 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 the partitions. Partitions are realizations of the CRP. The DP induces an exchangeable distribution over partitions. The expected number of tables K occupied by first N customers is $\sum_{i=1}^{N}[\alpha/(\alpha+i-1)]$ (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 is obtained by integrating out the random probability measure drawn from the Dirichlet process and thus it describes the marginal distributions in terms of random partitions determined by K tables in a restaurant. Thus the DP may be viewed as the de Finetti measure of the CRP. Samples from the Dirichlet process are probability measures and samples from the CRP are partitions. Teh et al. (2006) proposed a further generalization as *franchised CRP* which corresponds to a hierarchical Dirichlet process in which the base distribution F_0 of the Dirichlet process is itself considered as having a Dirichlet process prior with hyper parameters, say, M^* and G_0. This was presented in Sect. 2.4.2 as *Chinese Restaurant Franchise.* In the CRP franchise, we have a group of restaurant serving a common set of dishes. Patrons stream in, sit at a table in one of the restaurants and share a common dish from the global menu across restaurants. This process describes marginals under a hierarchical DP when G_j and G_0 are integrated out.

4.8.2 *Indian Buffet Process*

In the CRP each object (patron) can possess only one feature (choose one table or order one dish), and was represented having a single entry of one in each row of the $N \times K$ matrix and rest of the entries as zeros, where rows represent objects and columns as features. However, in certain applications such as factorial or featural modeling, each object may possess an unlimited number of features, therefore a generalization of the above model is desired. This leads to the development of the *Indian Buffet process* proposed by Griffiths and Ghahramani (2006). 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 contrast to the CRP, here a patron

may choose any number of dishes and the number N of patrons is fixed. It can be viewed as a factorial analog of the CRP.

The IBP is essentially a process to define a prior distribution on the equivalence class of sparse binary matrices (entries of the matrices have binary responses) consisting of a finite number of rows and an unlimited number of columns. 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. For such factorial models, a set of latent Bernoulli variables are associated with each data point. The advantage is that one need not know the cardinality of features beforehand. Since the rows are exchangeable, they form an exchangeable sequence. It is shown that the underlying mixing de Finetti measure for such a sequence is the beta process, as the DP is for the CRP. The IBP also provides an algorithm to sample beta processes. An expanded review of the process and its applications can be found in their subsequent paper (Griffiths and Ghahramani 2011).

Let $\mathbf{Z}$ be a binary response matrix of N rows and an unlimited number of columns. Its elements z_{ik} denote the fact that object i possess 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 approach adopted 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, $z_{ik} \sim \text{Ber}(\mu_k), k = 1, 2, \ldots, K$ 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

$$\mathcal{P}(\mathbf{Z} = \mathbf{z}|—) = \prod_{k=1}^{K} \prod_{i=1}^{N} P(z_{ik}|\mu_k) = \prod_{k=1}^{K} \mu_k^{m_k} (1 - \mu_k)^{N - m_k}, \tag{4.8.6}$$

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

$$\mathcal{P}(\mathbf{Z} = \mathbf{z}) = \prod_{k=1}^{K} \frac{\frac{\alpha}{K} \Gamma\left(m_k + \frac{\alpha}{K}\right) \Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}. \tag{4.8.7}$$

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 is obvious. Patrons stream in one by one and taste dishes. 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

$$\mathcal{P}(\mathbf{Z} = \tilde{\mathbf{z}}) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!} P(\mathbf{Z} = \mathbf{z}), \tag{4.8.8}$$

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

$$\mathcal{P}(\mathbf{Z} = \tilde{\mathbf{z}}) = \frac{\alpha^{K_+}}{\prod_{h=0}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N - m_k)!\,(m_k - 1)!}{N!}, \tag{4.8.9}$$

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 ordered components $p_1 > p_2 > \cdots$.

A random draw from the Dirichlet process can be obtained by any one of the methods mentioned in Sect. 2.1. 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 (α). The $(n+1)$-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 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 (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

$$\mathcal{P}(\mathbf{Z}=\mathbf{z}) = \frac{\alpha^{K+}}{\prod_{n=1}^{N} K^{(n)}!} \exp\{-\alpha H_N\} \prod_{k=1}^{K+} \frac{(N-m_k)!\,(m_k-1)!}{N!}, \tag{4.8.10}$$

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. (4.8.9) (see Griffiths and Ghahramani 2006 for further details).

4.8.2.1 Stick-Breaking Construction of IBP

To sample a binary matrix from the distribution of $\mathbf{Z}$ 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 . These ordered μ_k's are given in terms of beta random variables θ_k's. Let $\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$ be a decreasing reordering of $\mu_1, \mu_2, \ldots, \mu_K$, where a $\text{Be}(\frac{\alpha}{K}, 1)$ prior is placed on each μ_k. As $K \to \infty$, Teh et al. (2007) construct a stick-breaking representation (slightly different from the earlier SB construction) of the IBP as follows. Let

$$\theta_k \overset{\text{iid}}{\sim} \text{Be}(\alpha, 1); \qquad \mu_k = \theta_k \mu_{k-1} = \prod_{l=1}^{k} \theta_l, \tag{4.8.11}$$

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 $\text{Be}(\alpha, 1)$, 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 $\theta_2 \sim \text{Be}(\alpha, 1)$ 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.

The connection of this process to the one for the Dirichlet process is as follows. If at stage k, we denote the length of the discarded piece as p_k, then we have

$$p_k = (1-\theta_k)\mu_{k-1} = (1-\theta_k) \prod_{l=1}^{k-1} \theta_l. \tag{4.8.12}$$

Now making a change of variable $V_k = 1 - \theta_k$, we have $V_k \overset{\text{iid}}{\sim} \text{Be}(1, \alpha)$ 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 IBP, 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 $\theta_k \sim \text{Be}(\alpha + k\sigma, 1 - \sigma)$ and $\mu_k = \prod_{l=1}^{k} \theta_l$.

A two-parameter generalization of IBP is derived by Ghahramani et al. (2007) by assigning a $\text{Be}\left(\frac{\alpha\beta}{K}, \beta\right)$ (instead of $\text{Be}\left(\frac{\alpha}{K}, 1\right)$) prior to each μ_k. In the metaphor of Indian buffet, the first customer starts at left of buffet lay out and samples Poisson(α) 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 parameter α reflects the average number of dishes tried (have features) by the customers. The expected overall total number of dishes (features) tried is a function of β and it increases as β increases (for fixed n). Therefore, the authors interpret β as the feature *repulsion* parameter.

Teh and Gorur (2009) introduce a three-parameter generalization of the IBP with power-law behavior similar to the stable beta process. 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 (α) 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 (αr) ,where $r = (\Gamma(1 + \beta)\Gamma(n - 1 + \beta + \sigma))/(\Gamma(n + \beta)\Gamma(\beta + \sigma))$. This is similar to the Pitman–Yor process (Sect. 3.4.2) where σ was subtracted from the number of customers seated around each table and added to the prospect of sitting at a new table. The mass parameter α controls the total number of dishes tried by the patrons, concentration parameter β controls the number of customers that will try each dish, and the stability parameter σ controls the power-law behavior of the process.

It is shown that the total number of dishes tried by n customers is $O(n^\sigma)$ and the proportion of dishes tried by m customers is asymptotically $O\left(m^{-1-\sigma}\right)$. However, the number of dishes each customer tries is simply a Poisson (α) distributed random variable. When $\sigma = 0$, it reduces to the two parameter IBP. This is akin to Pitman–Yor process discussed in the last chapter 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 two parameter IBP and show that their power-law IBP is a good model for word occurrences in document corpora. For such a model, let n be the number of documents in a corpus and let $Z_i(\{\theta\}) = 1$ if word type θ appears in document i, and 0 otherwise. Let $\mu(\{\theta\})$ be the appearance probability of word type θ among the documents. Now assume a stable-beta prior on μ and each document modeled as a conditionally independent Bernoulli process draw. The joint distribution of the

word appearance $Z_1, \ldots, Z_n$, integrating out μ results in the IBP joint probability given above.

IBP is connected to the beta process B (discussed in Sect. 4.6), in the same way the CRP is connected to the Dirichlet process. It is an iid mixture of the Bernoulli processes with mixing measure the beta process. 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 base measure, and given B, let Z_i be distributed as the Bernoulli process with parameter B (Sect. 4.6). That is, $B \sim \mathrm{BP}(c, B_0)$ and $Z_i|B \sim \mathrm{BeP}(B)$, for $i = 1, \ldots, n$ are independent Bernoulli draws from B. Integrating out B, we have the marginal predictive distribution as

$$Z_{n+1}|Z_1, \ldots, Z_n, c, B_0 \sim \mathrm{BeP}\left(\frac{c}{c+n}B_0 + \sum_{k=1}^{K}\frac{m_k}{c+n}\delta_{\omega_k}\right), \tag{4.8.13}$$

where m_k is the number of customers among n having tried dish k and ω_k stands for a patron selects dish k. Its interpretation in the terminology of IBP is as follows. Suppose B_0 is continuous (if not adjustment needs to be made as indicated earlier) and c is constant such that $\gamma = B_0(\Omega)$ is finite. Since $Z_1 \sim \mathrm{BeP}(B_0)$ and B_0 is continuous, Z_1 is a Poisson process (B_0), and the total number of features of Z_1 is $Z_1(\Omega) \sim P(\gamma)$. That is the first customer will taste Poisson (γ) 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. $U \sim \mathrm{BeP}(\sum_k \frac{m_k}{c+n}\delta_{\omega_k})$ and $V \sim \mathrm{BeP}(\frac{c}{c+n}B_0)$.$U$ 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 Poisson $\left(\frac{c\gamma}{c+n}\right)$ number of new dishes.

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

Broderick et al. (2013) give an excellent account of combinatorial stochastic representations for the feature modeling, which includes the beta process and the IBP, thus extending the similar representations for the partition modeling which include the Dirichlet process and the CRP.

4.8.3 Infinite Gamma-Poisson Process

In both the CRP and IBP, we are dealing with binary matrices with each row having a single entry of one in the case of CRP and multiple entries of ones in the case of IBP, rest of the entries being zeros. In practice this may be inefficient to model the generation mechanism of data such as images. Titsias (2008) proposed a model based on gamma-Poisson distribution, which may be seen as an obvious generalization of IBP in which multiple repetition or occurrences of the same feature

is allowed. That is an entry of one is replaced by an entry of positive integer. In restaurant analogy this means that a patron may go for the same dish, say chicken curry or nan, more than once. His approach in deriving the probability distribution over such matrices is similar to the IBP, which is to consider first K to be finite and then taking the limit as K goes to infinity. This produces the distribution over the equivalence classes of non-negative integer valued matrices, which is equivalent to the distribution over partitions of objects by the DP, using Ewen's (1972) distribution.

Now given K features, $\mathbf{Z}$ is an $N \times K$ matrix with non-negative integer valued entries z_{ik}. Now assume $z_{ik} \sim \text{Poisson}(\mu_k)$ with μ_k as a feature specific parameter, $k = 1, 2, \ldots, K$, for each $i = 1, 2, \ldots, N$. Thus

$$\mathcal{P}(\mathbf{Z} = \mathbf{z}|—) = \prod_{k=1}^{K}\prod_{i=1}^{N}\text{Poisson}(z_{ik}|\mu_k) = \prod_{k=1}^{K}\frac{\mu_k^{m_k}\exp(-N\mu_k)}{\prod_{i=1}^{N} z_{ik}!}, \tag{4.8.14}$$

where $m_k = \sum_{i=1}^{N} z_{ik}$ is the number of objects possessing feature k. Further assume $\mu_k \sim G(\frac{\alpha}{K}, 1)$. Integrating out parameters $\{\mu_k\}$, we get

$$\mathcal{P}(\mathbf{Z} = \mathbf{z}|\alpha) = \prod_{k=1}^{K}\frac{\Gamma\left(m_k + \frac{\alpha}{K}\right)}{\Gamma\left(\frac{\alpha}{K}\right)(N+1)^{m_k+\frac{\alpha}{K}}\prod_{i=1}^{N} z_{ik}!},$$

which shows that the columns of $\mathbf{Z}$ are independent. Note that this distribution is exchangeable since reordering rows of $\mathbf{Z}$ does not change the probability. Since the columns are independent, the expectation of sum of all elements of $\mathbf{Z}$ is $K\sum_{i=1}^{N}\mathcal{E}(z_{ik}) = \alpha N$, since

$$\mathcal{E}(z_{ik}) = \sum_{z_{ik}=0}^{\infty} z_{ik}\text{NB}\left(z_{ik}; \frac{\alpha}{K}, \frac{1}{2}\right) = \frac{\alpha}{K},$$

where $\text{NB}(z_{ik}; r, p)$ denotes a negative binomial distribution with parameters $r > 0$ and $0 < p < 1$. The expectation of sum of all elements of $\mathbf{Z}$ is independent of K and as K increases to infinity, the matrix $\mathbf{Z}$ gets sparser. α controls the sparsity of the matrix and may be treated as sparsity parameter. This can alternatively derived by letting $T_i \sim \text{Poisson}(\lambda)$, the vector

$$(z_{i1}, \ldots, z_{iK}) \sim \binom{T_i}{z_{i1}, \ldots, z_{iK}}\prod_{k=1}^{K}\theta_k^{z_{ik}}, i = 1, \ldots, N,$$

$(\theta_1, \ldots, \theta_K) \sim D\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$, and $\lambda \sim G(\alpha, 1)$. Integrating out $\theta's$ and λ yield the above probability. This process generates first a gamma random variable and multinomial parameters and then samples the rows of $\mathbf{Z}$ independently by using the Poisson-multinomial pair.

Now in order to define a distribution over infinite dimensional non-negative integer valued matrices, the author for the same reason as in the IBP considers a class of equivalence matrices and shows that

$$\mathcal{P}(\mathbf{Z}=\tilde{\mathbf{z}}) = \frac{K!}{\sum_{h=0}^{c^N-1} K_h!} \prod_{k=1}^{K} \frac{\Gamma\left(m_k+\frac{\alpha}{K}\right)}{\Gamma\left(\frac{\alpha}{K}\right)(N+1)^{m_k+\frac{\alpha}{K}} \prod_{i=1}^{N} z_{ik}!}, \tag{4.8.15}$$

where c is a sufficiently large integer such that $z_{ik} \leq c-1$ holds, and $h = (z_{1k}, \ldots, z_{Nk})$ as the integer number associated with column k that is expressed in a numerical system with basis c. Now taking the limit $K \to \infty$ and using the same strategy used in the case of Dirichlet-multinomial pair, the author show that

$$\mathcal{P}(\mathbf{Z}=\mathbf{z}|\alpha) = \frac{1}{\sum_{h=0}^{c^N-1} K_h!} \frac{\alpha^{K_+}}{(N+1)^{m+\alpha}} \prod_{k=1}^{K_+} \frac{(m_k-1)!}{\prod_{i=1}^{N} z_{ik}!}, \tag{4.8.16}$$

where $m = \sum_{k=1}^{K_+} m_k$. This probability defines an exchangeable joint distribution over non-negative integer valued matrices with infinitely many columns in a left-ordered form.

Similar to the cases of CRP and IBP, the above distribution can be derived from a stochastic process that constructs the matrix $\mathbf{Z}$ sequentially as the data arrive one by one in a fixed order. However, an additional step is required here at each stage compared to the IBP. It is not enough to sample the number of dishes tasted alone, but also how frequently each dish is tasted. This is accomplished by using the Ewens (1972) distribution.

At the start, all features are unrepresented. Draw an integer number r_1 from the negative binomial distribution $\mathrm{NB}(r_1;\alpha,\frac{1}{2})$ which has the mean value α. It is the total number of feature occurrences for the first data point. Now given r_1, select randomly a partition $(z_{1k}, \ldots, z_{1K_1})$ of r_1 such that $z_{1k} + \cdots + z_{1K_1} = r_1$, and $1 \leq K_1 \leq r_1$ according to Ewens distribution given by

$$p\left(z_{1k}, \ldots, z_{1K_1}\right) = \alpha^{K_1} \frac{\Gamma(\alpha)}{\Gamma(r_1+\alpha)} \frac{r_1!}{\prod_{j=1}^{K_1} z_{1j}} \prod_{i=1}^{s} \frac{1}{a_i^{(1)}!},$$

where s is the number of distinct integers in the set $\{z_{1k}, \ldots, z_{1K_1}\}$ and $a_i^{(1)}$ is the multiplicity of integer i in the partition $(z_{1k}, \ldots, z_{1K_1})$. At the n-th stage, let K_{n-1} denote the number of represented features so far. For each $k \leq K_{n-1}$, draw z_{nk} according to $\mathrm{NB}(z_{nk}; m_k, \frac{n}{n+1})$ which has mean value $\frac{m_k}{n}$, where $m_k = \sum_{i=1}^{n-1} z_{ik}$ is the popularity measure of the k-th feature. Having sampled all the represented features, draw $r_n \sim \mathrm{NB}(r_n;\alpha,\frac{n}{n+1})$ of unrepresented features, and partition it by drawing from the Ewens' distribution as was done at stage 1. This process produces

the distribution

$$\mathcal{P}(\mathbf{Z} = \mathbf{z}|\alpha) = \frac{1}{\prod_{n=1}^{N}\prod_{i=1}^{r_n} a_i^{(n)}!} \frac{\alpha^{K+}}{(N+1)^{m+\alpha}} \prod_{k=1}^{K+} \frac{(m_k - 1)!}{\prod_{n=1}^{N} z_{nk}!}, \tag{4.8.17}$$

where $\{a_i^{(n)}\}$ are the integer-multiplicities for the n-th data point. Note that this distribution does not have the same form as (4.8.16) since it depends on the order the data arrives. Like in the case of IBP, if we consider only the left-ordered class of matrices generated then we obtain the distribution (4.8.16).

Titsias gives an MCMC algorithm for sampling the posterior distribution of $\mathbf{Z}$ given the data, using mainly the Gibbs type sampling from conditional posterior distributions.

Chapter 5
Tailfree Processes

In Chap. 1 it was indicated that there is a third method of constructing priors for a random probability measure P, which is based on an independence property of a sequence of nested partitions of the real line R. In this chapter we present such priors called *tailfree* and *Polya tree* processes and their properties, and point out their advantages and shortcomings in practical applications. In addition, a bivariate extension of the Polya tree process is also presented in Sect. 5.3.

5.1 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 (1964, 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 and 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. 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. Additional shortcomings are pointed out in property 8. Furthermore, it is difficult to derive resulting expressions in closed form and the

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1_5

parameters involved are difficult to interpret adequately. 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 the Dirichlet priors.

5.1.1 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 $\cup_0^\infty \pi_m$ 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 5.1 (Ferguson) The distribution of a random probability P on $(R, \mathcal{B})$ is said to be tailfree with respect to $\{\pi_m\}$ if there 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 $B_1 \supset B_2, \ldots, \supset B_m$, then $P(B_m) = \prod_{j=1}^{m} V_{j,B_j}$.

Another simple 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 $P\left(B_{mj}|B_{m-1,j(m-1)}\right)$ where $B_{m-1,j(m-1)}$ 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 $\pi_0 = (0, 1],\ \pi_1 = \{(0, p_1], (p_1, 1]\},\ \pi_2 = \{(0, p_1], (p_1, p_2], (p_2, 1]\} \ldots$.

An important special case of tailfree processes on $(0, 1]$ is when the partitions points are taken as the dyadic rationals.

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 $Y, 0 \le Y \le 1$ 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

$$P(B) = \left(\prod_{j=1,\epsilon_j=0}^{m} Y_{\epsilon_1\epsilon_2\ldots\epsilon_{j-1}} \right) \left(\prod_{j=1,\epsilon_j=1}^{m} (1 - Y_{\epsilon_1\epsilon_2\ldots\epsilon_{j-1}}) \right), \tag{5.1.1}$$

where for $j = 1$, $Y_{\epsilon_1\epsilon_2\ldots\epsilon_{j-1}}$ stands for Y. For example, $P\left(\left(\frac{1}{2}, \frac{5}{8}\right]\right) = (1-Y)\,Y_1 Y_{10}$. 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 $P(B_{\epsilon_1\epsilon_2\ldots\epsilon_m \underbrace{0000\ldots}_{j \text{ terms}}}) = P(B_{\epsilon_1\epsilon_2\ldots\epsilon_m}) \prod_j Y_{\epsilon_1\epsilon_2\ldots\epsilon_m \underbrace{0000\ldots}_{j \text{ terms}}} \cdots \xrightarrow{\text{a.s.}} 0$.

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 $Y_{\epsilon_1\epsilon_2\ldots\epsilon_m} \sim \text{Be}(\alpha_{\epsilon_1\epsilon_2\ldots\epsilon_m 0}, \alpha_{\epsilon_1\epsilon_2\ldots\epsilon_m 1})$, for some suitable nonnegative real numbers α's, the process yields a Polya tree process, discussed in the next section.

5.1.2 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 $\text{Be}(\alpha(A_{mi}), \alpha(R) - \alpha(A_{mi}))$. Therefore we have to show that there exist random variables Z's such that $\prod_{j=1}^{m} Z_{ji}$ is also distributed as $\text{Be}(\alpha(A_{mi}), \alpha(R) - \alpha(A_{mi}))$. 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 $\prod_{j=1}^{m} Z_{ji} \sim \text{Be}(\alpha(A_{mi}), \alpha(R) - \alpha(A_{mi}))$ as was to be shown.

2. 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.
3. Tailfree processes are conjugate. Ferguson's (1974) Theorem 2.2 is

   Theorem 5.2 *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 in the definition, given $X_1, \ldots X_n$ can easily be calculated.
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 partitions 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 partition π_m 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}, ∞).

5.2 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. As in the case of 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. Paddock et al. (2003) indicate multidimensional extensions of the Polya tree. They also propose randomized Polya trees to soften the effect of partition points

5.2.1 Definition

Let $E = \{0, 1\}$, $E^0 = \emptyset$, E^m be the m-fold product $E \times \cdots \times E$, $E^* = \cup_0^\infty E^m$ 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 π_0, π_1, ... be a sequence of partitions such that $\cup_0^\infty \pi_m$ 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 $\epsilon = \epsilon_1 \cdots \epsilon_m \in E^*$, let $B_{\epsilon 0}$ and $B_{\epsilon 1}$ be the two pieces into which B_ϵ splits. Degenerate splits are allowed, for example, $B_\epsilon = B_{\epsilon 0} \cup \emptyset$. Here the partition π_m may be viewed as a generalization of Ferguson's set π_m of dyadic intervals of length 2^{-m} mentioned in the previous section.

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 hold:

(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 $\epsilon \in E^m$,

$$P(B_{\epsilon_1 \cdots \epsilon_m}) = \left(\prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_1 \cdots \epsilon_{j-1}} \right) \left(\prod_{j=1;\epsilon_j=1}^{m} \left(1 - Y_{\epsilon_1 \cdots \epsilon_{j-1}}\right) \right), \tag{5.2.1}$$

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 $\epsilon = 0$, 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 $P|\theta_i$. 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.

5.2.2 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 $B_{\epsilon_1 \cdots \epsilon_m 0} = [m\Delta s, (m+1)\Delta s)$ and $B_{\epsilon_1 \cdots \epsilon_m 1} = [(m+1)\Delta s, \infty)$ where $\epsilon_i = 1$ for all $i = 1, 2, \ldots, m$. Let G be a measure on Ω and set $\alpha_{m-10} = \gamma_{m-1} G(B_{\epsilon_1 \cdots \epsilon_m 0})$ and $\alpha_{m-11} = \gamma_{m-1} G(B_{\epsilon_1 \cdots \epsilon_m 1})$,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 $Y_m \sim \text{Be}(\alpha_{m-10}, \alpha_{m-11})$, and let $X_m = Y_m \Pi_{j=1}^{m-1}(1 - Y_j)$. Finally set $A(0) = 0$ and $A(t) = \sum_{m\Delta s \le t} Y_m$. Then A can be considered as a beta process (Hjort 1990) with parameter $c(\cdot) = \gamma(\cdot) G[\cdot, \infty)$ and $A_0(\cdot) = \int_0^{(\cdot)} dG(s)/G[s, \infty)$ viewed at an incremental level of Δs (Walker and Muliere 1997b; Muliere and Walker 1997). The corresponding distribution function results from $F(t) = 1 - \Pi_{m\Delta s \le t}(1 - Y_m) = \sum_{j\Delta s \le t} X_j$.
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. 6.5.4).
4. 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 $\pi_m = \{((i-1)/2^m, i/2^m]; i = 1, \ldots, 2^m\}, m = 1, 2, \ldots..$ As before, let $.\epsilon_1\epsilon_2 \ldots .\epsilon_m$ denote the binary expansion of dyadic rational $\sum_{j=1}^m \epsilon_j .2^{-j}$, where ϵ_j is zero or one, and for short, write $\epsilon_m = \epsilon_1 \cdots \epsilon_m$. If all the Y random variables defined there are taken to be mutually independent distributed as $Y_{\epsilon_{m-1}} \sim \text{Be}\left(\alpha_{\epsilon_{m-1}0}, \alpha_{\epsilon_{m-1}1}\right)$, for nonnegative real numbers α's, then the posterior distributions of Y variables given a sample of size n have the same structure, with $Y_{\epsilon_{m-1}} \sim \text{Be}\left(\alpha_{\epsilon_{m-1}0} + r, \alpha_{\epsilon_{m-1}1} + n - r\right)$, where r is the number of observations falling in the interval $(.\epsilon_{m-1}0, .\epsilon_{m-1}1]$ and $n - r$ falling in the interval $(.\epsilon_{m-1}1, .\epsilon_{m-1}1 + 2^{-m}]$. He points out that the choice of α's have the following consequences.

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

5. 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 since it is a discrete measure. That is, Lavine 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 $\mathcal{P}\left(\text{essSup}_\theta \left|\log\left(p\left(\theta\right)/q\left(\theta\right)\right)\right| < \epsilon\right) > \eta$.
6. Polya trees have an advantage that for some sampling situations where the posterior turns out to be a mixture of Dirichlet processes when a Dirichlet prior is used, Polya trees lead just to a single Polya tree.
7. 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.
8. It enables one to design partitions so that the random distribution centers around a desired known distribution.
9. 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. Second, because of the fixed partitioning, discontinuities are introduced in the predictive distributions. In fact Ferguson (1974) had pointed that a PT can be used to construct a density with respect to the Lebesgue measure that exists with probability one, but it will have discontinuities at all partition points with probability 1. They will have more dramatic effect in higher dimensions.

   To soften the effect of partition points however, Paddock et al. (2003) proposed an interesting new strategy. They introduced randomized Polya trees in which the partition points are no longer taken to be fixed but are random. The randomness is implemented by fudging the bifurcation points via auxiliary random variables, one per each level of the tree, in such a way that the endpoints of subintervals are no longer fixed dyadic rationals but are random close to the dyadic rationals. Thus the subintervals at each level are no longer equal in length. Now the random variables Y are defined in the usual manner. This scheme can be implemented for Polya trees with any partition points and not necessarily restricted to dyadic rationals. It induces smoothness in the resulting distribution.

   Another sour point is that the Polya tree is limited in its extension to higher dimension, although extension to bivariate is not so bad and is included in this chapter later on.
10. 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 $\alpha\left(\cdot\right) = MF_0\left(\cdot\right)$ and $M = \alpha\left(R\right)$. Therefore, $\mathcal{E}\left[F\left(\cdot\right)\right] = F_0\left(\cdot\right) = \alpha\left(\cdot\right)/\alpha\left(R\right)$. Similarly in the case of Polya tree, we can define a probability measure $\beta = \mathcal{E}\left[P\right]$ by defining $\beta\left(B\right) = \mathcal{E}\left[P\left(B\right)\right]$ 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.

$$\beta(B_\epsilon) = \mathcal{E}\left[\left(\prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_1\cdots\epsilon_{j-1}}\right)\left(\prod_{j=1;\epsilon_j=1}^{m}(1-Y_{\epsilon_1\cdots\epsilon_{j-1}})\right)\right]$$
$$= \prod_{j=1;\epsilon_j=0}^{m} \frac{\alpha_{\underline{\epsilon}_{j-1}0}}{\alpha_{\underline{\epsilon}_{j-1}0}+\alpha_{\underline{\epsilon}_{j-1}1}} \prod_{j=1;\epsilon_j=1}^{m} \frac{\alpha_{\underline{\epsilon}_{j-1}1}}{\alpha_{\underline{\epsilon}_{j-1}0}+\alpha_{\underline{\epsilon}_{j-1}1}} \tag{5.2.2}$$

defines β on the elements of $\cup\pi_m$ 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

$$\mathcal{P}[\theta_i \in B_\epsilon] = \mathcal{E}[P[B_\epsilon]] = \mathcal{E}[P(B_{\epsilon_1})P(B_{\epsilon_1\epsilon_2}|B_{\epsilon_1})\cdots P(B_\epsilon|B_{\epsilon_1\ldots\epsilon_{m-1}})]$$
$$= \frac{\alpha_{\epsilon_1}}{\alpha_0+\alpha_1}\cdots\frac{\alpha_{\underline{\epsilon}_m}}{\alpha_{\underline{\epsilon}_{m-1}0}+\alpha_{\underline{\epsilon}_{m-1}1}}. \tag{5.2.3}$$

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, he 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).

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

$$\mathcal{P}\left[\theta_{n+1} \in B_{\underline{\epsilon}_m}|\text{data}\right] = \frac{\alpha_{\epsilon_1}+n_{\epsilon_1}}{\alpha_0+\alpha_1+n}\cdots\frac{\alpha_{\underline{\epsilon}_m}+n_{\underline{\epsilon}_m}}{\alpha_{\underline{\epsilon}_{m-1}0}+\alpha_{\underline{\epsilon}_{m-1}1}+n_{\underline{\epsilon}_{m-1}}}, \tag{5.2.4}$$

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

12. Dråghici 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 be chosen so that $B_1 = (\theta_1, \infty), B_{11} = (\theta_2, \infty), \ldots, B_{11\ldots1} = (\theta_k, \infty)$. Then $P|\text{data} \sim \text{PT}(\Pi, A^*)$ 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.

The parameters α's may be chosen according to how quickly the updated predictive distribution moves away from the prior predictive distribution. If α's are large it will be close 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.

14. An anonymous reviewer has brought to my attention a paper of Paddock et al. (2003) in which it is indicated how the Polya tree can be extended to multidimensional. As an application, the authors give computational scheme to simulate conditional predictive distribution of a vector of random variables $X_1, \ldots, X_k$ given $X_{k+1}, \ldots, X_m, m > k$ based on Polya tree prior which is restricted to a finite level n. The procedure is like the Gibbs sampler. A bivariate Polya tree (Phadia 2007) is presented in the next section and Bayesian estimators with respect to this prior are given in Chaps. 6 and 7.

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 $P, P \sim PT\left(\Pi_\beta, \mathcal{A}_\beta\right)$. 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 $k = 1, 2, \ldots, m$ represent the number of observations in $B_{\epsilon_1 \ldots \epsilon_k}$ (where $B_{\epsilon_1 \ldots \epsilon_m} = B_{mt}$) given that there are n_j observations in B_{mj} (for $j = 1, \ldots, t$).

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

$$\mathcal{P}\left[\theta_{n+1} \in B_{mt} | n_1 \in B_{m1}, \ldots, n_t \in B_{mt}\right] = \frac{\alpha_{\epsilon_1} + r_{1t}}{\alpha_0 + \alpha_1 + n} \frac{\alpha_{\epsilon_1 \epsilon_2} + r_{2t}}{\alpha_{\epsilon_1 0} + \alpha_{\epsilon_1 1} + r_{1t}} \cdots \frac{\alpha_{\underline{\epsilon}_m} + r_{mt}}{\alpha_{\underline{\epsilon}_{m-1} 0} + \alpha_{\underline{\epsilon}_{m-1} 1} + r_{m-1t}}, \tag{5.2.5}$$

where $\alpha_{\epsilon_1 \ldots \epsilon_m}$ is written as $\alpha_{\underline{\epsilon}_m}$.

5.2.3 *Finite and Mixtures of Polya Trees*

Just as Antoniak (1974) (see Sect. 2.3) 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.

Definition 5.3 (Lavine) The distribution of a random probability P is said to be a mixture of Polya trees if there exists 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 $B \in \sigma(\mathfrak{X}), \mathcal{P}(P \in B) = \int \mathcal{P}(P \in B|u) H(du)$. 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 just as the case was in mixtures of Dirichlet processes. 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. 2.4.1, and Ishwaran and James 2001). Similarly, in using the Polya Tree prior, truncation is recommended by Lavine (1994). This is possible by choosing α's appropriately so that they increase rapidly towards 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 accuracy. For example, up to level M, define random distribution G according to the Polya tree scheme and there after

it may be defined either as uniform distribution 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 and give an 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 $P(B_{\epsilon_1 \cdots \epsilon_M})$ for each $\epsilon_1 \cdots \epsilon_M$ according to (5.2.3). Then one of the 2^M sets is picked according to the random weights $P(B_{\epsilon_1 \cdots \epsilon_M})$ 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. Details may be found in their paper.

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 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 justified the efficiency of using mixtures of Polya trees alternative to using parametric models and provided computational strategies to carry out the analysis and illustrated them by considering several examples.

There are numerous papers published since 2006 demonstrating the utility of using mixtures of Polya trees just like mixtures of the Dirichlet process, 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 analysis (Hanson 2007), and multivariate mixtures of Polya trees are used for modelling ROC data (Hanson et al. 2008). For an introduction and some applications, the reader is referred to a paper by Christensen et al. (2008).

To soften the effect of partition points that play part in inferential outcomes, Paddock et al. (2003) devised a randomized Polya tree, as mentioned earlier. The paper also contains procedure to simulate observations from predictive distribution and illustrative examples.

5.3 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 encountered problems. In some cases the estimators of a distribution function (or a survival function) obtained are not proper distribution functions (or of 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 straightforward manner. See, for example, Ferguson (1973), Dalal and Phadia (1983), Phadia and Susarla (1983), among others, who assign a Dirichlet process prior for an unknown bivariate distribution function defined on $R^2 = R \times R$ or $R_+^2 = R^+ \times R^+$ 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 $\mathcal{R}^* = \mathcal{R}^+ \times \{0, 1\}$ and $\mathcal{B}^* = \mathcal{B} \times \{\phi, \{0\}, \{1\}, \{0, 1\}\}$, $\mathcal{B}$ is a Borel field on $\mathcal{R}^+$ 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, Pruitt (1992) 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), Paddock et al. (2003) and Phadia (2007). Recall that in the univariate case the Bayes 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 labelled as "essential 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 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 estimation of 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 closed but otherwise unknown. Their closeness is modelled by defining a parameter $\rho = \text{corr}\left(F_1(A), F_2(A)\right) \geq 0$ 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 distributions and show how to find their posterior distributions. Their prior utilizes the Dirichlet-multinomial point process introduced by Lo (1986) (Sect. 2.5). 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 Chaps. 6 and 7. This extension is presented here. An anonymous reviewer has brought to my attention a paper of Paddock et al. (2003) in which a similar construction is suggested for Euclidean space R^k. Here more details are presented in the case of bivariate process.

5.3.1 Bivariate Tailfree Process

Recall the definition of a tailfree process presented in Sect. 5.1. 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 $\exists$ 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 $B_1 \supset B_2, \ldots, \supset B_m$, 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. 5.1 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, $B_{11}, \ldots, B_{14}, \ldots, B_{41}, \ldots, B_{44}$, each of size $1/4$ and the process is continued. Now let $\pi_0 = \{B_0\}, \pi_1 = \{B_1, B_2, B_3, B_4\}, \pi_2 = \{B_{11}, \ldots, B_{14}, \ldots, B_{41}, \ldots, B_{44}\}, \ldots,$ $\pi_m = \{B_{c_1 c_2, \ldots c_m},$ where $c_i = 1, 2, 3, 4$ for $i = 1, 2, \ldots, m\}, m = 1, 2, \ldots .\ \pi_m$ will have 4^{m-1} 4-tuple subsquares of size 2^{-m}. Note that $B_{c_1} \supset B_{c_1 c_2} \supset B_{c_1 c_2 c_3} \supset \ldots .$ Thus $\Pi = \{\pi_m;\ m = 0, 1, \ldots\}$ forms a sequence of nested partitions.

The subsquares $B_{c_1 c_2, \ldots c_m}$ 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 $(.e_1 e_2 \ldots .e_m, .e_1^{'} e_2^{'} \ldots .e_m^{'})$, where e_i and $e_i^{'}$ take values 0 or 1. For example, the bottom left end corner of B_{32} is $\left(\frac{1}{2}, \frac{3}{4}\right) = (0.10, 0.11)$. To further identify the four subsquares $B_{c_1 c_2, \ldots c_m 1}, B_{c_1 c_2, \ldots c_m 2}, B_{c_1 c_2, \ldots c_m 3}$, and $B_{c_1 c_2, \ldots c_m 4}$ of $B_{c_1 c_2, \ldots c_m}$ with their bottom left end corners $(.e_1 e_2 \ldots .e_m \cdot, .e_1^{'} e_2^{'} \ldots .e_m^{'} \cdot)$, 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 $(0.010, 0.100)$.

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 $\underline{c}_k = c_1 c_2, \ldots c_k$. Let $E^* = \cup_k E_k$ 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 $\cup_k E_k$ generates $\sigma\left((0,1]^2\right)$.

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 $B = B_{\underline{c}_2} \in \pi_2,\ c_i = 1, 2, 3, 4, i = 1, 2\}, \ldots, \{V_{m,B} = Z_{\underline{c}_m} = P(B_{\underline{c}_m} | B_{\underline{c}_{m-1}})$ for $B = B_{\underline{c}_m} \in \pi_m, c_i = 1, 2, 3, 4$ for $i = 1, 2, \ldots, m\}, \ldots \ldots .$ We take these families to be independent between levels, $m = 1, 2, \ldots .$ This way for an arbitrary set $B \in \pi_m$, $P(B) = P(B_{c_1 c_2, \ldots c_m})$, is the product of all the variables associated with the path in the tree from $[0,1]^2$ to $B_{c_1 c_2, \ldots c_m}$ so that $P(B_{c_1 c_2, \ldots c_m}) = \prod_{i=1}^{m} Z_{c_1 c_2, \ldots, c_i}, c_i \in \{1, 2, 3, 4\}, i = 1, 2, \ldots, m.$ For example, if $B_{2143} \in \pi_4, P(B_{2143}) = Z_2 Z_{21} Z_{214} Z_{2143}$. The random probability P is defined through the joint distribution of $P(B_{c_1 c_2, \ldots c_m})$ by assigning suitable distributions to Z's. Since the sets $B_{\underline{c}_m}$ are decreasing to an empty set ϕ, we should therefore have $P(B_{\underline{c}_m \underbrace{ii \ldots .i}})$ going to 0 for $i \in \{1, 2, 3, 4\}$. Thus the choice of distributions of Z's should be such that $P(B_{\underline{c}_m \underbrace{ii \ldots .i}} \ldots) = P(B_{\underline{c}_m}) \prod Z_{\underline{c}_m \underbrace{ii \ldots .i}} \ldots \xrightarrow{\text{a.s.}} 0.$

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 $((0,1]^2, \sigma\left((0,1]^2\right))$. The distribution of P will be tailfree with respect to the sequence of partitions, Π. Now Theorem 5.2 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 $\left(Z_{\underline{c}_{m-1,1}}, \ldots, Z_{\underline{c}_{m-14}}\right)$ at level m to be mutually independent each having a Dirichlet distribution with parameters $\left(\alpha_{\underline{c}_{m-1,1}}, \ldots, \alpha_{\underline{c}_{m-14}}\right)$ for some nonnegative real numbers α's. 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, $Z_{2143} \sim \text{Be}(\alpha_{2143}, \alpha_{2141} + \alpha_{2142} + \alpha_{2144})$. To ensure that $P(B_{\underline{c}_m}) \prod Z_{\underline{c}_m \underbrace{ii\ldots i}_{j}} \cdots \xrightarrow{\text{a.s.}} 0$, we place the following condition on α's.

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

$$\gamma_{\underline{c}_m \underbrace{ii\ldots i}_{j}} = \sum_{l=1}^{4} \alpha_{\underline{c}_m \underbrace{ii\ldots i}_{j} l}. \tag{5.3.1}$$

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.

$$(Z_{\underline{c}_m 1}, Z_{\underline{c}_m 2}, Z_{\underline{c}_m 3}, Z_{\underline{c}_m 4}) \mid X_1, X_2, \ldots X_n \sim D(\alpha_{\underline{c}_m 1} + N_{\underline{c}_m 1}, \ldots, \alpha_{\underline{c}_m 4} + N_{\underline{c}_m 4}), \tag{5.3.2}$$

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). The application of using Polya tree priors in Bayesian estimation is discussed in the next two chapters.

Chapter 6
Inference Based on Complete Data

6.1 Introduction

In the statistical problem of nonparametric Bayesian analysis we have a random probability P belonging to Π and having a particular prior distribution. Given $P = P$, 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 part use the Dirichlet process prior or its variants—Dirichlet Invariant process and mixtures of Dirichlet processes. In Chap. 7, we use other priors such as the neutral to the right processes which are more suited while dealing with censored 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 estimate of ϕ for the no-sample problem whenever possible 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 R. In the case of two-sample problem, we will have a second sample $Y_1, \ldots, Y_n$ from another distribution function, say, G defined on R. Both samples will be assumed to be independent. The loss functions used are a weighted integral squared error loss, denoted by L_1 for the distribution function and a squared error loss L_2 for its functionals, where

$$L_1(F, \widehat{F}) = \int (F(t) - \hat{F}(t))^2 dW(t); \quad L_2(\varphi, \hat{\varphi}) = (\varphi - \hat{\varphi})^2, \tag{6.1.1}$$

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

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1_6

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 $\bar{\alpha}(\cdot) = \alpha(\cdot)/\alpha(R)$, $M = \alpha(R)$ and $p_n = \alpha(R)/(\alpha(R) + n)$. In some applications, we use $\mathfrak{X}$ 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. Some 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, and beta-Stacy, although closed form of the results may not be guaranteed.

6.2 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.

6.2.1 Estimation of a CDF

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, 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, $F(t) \sim \text{Be}(\alpha(-\infty, t], \alpha(t, \infty))$. 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 2.6 of Sect. 2.1, we have $F|\mathbf{X} \sim \mathcal{D}(\alpha + \sum_{i=1}^n \delta_{X_i})$. 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)

$$\widehat{F}_\alpha(t) = \widehat{F}(t \mid X_1, \ldots, X_n) = \frac{\alpha(-\infty, t] + \sum_{i=1}^n \delta_{X_i}(-\infty, t]}{\alpha(R) + n}$$
$$= p_n \cdot F_0(t) + (1 - p_n) \cdot \widehat{F}_n(t), \text{ say,} \tag{6.2.1}$$

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 Baye's 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 (1983).

Remark 6.1 $M = \alpha(R)$ 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 Baye's 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 α_0 and $\alpha_k, k \geq 1$ on R, they show that if, along with the sequence $\alpha_k(R) \to 0$, we have $\text{Sup}_A |\overline{\alpha}_k(A) - \overline{\alpha}_0(A)| \to 0$ as $k \to \infty, A \in \mathcal{B}$, then $\mathcal{D}(\alpha_k) \to \delta_{Y_0}$, 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 its discreteness.

6.2.2 *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. 2.2), that is, $F \sim \mathcal{DGI}(\alpha)$, $\mathcal{G} = \{e, g\}$ with $e(x) = x, g(x) = 2\eta - x$. Then the Bayes estimate of F under the loss function L_1 is (Dalal 1979a)

$$\widehat{F}_{\alpha\eta}(t \mid X_1, \ldots, X_n) = \frac{\alpha(-\infty, t] + \frac{1}{2}\sum_{i=1}^n (\delta_{X_i}(-\infty, t] + \delta_{2\eta - X_i}(-\infty, t])}{\alpha(R) + n}$$
$$= p_n \cdot F_0(t) + (1 - p_n) \cdot \widehat{F}_{sn}(t), \tag{6.2.2}$$

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

6.2.3 Estimation of a CDF with MDP Prior

Let $G(\theta)$ stand for a random distribution function selected from a mixture of Dirichlet processes (Sect. 2.3) 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 $\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G$, and given θ_i let $X_{i1}, \ldots, X_{im_i}$ be a sample of size m_i from $F_{\theta_i}(x), i = 1, \ldots, n$. This is the so-called mixture model presented in Sect. 2.4.

The Bayes estimate of G under the L_1 loss function is given by $\widehat{G} = \mathcal{E}(G|\theta_1, \ldots, \theta_n)$ if θ_i's are observed directly, and $\widehat{G} = \mathcal{E}(G|X_{i1}, \ldots, X_{im_i})$ if X_{ij}'s are observed. One can use the formula given under property 3 or 4 of Sect. 2.3 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 reported in Kuo (1986b) Dey et al. (1998), Ibrahim, Chen, and Sinha (2001), and simulation procedures discussed in Sect. 2.4 have mitigated the problem to some extent.

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 $\mathcal{D}(MH_\theta)$ with specified $M > 0$. This leads to the mixture of Dirichlet processes priors, $F \in \int \mathcal{D}_{MH_\theta} \upsilon(d\theta)$. 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.

6.2.4 Empirical Bayes Estimation of a CDF

In the Sect. 6.2.1 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 that 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 6.2 In many hierarchical modeling, the parameters at intermediate stages are assumed to have certain distributions with some hyper parameters. This 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 the unknown parameters.

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 $p_j = \alpha(R)/(\alpha(R) + m_j)$, $j = 1, \ldots, n+1$. 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

$$L(F_{n+1}, \widetilde{F}_{n+1}) = \int \left(F_{n+1}(t) - \widetilde{F}_{n+1}(t)\right)^2 dW(t), \tag{6.2.3}$$

based on $\mathbf{X}_1, \ldots, \mathbf{X}_{n+1}$ is from Sect. 6.2.1 (suppressing the dependence on α),

$$\widetilde{F}_{n+1}(t) = p_{n+1}F_0 + (1 - p_{n+1})\widehat{F}_{n+1}(t). \tag{6.2.4}$$

The Bayes risk of $\widetilde{F}_{n+1}(t)$ with respect to $D(\alpha)$, denoted by $r(\alpha) = \mathcal{E}_{\mathcal{D}(\alpha)}\mathcal{E}_{F_{n+1}}[L(F_{n+1}, \widetilde{F}_{n+1})]$, is (Korwar and Hollander 1976)

$$r(\alpha) = r(\widetilde{F}_{n+1}, \alpha) = \frac{p_{n+1}}{\alpha\,(R) + 1} \int F_0(t)(1 - F_0(t))dW(t) = \frac{p_{n+1}}{\alpha\,(R) + 1}\sigma^2, \tag{6.2.5}$$

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 (6.2.4) and proposed the following empirical Bayes estimator of F_{n+1}:

$$\overline{F}_{n+1}(t) = p_{n+1}\frac{1}{n}\sum_{j=1}^{n} \widehat{F}_j(t) + (1 - p_{n+1})\widehat{F}_{n+1}(t). \tag{6.2.6}$$

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

$$r(\overline{F}_{n+1}, \alpha) = \mathcal{E}[L(F_{n+1}, \overline{F}_{n+1})] = r(\alpha)\left(1 + \frac{p_{n+1}}{n^2}\sum_{j=1}^{n} \frac{1}{1 - p_j}\right), \tag{6.2.7}$$

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 $0(n^{-1})$. 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 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, $0(n^{-1})$.

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

$$\widetilde{F}^*_{n+1}(t) = p_{n+1}\widehat{F}_0(t) + (1 - p_{n+1})\widehat{F}_{n+1}(t), \tag{6.2.8}$$

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

$$r(\widetilde{F}^*_{n+1},\alpha) = r(\alpha)\left[1+\frac{p_{n+1}}{\sum_{j=1}^{n+1}(1-p_j)}\right], \tag{6.2.9}$$

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 $r(\widetilde{F}^*_{n+1},\alpha) - r(\alpha) \leq r(\overline{F}_{n+1},\alpha) - r(\alpha)$ 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 $w_j^* = p_{n+1}(1-p_j)/\left(\sum_{j=1}^{n+1}(1-p_j)\right), j = 1,\ldots,n$, 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 $w_j = p_m/n$ $(j = 1,\ldots,n)$ and $w_{n+1} = 1-p_m$, with $m_1 = \cdots = m_n = m$ 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 $j = 1,\ldots,n+1$ 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 $w_1 = \cdots = w_n = 0$.

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

$$m/(1-F_n) \to \alpha(R) \text{ in probability as } n \to \infty. \tag{6.2.10}$$

Ghosh et al. (1989) provide an improvement over his estimator by including $\mathbf{X}_{n+1}$ as well in the F_n statistic. Let $\overline{X}_j = \sum_{i=1}^{m_j} X_{ji}/m_j$ be the mean of the sample values at jth 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

$$\text{MSW} = \sum_{j=1}^{n+1}\sum_{i=1}^{m_j}(X_{ji}-\overline{X}_j)^2 \Big/ \sum_{j=1}^{n+1}(m_j-1) \text{ and } \text{MSB} = \sum_{j=1}^{n+1}(\overline{X}_j-\overline{X})^2/n, \tag{6.2.11}$$

the usual within and between mean squares, respectively. Simple evaluations involving the Dirichlet process yield

$$E(\text{MSW}) = (\alpha(R)/(\alpha(R)+1))\,\sigma^2 \tag{6.2.12}$$

$$E(\text{MSB}) = (\alpha(R)/(\alpha(R)+1))\,\sigma^2 + \xi.\sigma^2/(n(\alpha(R)+1)), \tag{6.2.13}$$

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 $\alpha(R)$.

$$\hat{\alpha}^{-1}(R) = \max\{0,\ (\text{MSB}/\text{MSW} - 1)\, n/\xi\}, \tag{6.2.14}$$

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 $m_1 = \cdots = m_{n+1}$]. Substituting this estimator of $\alpha(R)$ in $p_j = (1 + m_j \alpha^{-1}(R))^{-1}$ they revise the estimate for F_0 as

$$\widetilde{F}_0(t) = \begin{cases} \sum_{j=1}^{n+1} (1 - \hat{p}_j) \widehat{F}_j(t) / \sum_{j=1}^{n+1} (1 - \hat{p}_j), & \text{if } \hat{\alpha}^{-1}(R) \neq 0 \\ \sum_{j=1}^{n+1} \widehat{F}_j(t)/(n+1), & \text{if } \hat{\alpha}^{-1}(R) = 0. \end{cases} \tag{6.2.15}$$

Finally for the case $\alpha(R)$ unknown, Ghosh et al. (1988) utilizing these estimators proposed $\widehat{\widetilde{F}}_{n+1}$ as an improved empirical Bayes estimator of F_{n+1}, where

$$\widehat{\widetilde{F}}_{n+1}(t) = \hat{p}_{n+1} \widetilde{F}_0(t) + (1 - \hat{p}_{n+1}) \widehat{F}_{n+1}(t), \tag{6.2.16}$$

and proved the asymptotic optimality of this estimator.

6.2.5 *Sequential Estimation of a CDF*

Ferguson (1982) derives the sequential estimator of F with respect to the loss function L_1 and prior $\mathcal{D}(\alpha)$, with $F_0 = \overline{\alpha}$ taken 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 $\mathcal{D}((M+n)\hat{F}_\alpha)$, where $\widehat{F}_\alpha$, as before, is the Bayes estimator of F under L_1, and the minimum Bayes risk is obtained as

$$\int \text{Var}(F(x)|\mathbf{X}) dW(x) = (1/(M+n+1)) \int_\alpha \widehat{F}_\alpha(x)(1 - \widehat{F}_\alpha(x)) dW(x). \tag{6.2.17}$$

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 provides 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 calls for stopping after the first n observations for which

$$\int_\alpha \widehat{F}_\alpha (1 - \widehat{F}_\alpha) dW \leq c(M+n+1)^2. \tag{6.2.18}$$

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 and 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 $\text{Var}(\mu|X_1, \ldots X_n) = \sigma_n^2/(M+n+1)$, where σ_n^2 is the variance of the distribution $\widehat{F}_n$, which can be expressed as

$$\sigma_n^2 = \int (x-\mu_n)^2 d\widehat{F}_n(x) = \left(M\sigma_0^2 + ns_n^2 + \frac{Mn}{M+n}(\overline{x}_n - \mu_0)^2 \right) / (M+n), \tag{6.2.19}$$

with σ_0^2 as the variance of F_0 and $ns_n^2 = \sum_1^n (x_i - \overline{x}_n)^2$. Then the 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.

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 priors. Clayton and Berry (1985) treat one-armed bandit problem and Clayton (1985) a sequential testing problem for the mean of a population.

6.2.6 *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 used was first to find an equalizer rule given by

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

and 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).

6.3 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) and others.

6.3.1 Tolerance Region

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

$$L(p,a) = pP((-\infty, a]) + qI_{(a,\infty)}(t_q), \tag{6.3.1}$$

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

$$\mathcal{E}(L(p,a)) = pu + q\int_0^q \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)}\, z^{uM-1}(1-z)^{(1-u)M-1}dz, \tag{6.3.2}$$

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 $X_1, \ldots, X_n$ of size n, the Bayes rule therefore is given by

$$\hat{a}_n(\mathbf{X}) = f(p,q,\alpha(R)+n)\text{th quantile of } \widehat{F}_n. \tag{6.3.3}$$

6.3.2 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) and Neath and Bodden (1997) treat this problem.

Recursive Method

Breth defines Bayesian confidence bands as follows.

Definition 6.3 (Breth) Suppose $F \sim \mathcal{D}(\alpha)$. Then if $\mathcal{P}\{F_L(t) \leq F(t) \leq F_U(t)$ for all $t\} = \nu_1\,(\nu_2)$ 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 $\nu_1\,(\nu_2)$. $(F_L(t), F_U(t))$are defined to be a pair of Bayesian confidence bands for the random distribution function F with prior (posterior) probability $\nu_1\,(\nu_2)$.

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 $0 = u_0 \leq u_1 \leq \cdots \leq u_m < 1$, $0 < v_1 \leq v_2 \leq \cdots \leq v_{m+1} = 1$. Further, let $I(x) = 1$ if $x \geq 0$ and 0 otherwise, and $J(x) = 1$ if $x > 0$ and 0 otherwise. For $-\infty = t_0 < t_1 < t_2 < \cdots < t_m < t_{m+1} = \infty$, 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 $\mathcal{P}\{F_L(t) \leq F(t) \leq F_U(t)$ for all $t\} = \mathcal{P}\{u_j \leq F(t_j) \leq v_j$ 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 $a_j = \alpha\left(t_j\right) - \alpha\left(t_{j-1}\right), j = 1, \ldots, m$. 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 $\mathcal{P}\{u_j \leq F(t_j) \leq v_j$ for all $j\}$ for fixed numbers $\{u_j\}$, $\{v_j\}$, and $\{t_j\}$ when F is a Dirichlet process, and gives details on calculations that are needed in practical applications. 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 F. 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 (1971) 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, $\theta \sim G$, $P|\theta \sim \mathcal{D}(\alpha_\theta)$. 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

$$\frac{1}{N}\sum_{i=1}^{N} I\{F_L(t) \le F_i(t) \le F_U(t), \quad t \in R\} \ge 1-\gamma. \tag{6.3.4}$$

In choosing the "best" bounds, the following two criteria are used:

$$\text{(i) } \min\{\max_t [F_U(t) - F_L(t)]\} \text{ and (ii) } \min\left\{\int [F_U(t) - F_L(t)]dW(t)\right\}. \tag{6.3.5}$$

The minimum is taken over all functions F_L and F_U such that (6.3.4) is satisfied. For the process of choosing F_L and F_U, they give two algorithms and discuss their implementation, and also provide a numerical example to illustrate the procedure.

Bayesian Bootstrap Method

Hjort (1985), on the other hand, uses a Bayesian bootstrap method to construct confidence intervals for a function $\theta(F)$ of an unknown F. Let $F \sim \mathcal{D}(\alpha)$ and $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$. Then $F|\mathbf{X} \sim \mathcal{D}(\alpha + \sum_1^n \delta_{x_i})$ which can be written as $\mathcal{D}(MF_0 + n\widehat{F}_n)$. 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 $G^{-1}(p) = \inf\{t : G(t) \ge p\}$. Now a large set of values of $\theta(F)$ can be generated via Monte Carlo and then G can be approximated, permitting the calculation of $G^{-1}(p)$.

6.4 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, and variance are derived using the Dirichlet process priors.

6.4.1 Estimation of the Mean

The Bayesian estimation of the mean $\mu = \int x dP(x)$ with respect to the Dirichlet process prior and under the squared error loss L_2 was considered in Ferguson (1973). 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 6 of Sect. 2.1.2, is $\hat{\mu} = \mathcal{E}_{\mathcal{D}(\alpha)} \int x dP(x) = \int x d\alpha(x)/\alpha(R) = \mu_0$. 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

$$\hat{\mu}_{\alpha n}(\mathbf{X}) = (\alpha(R)+n)^{-1} \int x d(\alpha(x) + \sum_{i=1}^{n} \delta_{X_i}(x)) = p_n \mu_0 + (1-p_n)\overline{X}_n, \quad (6.4.1)$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. The Bayes estimator thus is like $\hat{F}_\alpha$, a convex combination of the prior guess at μ, namely μ_0, and the sample mean. As $\alpha(R) \to 0$, $\hat{\mu}_{\alpha n} \to \overline{X}_n$, and as $\alpha(R) \to \infty$, $\hat{\mu}_n \to \mu_0$. The Bayes risk of $\hat{\mu}_n$ 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. 2.1.2. More generally, let Z be a measurable real valued function defined on $(R, \mathcal{B})$ and $\theta = \int Z dP$. If $P \sim \mathcal{D}(\alpha)$ and $\theta_0 = \int Z d\alpha/\alpha(R) < \infty$, then the Bayes estimator of θ under the loss L_2 is given by

$$\hat{\theta}_{\alpha n}(\mathbf{X}) = p_n \theta_0 + (1-p_n)\frac{1}{n}\sum_{i=1}^{n} Z(X_i). \quad (6.4.2)$$

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.

6.4.1.1 Empirical Bayes Estimation of the Mean

This can be dealt with in the same manner as the distribution function in Sect. 6.2.4 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

$$\hat{\mu}_\alpha = p_{n+1}\mu_0 + (1-p_{n+1}) \sum_{i=1}^{m_{n+1}} X_{n+1,i}/m_{n+1}. \quad (6.4.3)$$

The Bayes risk of $\hat{\mu}_\alpha$ 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 $\hat{\mu}_n$ has the Bayes risk as $r(\hat{\mu}_n, \alpha) = (1 + \alpha(R)/\sum_{i=1}^{n+1} m_i) r(\alpha)$. Ghosh et al. (1988) estimate μ_0 from the past as well as current sample data as $\hat{\mu}_0 = \sum_{j=1}^{n+1}(1-p_j)\overline{X}_j / \sum_{j=1}^{n+1}(1-p_j)$ and plugs in $\hat{\mu}_\alpha$ of

(6.4.3). The resulting estimator is denoted as $\hat{\mu}_{n+1}$ and its Bayes risk is

$$r(\hat{\mu}_{n+1}, \alpha) = r(\alpha) + p_{n+1}^2 \sigma^2 / \sum_{j=1}^{n+1} (1 - p_j). \tag{6.4.4}$$

They have shown that $\hat{\mu}_{n+1}$ 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. 6.2.4.

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.

6.4.2 *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 defined by

$$\text{Var } P = \int x^2 dP(x) - \left(\int x dP(x) \right)^2, \tag{6.4.5}$$

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 Var P for the no-sample problem is the posterior mean

$$\begin{aligned} \mathcal{E}\,\text{Var } P &= \mathcal{E} \int x^2 dP(x) - \mathcal{E} \left(\int x dP(x) \right)^2 = (\sigma_0^2 + \mu_0^2) - \left(\frac{\sigma_0^2}{\alpha(R) + 1} + \mu_0^2 \right) \\ &= \frac{\alpha(R)}{\alpha(R) + 1} \sigma_0^2, \end{aligned} \tag{6.4.6}$$

where μ_0 is as defined above in Sect. 6.4.1 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 Var P as

$$\begin{aligned}\hat{\sigma}_n^2(\mathbf{X}) &= \frac{\alpha(R)+n}{\alpha(R)+n+1}\mathrm{Var}(\hat{F}_\alpha) \\ &= \frac{\alpha(R)+n}{\alpha(R)+n+1}(p_n\sigma_0^2 + (1-p_n)\,s_n^2 + p_n(1-p_n)(\mu_0-\overline{X}_n)^2) \\ &= \frac{\alpha(R)+n}{\alpha(R)+n+1}\left(p_n\sigma_0^2 + (1-p_n)\left(p_n\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2 + (1-p_n)\,s_n^2\right)\right),\end{aligned} \tag{6.4.7}$$

where $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. The last equality expresses $\hat{\sigma}_n^2$ 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, $\hat{\sigma}_n^2$ 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 $(\mathrm{Var}P - \hat{\sigma}^2)^2/(\mathrm{Var}P)^2$.

6.4.3 *Estimation of the Median*

Next consider the problem of estimation of the median θ defined as $\theta = med\ P$. Ferguson (1973) derived the Bayes estimator under the absolute error loss, $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$. 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,

$$\text{med (dist. med P)} = \text{med } \mathcal{E}P. \tag{6.4.8}$$

Thus any number t satisfying

$$\frac{\alpha((-\infty,t))}{\alpha(R)} \le \frac{1}{2} \le \frac{\alpha((-\infty,t])}{\alpha(R)} \tag{6.4.9}$$

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 $\hat{\theta} = median\ of\ F_0$ and for the sample of size n, it is

$$\hat{\theta}_{\alpha n} = \text{median of } \widehat{F}_\alpha, \tag{6.4.10}$$

where $\widehat{F}_\alpha$ is the Bayes estimate of F derived in Sect. 6.2.1.

Doss (1985a,b) 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 the intervals $(-\infty, 0)$ 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, $F \sim \mathcal{D}^*(\alpha)$ 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

$$d\upsilon(\theta|\mathbf{X}) = \kappa(\mathbf{X})\,\Pi^*[f_0(X_i-\theta)]\Psi(\mathbf{X},\theta)\,d\upsilon(\theta), \tag{6.4.11}$$

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, Π^* 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. 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. Doss 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. 4.2.

6.4.4 *Estimation of the qth Quantile*

Ferguson (1973) extends the estimation of the median to the qth quantile of P, denoted by t_q: $P((-\infty, t_q)) \le q \le P((-\infty, t_q])$, for $0 < q < 1$. The qth quantile of $P \sim \mathcal{D}(\alpha)$ is unique with probability 1, so that t_q is a well-defined random variable. He considers the following loss function:

$$\begin{aligned} L(t_q, \hat{t}_q) &= p(t_q - \hat{t}_q) \qquad \text{if } t_q \ge \hat{t}_q \\ &= (1-p)(t_q - \hat{t}_q) \quad \text{if } t_q < \hat{t}_q, \end{aligned} \tag{6.4.12}$$

for some $p, 0 < p < 1$. For this loss, any pth quantile of the distribution of t_q is a Bayes estimate of t_q. The distribution of t_q is

$$\mathcal{P}\{t_q \leq t\} = \mathcal{P}\{F(t) > q\} = \int_q^1 \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)} z^{uM-1}(1-z)^{(1-u)M-1} dz, \tag{6.4.13}$$

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 pth 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 uth quantile of F_0,

$$\hat{t}_q = u(p, q, \alpha(R))\text{th quantile of } F_0, \tag{6.4.14}$$

and for the sample of size n, it is

$$\hat{t}_q(\mathbf{X}) = u(p, q, (\alpha(R) + n))\text{th quantile of } \widehat{F}_\alpha. \tag{6.4.15}$$

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,b) extends his own results of estimating the median to the estimation of quantiles as well, and discusses their properties.

6.4.5 *Estimation of a Location Parameter*

Considered the following model for sample observations. Let $Y = \eta + \varepsilon$, 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 an 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 $G \sim \mathcal{D}(MF_0)$, 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 (1979b) 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, he assumes F_η to be distributed according to the Dirichlet Invariant process, $\mathcal{DGI}(\alpha)$ and obtains a Bayes estimate $\hat{\eta}(\mathbf{y}) = \mathcal{E}_{\eta|\mathbf{y}}(\eta)$ 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, $\hat{\eta}(\mathbf{y})$ 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 $F_\eta \sim \mathcal{DGI}(\alpha)$. Then $\mathcal{E}(F_\eta) = F_0$ and the marginal conditional distribution of Y_1 given η is F_0. Since υ is a 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 $F_\eta|y_1, \eta \sim \mathcal{DGI}\left(\alpha + \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{2\eta - y_1}\right)$ 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. 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 on 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 involving simulation should make it simpler 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.

6.4.6 *Estimation of P(Z > X + Y)*

Zalkikar, Tiwari, and Jammalamadaka (1986) considered the problem of estimation of the parameter

$$\Delta(F) = P(Z > X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(x + y) dF(x) dF(y), \tag{6.4.16}$$

where it is assumed that the random variables X, Y, and Z are independent from the same 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 $F \sim \mathcal{D}(\alpha)$ 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:

$$\widehat{\Delta}(F) = \frac{M+m}{(M+m)^{(3)}}\left[2(M+m+1) + \widehat{F}_\alpha(0-) + (M+m)\int_{-\infty}^{\infty} S_\alpha(2y)d\widehat{F}_\alpha(y) + (M+m)^2\Delta(\widehat{F}_\alpha)\right], \tag{6.4.17}$$

where $a^{(k)} = a(a+1)\ldots(a+k-1)$ and $S_\alpha = 1 - \widehat{F}_\alpha$.

When $M \to 0$, the estimator reduces to an estimator which is asymptotically equivalent to the U-statistic,

$$U_m = \frac{1}{m(m-1)(m-2)}\sum I[X_i > X_j + X_k], \tag{6.4.18}$$

where the summation is taken over all $m(m-1)(m-2)$ distinct triplets (i, j, k), $1 \le i, j, k \le m$.

By using the earlier mentioned technique (Sect. 6.2.4), 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 $0(n^{-1})$.

6.5 Other Applications

There are many other applications that have not been covered here. For example, Lo (1988) 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 subsequent 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 (1985).

We do, however, present a few interesting applications.

6.5.1 Bayes Empirical Bayes Estimation

In a general empirical Bayes setting (or mixture models), 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

$$G|\mathbf{X} \sim \int \mathcal{D}\left(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}\right) dH(\boldsymbol{\theta}|\mathbf{X}). \tag{6.5.1}$$

If we have unconditional marginal distribution of $\boldsymbol{\theta}$, then $dH(\boldsymbol{\theta}|\mathbf{X})$ can be expressed as

$$dH(\boldsymbol{\theta}|\mathbf{X}) = \prod_{j=1}^{n} f_j\left(x_j|\theta_j\right) dH\left(\boldsymbol{\theta}\right) \Bigg/ \left(\int \prod_{j=1}^{n} f_j\left(x_j|\theta_j\right) dH\left(\boldsymbol{\theta}\right)\right). \tag{6.5.2}$$

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,b) was able to express the Bayes estimator of θ_i under the loss $\sum_{i=1}^{n}(\theta_i - \hat{\theta})^2$ in a concise form as a ratio of two n-dimensional integrals as follows:

$$\hat{\theta}_i = \mathcal{E}(\theta_i|\mathbf{X}) = \frac{\int \ldots \int_{R^n} (\theta_i \prod_{i=1}^{n} f_i(x_i|\theta_i)) \prod_{i=1}^{n} (\alpha + \sum_{j=1}^{i-1} \delta_{\theta_j})(d\theta_i)}{\int \ldots \int_{R^n} (\prod_{i=1}^{n} f_i(x_i|\theta_i)) \prod_{i=1}^{n} (\alpha + \sum_{j=1}^{i-1} \delta_{\theta_j})(d\theta_i)} \tag{6.5.3}$$

for all $i = 1, 2, \ldots, n$. Still it is hard to evaluate these integrals. She overcomes this problem by decomposing each of the multidimensional 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.

6.5.2 *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 subjects 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 used 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. 7.2.8) 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 [see Eq. 7.2.21]. It is 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, $\hat{F}(t_1) = (\alpha(0, t_1] + r_1) / (\alpha(0, \infty) + n_1)$. 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:

$$\hat{F}(t_1) = \sum_{i=0}^{r_2} \sum_{j=0}^{n_1-r_1} a_{ij} \frac{\beta_1 + r_1 + i}{M + n_1 + n_2}, \tag{6.5.4}$$

$$\hat{F}(t_2) = \sum_{i=0}^{r_2} \sum_{j=0}^{n_1-r_1} b_{ij} \frac{\beta_1 + \beta_2 + n_1 + r_2 - j}{M + n_1 + n_2}, \tag{6.5.5}$$

and for other values of t, $\hat{F}(t)$ is obtained by the linear interpolation. Here

$$a_{ij} = b_{ij} / \sum_{i=0}^{r_2} \sum_{j=0}^{n_1-r_1} b_{ij} \text{ and} \tag{6.5.6}$$

$$b_{ij} = \binom{n_1 - r_1}{j} \binom{r_2}{i} \times \frac{\Gamma(\beta_1 + r_1 + i)\,\Gamma(\beta_2 + n_1 - r_1 + r_2 - i - j)\,\Gamma(\beta_3 + n_2 - r_2 + j)}{\Gamma(\beta_1)\,\Gamma(\beta_2)\,\Gamma(\beta_3)}, \tag{6.5.7}$$

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 $\int (F - \hat{F})^2 dW$ 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 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 ith column replaced by the column $\left(\text{cov}\left(r_1, F(t_j)\right), \ldots, \text{cov}\left(r_L, F(t_j)\right)\right)^T$. 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 $F \sim \mathcal{D}(\alpha)$, the Bayes rule in this class at each dose level $t_j, j = 1, 2 \ldots, L$ is given by

$$\hat{F}(t_j) = F_0(t_j) + \sum_{i=1}^{L} n_i \hat{\lambda}_i(j)[r_i/n_i - F_0(t_i)], \tag{6.5.8}$$

and at $t, t_j < t < t_{j+1}$

$$\hat{F}(t) = \frac{F_0(t_{j+1}) - F_0(t)}{F_0(t_{j+1}) - F_0(t_j)} \hat{F}(t_j) + \frac{F_0(t) - F_0(t_j)}{F_0(t_{j+1}) - F_0(t_j)} \hat{F}(t_{j+1}), \tag{6.5.9}$$

where $\hat{\lambda}_i(j) = |D(i, t_j)|/|\text{cov}(r)|$.

Kuo also shows that $\hat{F}(t_j)$ is an asymptotically unbiased and consistent estimator of $F(t_j)$. As $M \to 0$, $\hat{F}(t_j) \to r_j/n_j$ and as $M \to \infty$, $\hat{F}(t_j) \to F_0(t_j)$. 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 (Barlow 1972, pp. 13–18) for obtaining the desired result.

In the case of M and F_0 unknown, empirical Bayes method of Sect. 6.2.4 may be used.

6.5.3 *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 arises in regression problems can also be handled in the same way. Let G be a distribution function on $[0, 1]$ and assume that $G \sim \mathcal{D}(\alpha)$. At chosen points $0 = t_0 < t_1 < \cdots < t_k \leq 1$, 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 the bioassay problem, the observations available at each value of l were from a binomial distribution, where as in the regression problem, they arise from some known distribution.

Consider the general linear model $Z = \mathbf{X}_{\cdot} + \epsilon$, 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 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.

As stated in Sect. 5.2, 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, whereas Hanson and Johnson (2002) recommend modeling the error distribution as a mixture of Polya trees. These approaches were mentioned in Sect. 5.2.

6.5.4 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 $\int x f(x)dx$. 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) approaches this problem by using a kernel representation of the density function, and assigning a Dirichlet prior to its mixing distribution 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 $G \sim \mathcal{D}(\alpha)$, 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

$$\mathcal{P}(G \in B|\mathbf{X}) = \frac{\int_B \prod_{i=1}^n \int_R K(x_i, u_i)G(du_i)\mathcal{D}_\alpha(dG)}{\int_{\mathcal{F}} \prod_{i=1}^n \int_R K(x_i, u_i)G(du_i)\mathcal{D}_\alpha(dG)}, \tag{6.5.10}$$

for all $B \in \mathcal{F}$. By repeated application of his lemma (interchanging the order of integration),

$$\int_{\mathcal{F}} \int_{R} h(u, G) G(du) \mathcal{D}_{\alpha}(dG) = \int_{R} \int_{\mathcal{F}} h(u, G) \mathcal{D}_{\alpha+\delta_u}(dG) \alpha(du)/\alpha(R), \tag{6.5.11}$$

he shows that

$$\mathcal{P}(G \in B|\mathbf{X}) = \frac{\int_{R^n} \mathcal{D}_{\alpha+\sum \delta_{u_i}}(B) \mu_{n,k,\alpha}(d\mathbf{u})}{\int_{R^n} \mu_{n,k,\alpha}(d\mathbf{u})}, \tag{6.5.12}$$

where

$$\mu_{n,K,\alpha}(C) = \int_{C} \prod_{i=1}^{n} K(x_i, u_i) \prod_{i=1}^{n} \left(\alpha + \sum_{j=1}^{i-1} \delta_{u_j} \right) (du_i) \tag{6.5.13}$$

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

$$\mathcal{E}(g(G)|\mathbf{X}) = \frac{\int_{R^n} g(G) \mathcal{D}_{\alpha+\sum \delta_{u_i}}(dG) \mu_{n,k,\alpha}(d\mathbf{u})}{\int_{R^n} \mu_{n,k,\alpha}(d\mathbf{u})}. \tag{6.5.14}$$

Now, by taking $g(G) = f(x|G)$ and simplifying, the posterior expectation $\hat{f}(x|G)$ of $f(x|G)$ is derived as

$$\hat{f}_{\alpha}(x|G) = \mathcal{E}(f(x|G)|\mathbf{X}) = p_n f_0(x) + (1 - p_n) \hat{f}_n(x), \tag{6.5.15}$$

which is a convex combination of prior guess $f_0(x)$ defined above, and a quantity $\hat{f}_n(x)$, to be defined below, which mirrors the sample distribution function, but is complicated.

Let $k = k(\pi)$ denote the number of cells C_i in the partition π of $\{1, 2, \ldots, m\}$; cardinality of the ith cell being m_i $i = 1, \ldots, k$; $g_i(u)$, $i = 1, \ldots, m$, are m positive or α-integrable functions;

$$\varphi(\pi) = \prod_{i=1}^{k} \left\{ (m_i - 1)! \int_{R} \prod_{l \in C_i} g_l(u) \alpha(du) \right\}, \tag{6.5.16}$$

and finally let $w(\pi) = \varphi(\pi)/\sum_{\pi} \varphi(\pi)$. Then it is shown $\hat{f}_n(x)$ to be

$$\hat{f}_n(x) = \frac{1}{n} \sum_{\pi} w(\pi) \sum_{i=1}^{k} m_i \left\{ \frac{\int_{R} K(x, u) \prod_{l \in C_i} K(x_l, u) \alpha(du)}{\int_{R} \prod_{l \in C_i} K(x_l, u) \alpha(du)} \right\}, \tag{6.5.17}$$

where the summation is taken over all partitions π of $\{1, 2, \ldots, m\}$. $\hat{f}_n$ serves as a Bayes estimate under the loss function $L(f, \hat{f}) = \int |f(x|G) - \hat{f}(x|G)|^2 W(dx)$, where W is a known weight function.

The choice of the kernel K and the parameter α of the prior is critical, Lo gives several examples of $K(x, u)$ and α and computes the Bayes estimators. Among them include kernels of type histogram, normal with location and/or scale parameters, symmetric and unimodal densities, decreasing densities, etc. If K is chosen to reflect the histogram model, the estimator reduces to the usual Bayes estimates of cell probabilities. Kuo's (1986b) Monte Carlo method may be adapted to carry out the calculations. 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 $\hat{f}(x|G)$ at the $(n + 1)$th stage. Under certain conditions, they prove the asymptotic optimality of the resulting estimator.

A different formulation of the density function was considered by Ferguson (1983). 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 $\xi_i \overset{\text{iid}}{\sim} G_0$. 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 $\mathcal{E}(\mathcal{D}(\alpha + nG_n)) = (MG_0 + nG_n)/(M + n)$,

$$\mathcal{E}(G(\xi)|x_1, \ldots, x_n) = p_n G_0(\xi) + (1 - p_n) \int \ldots \int G_n(\xi)\, dH(\xi_1, \ldots, \xi_n|x_1, \ldots, x_n), \tag{6.5.18}$$

and

$$\hat{f}(x) = \mathcal{E}(f(x)|x_1, \ldots, x_n) = p_n f_0(x) + (1 - p_n)\hat{f}_n(x), \tag{6.5.19}$$

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 $\hat{f}_n(x)$ is given by

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \int \cdots \int h(x|\xi_i)\, dH(\xi_1, \ldots, \xi_n|x_1, \ldots, x_n). \tag{6.5.20}$$

Following Lo, $\hat{f}_n(x)$ can be written as a ratio $h(x, x_1, \ldots, x_n)/h(x_1, \ldots, x_n)$, where

$$h(x_1, \ldots, x_n) = \frac{1}{M^{(n)}} \int \cdots \int \left(\prod_{n=1}^{n} h(x_i|\xi_i)\right) \prod_{n=1}^{n} d\left(MG_0 + \sum_{j=1}^{i-1} \delta_{\xi_j}\right)(\xi_i), \tag{6.5.21}$$

and computations are carried out by Kuo's (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 $Y_i|\mu_i \sim N(\mu_i, 1)$, $\mu_i|G \overset{\text{iid}}{\sim} G$, 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 and using a nonparametric Bayesian approach. Escobar also uses a DP prior for G. When G is known the Bayesian estimator is the posterior mean

$$\mathcal{E}(\mu_i|Y_i) = \frac{\int \mu_i \phi(Y_i - \mu_i)\, dG(\mu_i)}{\int \phi(Y_i - \mu_i)\, dG(\mu_i)}, \tag{6.5.22}$$

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 DP prior for G. Antoniak has shown that if the DP prior is used for G, then the posterior distribution of μ_i is a mixture of DPs. 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. Therefore, he introduces new a 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 the next observation $Y_{n+1}|Y_1,\ldots,Y_n$ which is a mixture of normals, $Y_{n+1}|Y_1,\ldots,Y_n \sim N\left(\mu_{n+1},\sigma^2_{n+1}\right)$. A usual practice is to put a parametric prior on vector $\boldsymbol{\nu} = \left(\mu_1,\ldots,\mu_n,\sigma^2_1,\ldots,\sigma^2_n\right)$. Ferguson models the common prior for $\nu_i = \left(\mu_i,\sigma^2_i\right)$ as a DP prior. Thus the data is considered as coming from a Dirichlet mixture of normals in contrast to Antoniak where the DP processes were mixed with respect to a parametric distribution $H(\theta)$, $\alpha_\theta \sim H(\theta)$. A particular case of $\left(\mu_i,\sigma^2_i\right) = \left(\mu_i,\sigma^2\right)$ has been studied (see West 1992) in which the μ_i's distribution is modeled as DP with a normal base measure.

In view of the discreteness of DP prior which induces multiplicities of observations, $\nu_{n+1}|\nu_1,\ldots,\nu_n$ will have distribution of the form given in property 21 of Sect. 2.1.2. 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 $N\left(\mu_i,\sigma^2_i\right)$, and then it is shown that the unconditional predictive distribution is given by $Y_{n+1}|Y_1,\ldots,Y_n \sim \int P\left(Y_{n+1}|\boldsymbol{\nu}\right) dP\left(\boldsymbol{\nu}|Y_1,\ldots,Y_n\right)$. Since the evaluation of $P(\boldsymbol{\nu}|Y_1,\ldots,Y_n)$ is difficult even in small samples, they use Monte Carlo approximation using extensions of the iterative technique developed by Escobar (1994).

6.5.5 *Estimation of the Rank of X_1 Among $X_1,\ldots,X_n$*

Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} F$. 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 $F \sim D(\alpha)$, Campbell and Hollander obtained the posterior mean,

$$\hat{G} = \mathcal{E}(G\,|x_1,\ldots,x_r) = G' + (n-r)\{\alpha'((-\infty,x_1)) + \frac{1}{2}\alpha'(\{x_1\})/\alpha'(R)\}, \tag{6.5.23}$$

where $\alpha' = \alpha + \sum_{i=1}^{r}\delta_{x_i}$. $\hat{G}$ 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.

6.6 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.

6.6.1 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 $L(F, \hat{F}) = \int_{R^2} (F - \hat{F})^2 dW$, 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

$$\begin{aligned}\hat{F}_\alpha(x, y) &= \frac{\alpha((-\infty, x] \times (-\infty, y]) + \sum_{i=1}^{n} \delta_{(X_i, Y_i)}((-\infty, x] \times (-\infty, y])}{\alpha(R^2) + n} \\ &= p_n \frac{\alpha((-\infty, x] \times (-\infty, y])}{\alpha(R^2)} + (1 - p_n) \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}((-\infty, x] \times (-\infty, y]).\end{aligned} \tag{6.6.1}$$

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.

6.6.2 Estimation of F w.r.t. a Tailfree Process Prior

In Chap. 5, 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 squared error 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 (see Sect. 5.3). Consider, for example, $(x, y) = (\frac{1}{2}, \frac{3}{4})$. Following the notation of Sect. 5.3,

$$\begin{aligned}\widehat{F}\left(\frac{1}{2}, \frac{3}{4}\right) = \mathcal{E}\left[F\left(\frac{1}{2}, \frac{3}{4}\right)\right] &= \mathcal{E}[P(B_1) + P(B_{21}) + P(B_{23})] \\ &= \mathcal{E}[Z_1 + Z_2 Z_{21} + Z_2 Z_{23}] \\ &= \mathcal{E}[Z_1 + Z_{21} + Z_{23}] \\ &= \frac{\alpha_1}{\gamma_1} + \frac{\alpha_2}{\gamma_2}\frac{\alpha_{21}}{\gamma_{21}} + \frac{\alpha_2}{\gamma_2}\frac{\alpha_{23}}{\gamma_{23}},\end{aligned}$$

where $\gamma_{\underline{c_m}} = \alpha_{\underline{c_m}1} + \alpha_{\underline{c_m}2} + \alpha_{\underline{c_m}3} + \alpha_{\underline{c_m}4}$ and $\underline{c_m} = c_1 c_2 \ldots c_m$.

On the other hand, if $(x, y) = (\frac{1}{3}, \frac{1}{2})$, say, then

$$\begin{aligned}\widehat{F}\left(\frac{1}{3}, \frac{1}{2}\right) = \mathcal{E}\left[F\left(\frac{1}{3}, \frac{1}{2}\right)\right] &= \mathcal{E}[P(B_{11} \cup B_{12}) + P(\cup_{i=1}^{2} \cup_{j=1}^{2} (B_{13ij} \cup B_{14ij}) + \ldots] \\ &= \mathcal{E}\left[Z_{11} + Z_{12} + \sum_{i=1}^{2}\sum_{j=1}^{2}(Z_{13ij} + Z_{14ij}) + \ldots\right] \\ &= \frac{\alpha_{11}}{\gamma_{11}} + \frac{\alpha_{12}}{\gamma_{12}} + \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{\alpha_{13ij}}{\gamma_{13ij}} + \frac{\alpha_{14ij}}{\gamma_{14ij}}\right) + \ldots.\end{aligned}$$

Now given a sample $\mathbf{X}$, all we have to do is to update the α's.

6.6.3 *Estimation of a Covariance*

The covariance of P is defined for $(x, y) \in R^2$ by the formula

$$\text{Cov } P = \int xy dP - \int x dP \int y dP. \tag{6.6.2}$$

Assuming the squared error loss L_2 and $P \sim \mathcal{D}(\alpha)$, Ferguson (1973) derived its Bayesian estimator. For the no-sample problem we have

$$\mathcal{E}(\text{Cov } P) = \frac{\alpha(R^2)}{\alpha(R^2) + 1}\sigma_{12}, \tag{6.6.3}$$

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

$$\widehat{\mathrm{Cov}P_\alpha} = \frac{\alpha(R^2)+n}{\alpha(R^2)+n+1}(p_n\sigma_{12} + (1-p_n)s_{12} + p_n(1-p_n)(\mu_1 - \overline{X}_n)(\mu_2 - \overline{Y}_n)), \tag{6.6.4}$$

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.

6.6.4 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 $\tau = \mathcal{E}(\mathrm{sign})(X-X')(Y-Y')$, 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

$$\begin{aligned} T_1 &= \{(x, y, x', y') : (x - x')(y - y') > 0\} \text{ and} \\ T_2 &= \{(x, y, x', y') : (x - x')(y - y') = 0\}. \end{aligned} \tag{6.6.5}$$

Since F is allowed to be discrete, a slight modification of Δ, namely,

$$\Delta = P_F\{(X - X')(Y - Y') > 0\} + \frac{1}{2} \cdot P_F\{(X - X')(Y - Y') = 0\} \tag{6.6.6}$$

is preferred. The rationale is that the tied pairs are evenly distributed among concordants $(X - X')(Y - Y') > 0$ and discordants $(X - X')(Y - Y') < 0$. 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,

$$\Delta = \Delta_F = \int \left(I_{T_1} + \frac{1}{2} \cdot I_{T_2}\right) d(F(x, y)F(x', y')). \tag{6.6.7}$$

Assuming $F \sim \mathcal{D}(\alpha)$, and α defined on $(R^2, \mathcal{B}^2)$, the Bayes estimator of Δ for the no-sample problem is given by

$$\widehat{\Delta}_{\alpha 0} = \mathcal{E}_{\mathcal{D}(\alpha)}(\Delta_F) = \int \left(I_{T_1} + \frac{1}{2} \cdot I_{T_2}\right) d\mathcal{E}_{\mathcal{D}(\alpha)}(F(x, y)F(x', y'). \tag{6.6.8}$$

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,

$$\widehat{\Delta}_{\alpha 0} = \frac{M}{M+1}\Delta_G + \frac{1}{2(M+1)}, \tag{6.6.9}$$

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

$$\begin{aligned}\alpha + \sum_{i=1}^{n} \delta_{(X_i,Y_i)} &= (M+n)\left[\frac{M}{M+n}Q + \frac{1}{M+n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}\right] \\ &= (M+n)Q^*, \text{ say.}\end{aligned} \tag{6.6.10}$$

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 Bayesian estimator is given by,

$$\begin{aligned}\widehat{\Delta}_{\alpha n} &= \frac{M+n}{M+n+1}\int\left(I_{T_1} + \frac{1}{2}\cdot I_{T_2}\right) d(p_n G + (1-p_n)\widehat{G}_n) \\ &\quad \times (p_n G + (1-p_n)\widehat{G}_n) + \frac{1}{2(M+n+1)} \\ &= \frac{M+n}{M+n+1}[p_n^2\Delta_G + 2p_n(1-p_n)\Delta(G,\widehat{G}_n) + (1-p_n)^2\Delta_{\widehat{G}_n}] \\ &\quad + \frac{1}{2(M+n+1)},\end{aligned} \tag{6.6.11}$$

where

$$\Delta_{\widehat{G}_n} = \frac{1}{n^2}\sum_{i,j=1}^{n}\left(I_{[(x-x')(y-y')>0]} + \frac{1}{2}I_{[(x-x')(y-y')=0]}\right), \tag{6.6.12}$$

and $\Delta(G, \widehat{G}_n) = \frac{1}{n}\sum_{i=1}^{n}\Delta_G(x_i, y_i)$ with

$$\Delta_G(x_i, y_i) = \left\{P_G[(X-x_i)(Y-y_i) > 0] + \frac{1}{2}P_G[(X-x_i)(Y-y_i) = 0]\right\}. \tag{6.6.13}$$

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 $0(n^{-1})$.

6.7 Estimation of a Function of P

The examples of Sects. 6.4 and 6.6 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 $h(X_1, \ldots, X_k)$ such that $\phi(P) = \int_{\mathfrak{X}^k} h(x_1, \ldots, x_k) \prod_{i=1}^{k} dP(x_i)$. The degree of an estimable parameter $\phi(P)$ is the least sample size for which there is such an h (Zacks 1971, p. 151). The Bayes and empirical Bayes estimation of estimable parameters of degree 1 and 2 under the loss function L_2 and with respect to the Dirichlet and Dirichlet Invariant processes as priors were investigated by Yamato (1977a,b), Tiwari (1981) and Tiwari and Zalkikar (1985). Their results are as follows.

6.7.1 Dirichlet Process Prior

Based on a random sample $\mathbf{X}$ from P, the Bayesian estimator $\hat{\phi}$ of ϕ under L_2 loss is given by the posterior mean $\mathcal{E}(\phi(P) \mid X_1, \ldots, X_n)$. In particular, suppose $\phi(P) = \phi_h(P)$ and $P \sim \mathcal{D}(\alpha)$, where

$$\phi_h(P) = \int_{\mathfrak{X}^k} h(x_1, \ldots, x_k) dP(x_1) \cdots dP(x_k), \tag{6.7.1}$$

and h is a symmetric measurable function from χ^k into R satisfying

$$\int_{\mathfrak{X}^k} |h(x_1, \ldots, x_k)| d\overline{\alpha}(x_1) \cdots d\overline{\alpha}(x_m) < \infty, \tag{6.7.2}$$

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

$$\int_{\mathfrak{X}^m} |h(x_1,\ldots,x_1,x_2,\ldots,x_2,\ldots,x_m,\ldots,x_m)|^2 \, d\overline{\alpha}(x_1)\cdots d\overline{\alpha}(x_m) < \infty, \tag{6.7.3}$$

for all possible combinations of arguments $(x_1,\ldots,x_1,x_2,\ldots,x_2,\ldots,x_m,\ldots,x_m)$, $m \leq k$, 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 $\hat{\phi}^0_{h,\alpha} = \mathcal{E}_{\mathcal{D}(\alpha)}(\phi_h(P))$, and for the sample $X_1,\ldots,X_n$ it is

$$\hat{\phi}^n_{h,\alpha} = \mathcal{E}_{\mathcal{D}(\alpha+\sum_{i=1}^n \delta_{x_i})}(\phi_h(P)) = \hat{\phi}^0_{h,\alpha+\sum_{i=1}^n \delta_{X_i}}. \tag{6.7.4}$$

Thus using this expression and property 9 of Sect. 2.1.2, Yamato (1977a,b) and Tiwari (1981) derived the following result. Based on a sample $X_1,\ldots,X_n$, the Bayes estimator of $\phi_h(P)$ with respect to the prior $\mathcal{D}(\alpha)$ and loss L_2 is given by

$$\hat{\phi}^n_{h,\alpha} = \sum_{C(\sum im_i = k)} \frac{k![\alpha(\mathfrak{X})+n]^{\sum m_i}}{\prod_{i=1}^k [i^{m_i}(m_i)!][\alpha(\mathfrak{X})+n]^{(k)}} \times \int_{\mathfrak{X}^{\sum m_i}} h(x_{11},\ldots,x_{1m_1},x_{21},\ldots,x_{2m_2},\ldots,x_{k1},..,x_{km_k}) \prod_{i=1}^k \prod_{j=1}^{m_i} d\widehat{F}_\alpha(x_{ij}), \tag{6.7.5}$$

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 $\hat{\phi}^n_{h,\alpha} \to \hat{\phi}_{h,\,\sum_{i=1}^n \delta_{x_i}}$ as $\alpha(\mathfrak{X}) \to 0$.

Also, if $h(x_1,\ldots,x_k)$ is such that it vanishes whenever two coordinates are equal, then

$$\hat{\phi}_{h,\,\sum_{i=1}^n \delta_{x_i}} = \frac{n(n-1)\cdots(n-k+1)}{n^{(k)}} U_{h,n}, \tag{6.7.6}$$

where $U_{h,n}$ is the usual U-statistic based on the sample $X_1,\ldots,X_n$. Yamato (1977b) has proved that the asymptotic distribution of $\hat{\phi}^0_{h,\,\sum_{i=1}^n \delta_{x_i}}$ 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, 1991a) as

$$\hat{\phi}_1(P) = p_n \int h(x)d\overline{\alpha}(x) + \frac{(1-p_n)}{n}\sum_{i=1}^n h(X_i) \tag{6.7.7}$$

and

$$\hat{\phi}_2(P) = \frac{M+n}{M+n+1}\left\{p_n^2\int h(x,y)d\overline{\alpha}(x)d\overline{\alpha}(y) + \frac{2p_n(1-p_n)}{n}\sum_{i=1}^{n}\int h(x,x_i)d\overline{\alpha}(x) + \frac{(1-p_n)^2}{n^2}\sum_{i\neq j}h(X_i,X_j)\right\}. \tag{6.7.8}$$

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

$$\varsigma = \int h(x,y;x',y')dP(x,y)dP(x',y'), \tag{6.7.9}$$

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

$$\hat{\varsigma}_\alpha = \frac{M+m}{M+m+1}[p_m^2\varsigma_{\overline{\alpha}} + 2p_m(1-p_m)\varsigma(\overline{\alpha},F_m) + (1-p_m)^2\varsigma_{F_m}] + \frac{1}{M+m+1}\left\{p_m\int h(x,y;x,y)d\overline{\alpha}(x,y) + \frac{(1-p_m)}{m}\sum_{i=1}^{m}h(X_i,Y_i;X_i,Y_i)\right\}, \tag{6.7.10}$$

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

$$T_1 = \{(x,y;x',y') : (x-x')(y-y') > 0\} \tag{6.7.11}$$

and

$$T_2 = \{(x,y;x',y') : (x-x')(y-y') = 0\}, \tag{6.7.12}$$

the results of Dalal and Phadia (1983) can be obtained.

6.7.2 *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

$$\hat{F}^*_\alpha(t) = p_n F_0(t) + (1 - p_n)\hat{F}^*_n(t), \tag{6.7.13}$$

where, $F_0(t) = Q((-\infty, t])$ and $\hat{F}^*_n(t)$ is η-symmetrized version of the empirical distribution,

$$\hat{F}^*_n(t) = \frac{1}{2n}\sum_{i=1}^{n} \delta_{X_i}((-\infty, t]) + \delta_{2\eta - X_i}((-\infty, t]). \tag{6.7.14}$$

$\hat{F}^*_\alpha$ is identical to the one obtained by Dalal (1979a).

He (Yamato 1987) also 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

$$\hat{\phi}^*_{1\alpha} = p_n \int h(x) dQ(x) + \frac{(1 - p_n)}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} h(g_j X_i), \tag{6.7.15}$$

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

$$\int_{\mathfrak{X}^2} h(x, y) dQ(x) dQ(y) < \infty, \qquad \int_{\mathfrak{X}} h(x, gx) dQ(x) < \infty \text{ for any } g \in \mathcal{G}, \tag{6.7.16}$$

and let $X_1, \ldots, X_n$ be a sample from P, $P \sim \mathcal{DGI}(\alpha)$. Then the Bayes estimate of φ_2 under L_2 loss is (Yamato 1986, 1987)

$$\hat{\phi}^*_{2\alpha} = \frac{M+n}{M+n+1}\left[p_n^2 \int_{\mathfrak{X}^2} h(x,y)dQ(x)dQ(y) + \frac{2p_n(1-p_n)}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} \times \int_{\mathfrak{X}} h(x, g_j X_i)dQ(x) + \frac{(1-p_n)^2}{n^2k^2}\sum_{i_1,i_2}\sum_{j_1,j_2} h(g_{j_1}X_{i_1}, g_{j_2}X_{i_2})\right] + \frac{1}{k(M+n+1)}\left[p_n \sum_{j=1}^{k}\int_{\mathfrak{X}} h(x, g_j x)dQ(x) + \frac{(1-p_n)}{nk}\sum_{i}\sum_{j_1,j_2} h(g_{j_1}X_i, g_{j_2}X_i)\right]. \tag{6.7.17}$$

If we let M go to zero, the above estimator reduces to

$$\hat{\phi}^{**}_2 = \frac{1}{n(n+1)k^2}\sum_{j_1,j_2}\left[\sum_{i_1,i_2} h(g_{j_1}X_{i_1}, g_{j_2}X_{i_2}) + \sum_{i} h(g_{j_1}X_i, g_{j_2}X_i)\right], \tag{6.7.18}$$

and if we replace Dirichlet Invariant with Dirichlet process, clearly the estimator reduces to

$$\hat{\varphi}^{**}_{2D} = \frac{1}{n(n+1)}\left[\sum_{i_1,i_2} h(X_{i_1}, X_{i_2}) + \sum_{i} h(X_i, X_i)\right]. \tag{6.7.19}$$

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

$$\theta = \int_{R^2} x_1x_2dP(x_1,x_2) - \int_R x_1dP(x_1,x_2)\int_R x_2dP(x_1,x_2) \tag{6.7.20}$$

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

$$\hat{\theta} = \frac{1}{n(n+1)} \begin{bmatrix} \#\text{ of } \{\text{pairs } (i,j) : (X_i - X_j)(Y_i - Y_j) > 0, 1 \le i < j \le n\} \\ +\#\text{ of } \{\text{pairs } (i,j) : (X_i - Y_j)(Y_i - X_j) > 0, 1 \le i < j \le n\} \\ +\#\{i : X_i = Y_i, 1 \le i \le n\} - n. \end{bmatrix} \tag{6.7.21}$$

6.7.3 *Empirical Bayes Estimation of $\phi(P)$*

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, 1991a) 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 $0(n^{-1})$ 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

$$\hat{\phi}_{1\alpha} = p_m \phi_0 + (1 - p_m) U_{n+1}, \tag{6.7.22}$$

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. 6.2.4 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

$$\hat{\phi}_{\alpha 1, n+1} = \hat{p}_m \sum_{i=1}^{n} \frac{U_i}{n} + (1 - \hat{p}_m) U_{n+1}, \tag{6.7.23}$$

where, for the samples $\mathbf{X}_i, i = 1, 2, \ldots, n$, $U_i = \frac{1}{m} \sum_{j=1}^{m} h(X_{ij})$, $\hat{p}_m = \widehat{M}_n/(\widehat{M}_n + m)$, $\widehat{M}_n = \max(0, m(F_n - 1)^{-1})$, and 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, the asymptotic optimality relative to α with rate of convergence $O(n^{-1})$ is also established.

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

$$\hat{\phi}_{\alpha 2,n+1}(P) = \frac{M+m}{M+m+1}\left[p_m^2 \frac{M+1}{M}\sum_{i=1}^{n}\frac{U_{2i}}{n} + 2p_m(1-p_m)\sum_{k=1}^{n}\sum_{i=1}^{n}\frac{U_{n+1,i,k}}{mn}\right.$$
$$\left. +(1-p_m)^2 \sum_{1\le j\ne k\le m}\frac{h(X_{n+1,j}, X_{n+1,k})}{m^2}\right], \tag{6.7.24}$$

where for the ith sample, $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,m}), i = 1, 2, \ldots, n$ and,

$$U_{2i} = \frac{1}{m(m-1)}\sum_{1\le j\ne k\le m} h(X_{i,j}, X_{i,k}), \tag{6.7.25}$$

$$U_{n+1,i,k} = \frac{1}{m}\sum_{j=1}^{m} h(X_{n+1,k}, X_{i,j}),\ i = 1, 2, \ldots, n. \tag{6.7.26}$$

Under the assumption that $\int h^2(x,y)d\overline{\alpha}(x)d\overline{\alpha}(y)$ exists and is finite, it is shown that the sequence $\{\hat{\phi}_{2,n+1}\}$ is asymptotically optimal relative to α with the rate of convergence $O(n^{-1})$.

The empirical Bayes estimation of $\phi_2(P)$ is also considered in Ghosh (1985) and exact Bayes risk for Bayes and empirical Bayes estimators are computed. It is shown 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, 1991a) use Zehnwirth's estimate for M and established similar result (See their paper for details) and also obtain the empirical Bayes estimator for ς [see (6.7.9)] and proved its asymptotic optimality with rate of convergence $O(n^{-1})$.

Remark 6.4 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. (1988), where the main idea was to use past as well as the current data in estimating the parameters of the Dirichlet process prior. It is shown that by doing this, we get improved estimators in terms of smaller risks.

6.8 Two-Sample Problems

Suppose we have two independent samples, $X_1, \ldots, X_{n_1}$ from F and $Y_1, \ldots, Y_{n_2}$ 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.

6.8.1 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 $F \sim \mathcal{D}(\alpha_1)$ and independently, $G \sim \mathcal{D}(\alpha_2)$. 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 $X_1, \ldots, X_{n_1} \sim F$ and $Y_1, \ldots, Y_{n_2} \sim G$, we update the estimate Δ_0 and obtain the Bayesian estimate as $\hat{\Delta} = \int \widehat{F}_{\alpha_1} d\widehat{G}_{\alpha_2}$, where $\widehat{F}_{\alpha_1}$ and $\widehat{G}_{\alpha_2}$ are Bayes estimators of F and G, respectively, as obtained in Equation (6.2.1). Simplifying further we get

$$\begin{aligned}\widehat{\Delta}_{\alpha_1\alpha_2}(\mathbf{X}, \mathbf{Y}) &= p_{1n_1} p_{2n_2} \Delta_0 + p_{1n_1}(1 - p_{2n_2}) \frac{1}{n_2} \sum_{i=1}^{n_2} F_0(Y_i) \\ &\quad + (1 - p_{1n_1}) p_{2n_2} \frac{1}{n_1} \sum_{i=1}^{n_1} G_0(X_i-) + (1 - p_{1n_1})(1 - p_{2n_2}) \frac{1}{n_1 n_2} U, \end{aligned} \tag{6.8.1}$$

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 $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_1})$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_2})$ for $i = 1, 2, \ldots, n$ and propose the following estimator:

$$\begin{aligned}\widehat{\Delta}_{\alpha_1\alpha_2 n}(\mathbf{X}, \mathbf{Y}) &= p_{1n_1} p_{2n_2} \frac{1}{n^2 n_1 n_2} \sum_{j=1}^{n} \sum_{k=1}^{n_2} \sum_{i=1}^{n} \sum_{l=1}^{n_1} I_{(-\infty, Y_{jk}]}(X_{il}) \\ &\quad + p_{1n_1}(1 - p_{2n_2}) \frac{1}{n n_1 n_2} \sum_{k=1}^{n_2} \sum_{i=1}^{n} \sum_{l=1}^{n_1} I_{(-\infty, Y_{n+1k}]}(X_{il}) \\ &\quad + (1 - p_{1n_1}) p_{2n_2} \frac{1}{n_1} \sum_{l=1}^{n_1} \left\{ 1 - \frac{1}{n n_2} \sum_{j=1}^{n} \sum_{k=1}^{n_2} I_{(-\infty, X_{n+1l}]}(Y_{jk}) \right\} \\ &\quad + (1 - p_{1n_1})(1 - p_{2n_2}) \frac{1}{n_1 n_2} \sum_{l=1}^{n_1} \sum_{k=1}^{n_2} I_{(-\infty, Y_{n+1k}]}(X_{n+1l}). \end{aligned} \tag{6.8.2}$$

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.

6.8.2 Estimation of the Difference Between Two CDFs

A measure of the difference between two distributions functions F and G, is defined by

$$d(F,G) = \int (F(t) - G(t))^2 \, d\left(\frac{F(t) + G(t)}{2}\right), \tag{6.8.3}$$

which is somewhat difficult to handle. However, if the distributions are continuous on R, then it can be written as

$$d(F,G) = \frac{4}{3} - \left[\int G(t)\, dF^2(t) + \int F(t)\, dG^2(t)\right]. \tag{6.8.4}$$

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 $L(d(F,G), \hat{d}(F,G)) = (d(F,G) - \hat{d}(F,G))^2$. 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 6.5 (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,

$$\mathcal{E}[d(F,G)|X_1, \ldots, X_{n_1} \text{ and } Y_1, \ldots, Y_{n_2}]. \tag{6.8.5}$$

To evaluate this expectation, he defines (pseudo Bayesian estimators)

$$\begin{aligned}\widehat{F}_{\alpha_1 n_1}(t) &= p_{n_1} F_0(t) + (1-p_{n_1})\,\widehat{F}_{n_1}(t),\\ \widehat{G}_{\alpha_2 n_2}(t) &= p_{n_2} G_0(t) + (1-p_{n_2})\,\widehat{G}_{n_2}(t),\end{aligned} \tag{6.8.6}$$

on the interval (a,b), with $\widehat{F}_{\alpha_1 n_1}(t) = \widehat{G}_{\alpha_2 n_2}(t) = 0$ for $t \le a$, and $\widehat{F}_{\alpha_1 n_1}(t) = \widehat{G}_{\alpha_2 n_2}(t) = 1$ for $t \ge b$, 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

$$\begin{aligned}\hat{d}_{\alpha_1\alpha_2}(F,G) = \frac{4}{3} &- \frac{\alpha_1(R)+n_1}{\alpha_1(R)+n_1+1}\int_a^b \widehat{G}_{\alpha_2 n_2,\Delta}(t)\,d\widehat{F}^2_{\alpha_1 n_1,\Delta}(t)\\ &- \frac{1}{\alpha_1(R)+n_1+1}\left\{\frac{2}{3}\int_a^b \widehat{G}_{\alpha_2 n_2,\Delta}(t)\,d\widehat{F}_{\alpha_1 n_1,\Delta}(t)\right.\\ &\left.+\frac{1}{3}\sum_1^{k-1}\widehat{G}_{\alpha_2 n_2}(t_{i+1})[\widehat{F}_{\alpha_1 n_1}(t_{i+1}) - \widehat{F}_{\alpha_1 n_1}(t_i)]\right\} \qquad (6.8.7)\\ &- \frac{\alpha_2(R)+n_2}{\alpha_2(R)+n_2+1}\int_a^b \widehat{F}_{\alpha_1 n_1,\Delta}(t)\,d\widehat{G}^2_{\alpha_2 n_2,\Delta}(t)\\ &- \frac{1}{\alpha_2(R)+n_2+1}\left\{\frac{2}{3}\int_a^b \widehat{F}_{\alpha_1 n_1,\Delta}(t)\,d\widehat{G}_{\alpha_2 n_2,\Delta}(t)\right. \qquad (6.8.8)\\ &\left.+\frac{1}{3}\sum_1^{k-1}\widehat{F}_{\alpha_1 n_1}(t_{i+1})[\widehat{G}_{\alpha_2 n_2}(t_{i+1}) - \widehat{G}_{\alpha_2 n_2}(t_i)]\right\}. \qquad (6.8.9)\end{aligned}$$

Finally, taking the limit $\|\Delta\pi\| \to 0$, the above estimator reduces to

$$\begin{aligned}\hat{d}_{\alpha_1\alpha_2}(F,G) = \frac{4}{3} &- \frac{1}{\alpha_1^*+1}\left\{\int_a^b \widehat{G}_{\alpha_2 n_2}(t)\,d\widehat{F}_{\alpha_1 n_1}(t) + \alpha_1^*\int_a^b \widehat{G}_{\alpha_2 n_2}(t)\,d\widehat{F}^2_{\alpha_1 n_1}(t)\right\}\\ &- \frac{1}{\alpha_2^*+1}\left\{\int_a^b \widehat{F}_{\alpha_1 n_1}(t)\,d\widehat{G}_{\alpha_2 n_2}(t) + \alpha_2^*\int_a^b \widehat{F}_{\alpha_1 n_1}(t)\,d\widehat{G}^2_{\alpha_2 n_2}(t)\right\},\end{aligned} \tag{6.8.10}$$

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 its support. In general, when F and G are continuous, the author proposes an estimator of $d(F,G)$ as $\hat{d}(F,G)$ 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 (6.8.10) 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 $\hat{d}_{\alpha_1\alpha_2}(F, G)$ (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.

6.8.3 *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 $F \sim \mathcal{D}(\alpha)$. Under the loss function L_1, the Bayes estimator for the no-sample problem is found by minimizing the integrand of

$$\mathcal{E}(L(\Delta, \hat{\Delta})) = \int \mathcal{E}(\Delta(x) - \hat{\Delta}(x))^2 dW(x) \tag{6.8.11}$$

yielding $\hat{\Delta}_0(x) = \mathcal{E}(\Delta(x)) = \mathcal{E}\{G^{-1}F(x)\} - x$. 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 $\hat{\Delta}_0$ is

$$\begin{aligned}\hat{\Delta}_0(x) &= \frac{1}{\lambda \cdot B(\alpha', \beta')} \cdot \int_0^1 \sum_{j=1}^{\infty} \frac{y^{\alpha'+j-1}(1-y)^{\beta'-1}}{j} dy - x \\ &= \frac{1}{\lambda} \cdot \sum_{j=1}^{\infty} \frac{B(\alpha'+j, \beta')}{j \cdot B(\alpha', \beta')} - x, \end{aligned} \tag{6.8.12}$$

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 $\hat{\Delta}_0(x)$ 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.

6.9 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.

6.9.1 Testing $H_0 : F \leq F_0$

Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random sample from the distribution function F and let F_0 be a known distribution function. Consider the problem of testing hypothesis $H_0 : F \leq F_0$ against the alternative $H_1 : F \not\leq F_0$ when the loss function L is given by

$$L(F, a_0) = \int (F - F_0)^+ dW \quad \text{and} \quad L(F, a_1) = \int (F - F_0)^- dW, \tag{6.9.1}$$

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 known weight function, $\mathrm{a}^+ = \max\{a, 0\}$ and $\mathrm{a}^- = -\min\{a, 0\}$ for any real number a. Assume $F \sim \mathcal{D}(\alpha)$. 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

$$r_m(\delta, \alpha) = \int \mathcal{E}\,[L(F, a_0) - L(F, a_1) \mid \mathbf{X}]\, \delta(\mathbf{X})\, d\, Q_m(\mathbf{X}) + \mathcal{E}\,[L(F, a_1)], \tag{6.9.2}$$

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

$$r_m^*(\alpha) = \int_{[\Delta_m(\mathbf{X}) \leq 0]} \Delta_m(\mathbf{X})\, d\, Q_m(\mathbf{X}) + \mathcal{E}\,[L(F, a_1)]. \tag{6.9.3}$$

If α is known, $\Delta_m(\mathbf{X})$ can be easily evaluated since for each u, $F(u)|\mathbf{X} = \mathbf{x} \sim \mathrm{Be}(\alpha(-\infty, u] + \sum_{i=1}^m I[X_i \leq u], \alpha(u, \infty) + \sum_{i=1}^m I[X_i > u])$.

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 ith stage. Then an empirical Bayes rule at the $(n+1)$th stage is given by

$$\xi_{n+1}(\mathbf{X}_{n+1}) = \mathcal{P}\{\text{accepting } a_0 \mid \mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{X}_{n+1}\}, \tag{6.9.4}$$

Let $\widehat{\Delta}_n(\mathbf{X}_{n+1})$ be an estimate based on $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ of $\Delta_{m_{n+1}}(\mathbf{X}_{n+1}) = \int \mathcal{E}\,[F_{n+1}(u) - F_0(u) \mid \mathbf{X}_{n+1}]\, dW(u)$ given by

$$\widehat{\Delta}_n(\mathbf{x}_{n+1}) + \int F_0 dW = \int \frac{\{\hat{\alpha}(-\infty, u] + \sum_{i=1}^{m_{n+1}} I[X_{n+1,i} \le u]\}dW(u)}{(1 + m_{n+1})}, \tag{6.9.5}$$

where $\hat{\alpha}(-\infty, u] = n^{-1} \sum_{j=1}^{n} m_j^{-1} \sum_{i=1}^{m_j} [X_{j,i} \le u]$. Let $\xi_n(\mathbf{x}_{n+1}) = I[\widehat{\Delta}_n(\mathbf{X}_{n+1}) \le 0]$, and let $r_{n+1}(\xi_n)$ denote the risk of using ξ_n to decide about F_{n+1}. Then it is proved that $r_{n+1}(\xi_n) - r^*_{m_{n+1}}(\alpha) \le n^{-\frac{1}{2}}$.

When $\alpha(R)$ is unknown, they estimate it by a consistent estimator $\hat{\alpha}(R) = (\log m_n)^{-1}\{\#$ of distinct observations in $\mathbf{X}_n\}$ (see property 19 of Sect. 2.1.2). Let ξ_n^* be the rule obtained by substituting this estimator in ξ_n with

$$\widehat{\Delta}_n(\mathbf{X}_{n+1}) + \int F_0 dW = \int \frac{\{\hat{\alpha}(R)\hat{\alpha}(-\infty, u] + \sum_{i=1}^{m_{n+1}} [X_{n+1,i} \le u]\}dW(u)}{(\hat{\alpha}(R) + m_{n+1})}. \tag{6.9.6}$$

When α is nonatomic and $m_{n+1} \to \infty$ as $n \to \infty$, it is shown that $r_{n+1}(\xi_n^*) - r^*_{m_{n+1}}(\alpha) = O((m_{n+1})^{-1}(\min\{\log m_n, n\})^{-1/2})$.

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.

6.9.2 *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:

$$\begin{aligned} H_0 &: F(x, y) \ge F_X(x)F_Y(y) \text{ for all } (x, y) \text{ in } R^2 \\ H_1 &: F(x, y) < F_X(x)F_Y(y) \text{ for all } (x, y) \text{ in } R^2, \end{aligned} \tag{6.9.7}$$

under the loss function

$$L(F, a_0) = \int (F(x, y) - F_X(x)F_Y(y))^- dW(x, y)$$
$$L(F, a_1) = \int (F(x, y) - F_X(x)F_Y(y))^+ dW(x, y), \tag{6.9.8}$$

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

$$\theta(\mathbf{x}, \mathbf{y}) = I_{[\Delta_n(\mathbf{x},\mathbf{y})]}, \tag{6.9.9}$$

where

$$\Delta_n(\mathbf{x}, \mathbf{y}) = \mathcal{E}[L(F, a_0) - L(F, a_1) \mid (\mathbf{X}, \mathbf{Y})]$$
$$= \int [\mathcal{E}(F(x', y') - F_X(x')F_Y(y')) \mid (\mathbf{X}, \mathbf{Y})] dW(x', y'). \tag{6.9.10}$$

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

$$\frac{MG_0(x', y') + \sum_{i=1}^n \delta_{(x_i,y_i)}((-\infty, x'] \times (-\infty, y'])}{M + n} - \frac{G^*(x', y') + MG_X^*(x')G_Y^*(y')}{M + n + 1}, \tag{6.9.11}$$

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.

6.9.2.1 Testing the Hypothesis $H_0 : F \leq G$ Against the Alternative $H_1 : F \not\leq G$

An analog of the test discussed in Sect. 6.9.1 in a two-sample situation is to test the hypothesis $H_0 : F \leq G$ against the alternative $H_1 : F \not\leq G$. This topic is covered more generally in Sect. 7.6 based on randomly right-censored samples. Its application to the uncensored data as a special case is obvious and therefore it will not be presented here.

6.9.3 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 $\mathfrak{X} = \mathbb{N}$ to be a set of nonnegative integers and $\sigma(\mathbb{N})$ 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 $\sigma(\mathbb{N})$ 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 nonrandomized decision rule which minimizes the expected loss. First note that $F_i|\mathbf{X}_i \sim \mathcal{D}\left(\alpha_i^*\right)$, where $\alpha_i^* = \left(\alpha_{i0}^*, \alpha_{i1}^*, \ldots\right)$, 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

$$r_i(\boldsymbol{\pi}, \boldsymbol{\alpha}) = \sum_{j=1}^{k} L(i, j) P(j|Y_1, \ldots, Y_n) = \sum_{j=1}^{k} L(i, j) \frac{\pi_j P(\mathbf{Y}|j)}{\sum_{j=1}^{k} \pi_j P(\mathbf{Y}|j)}, \tag{6.9.12}$$

where

$$P(\mathbf{Y}|j) = \prod_{l=0}^{\infty} \frac{\alpha_{jl}^{*(k_l)}}{\alpha_l^{*(n)}}, \quad a^{(n)} = a(a+1)\cdots(a+n-1), n > 0, \tag{6.9.13}$$

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 7
Inference Based on Incomplete Data

7.1 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 nonnegative random variables defined on R^+ also or positive real numbers (to be specified in the sequel). 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 its 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, and survival data analysis (Gross and Clark 1975). However, its application and usefulness in the analysis of survival data arising in clinical research have 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 is

$$\widehat{S}_{\mathrm{PL}}(u) = \frac{N^+(u)}{n} \prod_{j=1}^{n} \left(\frac{N^+(Z_j) + \lambda_j}{N^+(Z_j)} \right)^{I\left[\delta_i = 0, Z_j \leq u\right]}, \tag{7.1.1}$$

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1_7

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, hypothesis testing, etc.

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 presented here. 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 as well as beta-Stacy process, among others, and their treatment is considered in more detail and formulas are presented.

As in the previous chapter, estimation of the distribution (survival) function is of prime interest and therefore it is considered first in Sect. 7.2 assuming the Dirichlet process prior. Also estimation of the survival function under different censoring schemes is presented. In Sect. 7.3, the estimation of the survival function based on other types of priors is considered. In a slight digression, a linear Bayes estimation of the survival function is presented in Sect. 7.4. Various other estimation problems are included in Sect. 7.5. Section 7.6 deals with a hypothesis testing problem and finally, estimation of the survival function incorporating covariates is presented in the last section (Sect. 7.7).

7.2 Estimation of an 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.

7.2.1 Estimation Based on Right Censored Data

Based on the data $(\mathbf{Z}, \boldsymbol{\delta})$, we shall obtain in this section the Bayesian estimation of the survival function $S(t)$, under the loss function L_1 used earlier

$$L_1(S, \widehat{S}) = \int_0^\infty (S(t) - \widehat{S}(t))^2 W(t), \tag{7.2.1}$$

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 $(F, X_1, \ldots, X_n)$. Susarla and Van Ryzin's strategy was to take care of real observations first by noting that the posterior distribution of F given these observations is again the Dirichlet process with updated parameter of the DP. Then the censored observations were dealt with the updated parameter.

Observe that $(\mathbf{Z}, \boldsymbol{\delta})$ are invariant for any permutation of observed pairs $(\delta_1, Z_1), \ldots, (\delta_n, Z_n)$. 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 distribution of F given $(1, Z_1), \ldots, (1, Z_k)$ which is $\mathcal{D}(\alpha_k)$. Thus, we have

$$\widehat{S}_\alpha(u) = \mathcal{E}_{\mathcal{D}(\alpha_k)}\{S(u) \mid (0, Z_{k+1}), \ldots, (0, Z_n)\}. \tag{7.2.2}$$

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 $Z_{(m+1)} = \infty$:

$$\widehat{S}_\alpha(u) = \frac{\alpha(u, \infty) + N^+(u)}{\alpha(R^+) + n} \prod_{j=k+1}^{l} \frac{\alpha[Z_{(j)}, \infty) + N^+(Z_{(j)}) + \lambda_j}{\alpha[Z_{(j)}, \infty) + N^+(Z_{(j)})}, \tag{7.2.3}$$

where $N^+(u)$ is the number of observations greater than u. Alternatively, it can be written as

$$\widehat{S}_\alpha(u) = \frac{\alpha(u, \infty) + N^+(u)}{\alpha(R^+) + n} \prod_{j=1}^{n} \left(\frac{\alpha[Z_j, \infty) + N^+(Z_j) + \lambda_j}{\alpha[Z_j, \infty) + N^+(Z_j)} \right)^{I[\delta_i = 0, Z_j \le u]}. \tag{7.2.4}$$

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 everywhere.
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 $Z_{(l)} \le u < Z_{(l+1)}$, with $l = k, \ldots, m$, $Z_{(k)} = 0$ and $Z_{(m+1)} = \infty$, the conditional p-th moment of $S(u)$ given the data is shown to be

$$\mathcal{E}\left[S^p(u) \mid (\boldsymbol{\delta}, \mathbf{Z})\right] = \prod_{s=0}^{p-1} \left\{ \frac{\alpha(u,\infty) + s + N^+(u)}{\alpha(R^+) + s + n} \times \prod_{j=k+1}^{l} \frac{\alpha[Z_{(j)},\infty) + s + N^+(Z_{(j)}) + \lambda_j}{\alpha[Z_{(j)},\infty) + s + N^+(Z_{(j)})} \right\}, \quad (7.2.5)$$

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 7.1 (Blum and Susarla) *Let G_j be absolutely continuous or discrete distribution for $j = 1, \ldots, n$ and let $Y_1, \ldots, Y_n$ be independent of $(F, X_1, \ldots, X_n)$. Then the posterior distribution of P given $(\mathbf{Z}, \boldsymbol{\delta})$ is a mixture of Dirichlet processes with transition measure $\beta(\cdot) + \sum_{j=1}^{k-1} \mu_j(\cdot) + I.(u)$ 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*

$$\mu_l(B) = \frac{\beta(B \cap [Z_{n-k+l}, Z_{n-k+l-1}))}{\beta([Z_{n-k+1}, \infty)) + l - 1} \times \sum_{j=1}^{l-1} \frac{\beta\left(B \cap [Z_{n-k+j}, Z_{n-k+j-1})\right)}{\beta([Z_{n-k+1}, \infty)) + j - 1} \prod_{i=j}^{l-1} \frac{\beta([Z_{n-k+i}, \infty)) + i}{\beta([Z_{n-k+i+1}, \infty)) + i} \quad (7.2.6)$$

for $l = 2, \ldots, k$.

If we assume that G_i's are identical to G, that is, $Y_i \sim G, i = 1, 2, \ldots, n$, and that G is a fixed unknown continuous distribution on R^+, then in equation (7.2.4) $\lambda_j = 1$ for all j and the Bayes estimator $\widehat{S}_\alpha$ becomes

$$\widehat{S}_\alpha(u) = \frac{\alpha(u,\infty) + N^+(u)}{\alpha(R^+) + n} \prod_{j=1}^{n} \left(\frac{\alpha[Z_j,\infty) + N^+(Z_j) + 1}{\alpha[Z_j,\infty) + N^+(Z_j)} \right)^{I[\delta_i=0, Z_j \le u]}. \tag{7.2.7}$$

This representation of $\widehat{S}_\alpha$ is used in proving asymptotic properties of the Bayes estimator. It is shown (Susarla and Van Ryzin 1978b,c) that $\widehat{S}_\alpha$ is almost surely consistent with a convergence rate of $O\left(\log n / n^{\frac{1}{2}}\right)$, 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. 7.5.3.

7.2.2 *Empirical Bayes Estimation*

Susarla and Van Ryzin (1978a) also considered the empirical Bayes approach in estimating the survival function. For simplicity we consider the case of sample size one only. The general case is straightforward. 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 $\hat{\alpha}_n$ 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

$$\widehat{S}_{n+1}(u) = \begin{cases} \kappa(1 + \hat{\alpha}_n(u,\infty)) & \text{for } u < Z_{n+1} \\ \kappa\hat{\alpha}_n(u,\infty) \left(\frac{\hat{\alpha}_n(Z_{n+1},\infty) + I[\delta_{n+1}=0]}{\hat{\alpha}_n(Z_{n+1},\infty)} \right) & \text{for } u \ge Z_{n+1} \end{cases}, \tag{7.2.8}$$

where $\kappa = \left(1 + \alpha(R^+)\right)^{-1}$ and

$$\frac{\hat{\alpha}_n(u,\infty)}{\alpha(R^+)} = \frac{1}{nG_j(u)} \sum_{j=1}^{n} I[Z_j > u]. \tag{7.2.9}$$

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 $\hat{\alpha}_n$ is replaced by

$$\frac{\hat{\alpha}_n(u,\infty)}{\alpha(R^+)} = \frac{N^+(u)}{n}\prod_{j=1}^{n}\left(\frac{N^+(Z_j)+1+c}{N^+(Z_j)+c}\right)^{\left[\delta_j=0,\, Z_j\le u\right]}, \tag{7.2.10}$$

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.

7.2.3 *Estimation Based on a Modified Censoring Scheme*

In extending the results of Susarla and 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 $X \ge x$. 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 $X_{m_1+m_2+1} \ge x_{m_1+m_2+1}, \ldots, X_{m_1+m_2+m_3} \ge x_{m_1+m_2+m_3}$, variables 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). (Lo 1993a, Lemma 7.1, 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. The vectors $\mathbf{u}, \boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{\mu}$ will be referred to as the *data.* If

$F \sim \mathcal{D}(\alpha)$, then the posterior expectation of the survival function $S(t)$ is

$$\mathcal{E}(S(t)|\text{data}) = \frac{\alpha(t,\infty) + h_{j(t)}}{\alpha(R) + n} \prod_{i=1}^{j(t)} \frac{(\alpha[u_i,\infty) + h_{i-1})(\alpha(u_i,\infty) + h_i + \lambda_i)}{(\alpha(u_i,\infty) + h_i)(\alpha[u_i,\infty) + h_i + \lambda_i + \delta_i)}. \tag{7.2.11}$$

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 $i = 1, \ldots, k$.

7.2.4 *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 $Z_{(k)}$, $1 \le k \le n$, and then in effect, the statistician has at his disposal only the data

$$\{(\delta_i^*, Z_{(i)}), 1 \le i \le k,\ Z_{(r)} > Z_{(k)},\ r = k+1, \ldots n\}, \tag{7.2.12}$$

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 which 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. 7.2.1 gives

$$\widehat{S}_k(u) = \mathcal{E}\ \{S(u) \mid (0, Z_{l+1}), \ldots, (0, Z_k), 1 \le l \le k,\ Z_{(r)} > Z_{(k)},\ r = k+1, \ldots n\}, \tag{7.2.13}$$

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 (7.2.13) becomes

$$\begin{aligned}\widehat{S}_k(u) = {} & \frac{\alpha(u,\infty) + N^+(u) + (n-k)I[Z_{(k)} > u]}{\alpha(R^+) + n} \\ & \times \prod_{j=1}^{k} \left(\frac{\alpha(Z_{(j)},\infty) + N^+\left(Z_{(j)}\right) + (n-k) + 1}{\alpha(Z_{(j)},\infty) + N^+\left(Z_{(j)}\right) + (n-k)} \right)^{I\left[\delta_j^* = 0, Z_{(j)} \le u\right]} \\ & \times \left(\frac{\alpha(Z_{(k)},\infty) + (n-k)}{\alpha(Z_{(k)},\infty)} \right)^{\left[Z_{(k)} \le u\right]}. \end{aligned} \tag{7.2.14}$$

Note that by setting $k(n) = n$ in (7.2.14) yields Susarla and Van Ryzin (1976) estimator (7.2.4) with $\lambda_j = 1$ for all j. Also, if there are no observed censoring times in the data of (7.2.12), estimator (7.2.14) reduces to

$$\widehat{S}(u) = \left\{ \frac{\alpha(u,\infty) + N^+(u) + (n-k)I[Z_{(k)} > u]}{\alpha(R^+) + n} \right\} \times \left(\frac{\alpha(Z_{(k)},\infty) + (n-k)}{\alpha(Z_{(k)},\infty)} \right)^{I\left[Z_{(k)} \le u\right]}, \tag{7.2.15}$$

which in turn may be viewed as a generalization of Ferguson's (1973) estimator $\big(\alpha(u,\infty) + N^+(u)\big) / \big(\alpha(R^+) + n\big)$, when censoring is absent and $k(n) = n$. For $u \le Z_{(k)}$, formula (7.2.14) yields the (Kaplan and Meier 1958) Product-Limit estimator, which itself reduces to the empirical survival function in the absence of censoring.

7.2.5 *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) treated 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

$$\widehat{S}_\alpha(u) = \frac{\alpha(u,\infty) + N^+(u)}{\alpha(R^+) + n} \prod_{j=1}^{n} \left(\frac{\alpha[Z_j,\infty) + N^+(Z_j) + \lambda_j}{\alpha[Z_j,\infty) + N^+(Z_j)} \right)^{I\left[\delta_j = 0,\, Z_j \le u\right]/\lambda_j}. \tag{7.2.16}$$

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 setup of Sect. 6.2.4, they derive an empirical Bayes estimator and show that it is asymptotically optimal.

7.2.6 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:

$$\widehat{S}(u) = \frac{\alpha(u,\infty) + n(S_n(u) - \overline{G}_n(u))}{\alpha(R^+)} \prod_{i:T_i<u}^{n} \frac{\alpha(T_i,\infty) + n(S_n(T_i^-) - \overline{G}_n(T_i^-))}{\alpha(T_i,\infty) + n(S_n(T_i^-) - \overline{G}_n(T_i))}, \tag{7.2.17}$$

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 $\alpha(R^+) \to 0$, it is shown 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. The weak convergence of the above estimator and a numerical example are furnished.

7.2.7 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 $S_Z(t) \sim \mathcal{D}(\alpha)$ and $\theta \sim \text{Be}(a, b)$, a beta distribution with parameters a and b, and that $S_Z(t)$ and θ are independent. Then

the posterior distributions are $S_Z(t) \mid (\mathbf{Z}, \boldsymbol{\delta}) \sim \mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{Z_i})$ and $\theta \mid (\mathbf{Z}, \boldsymbol{\delta}) \sim \text{Be}(a + N_u, b + n - N_u) = \text{Be}(a^*, b^*)$, 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

$$\widehat{S}_X(t) = E[(S_Z(t))^{\theta} \mid (\mathbf{Z}, \boldsymbol{\delta})] = \mathcal{E}_{\text{Be}(a^*, b^*)} \left[\mathcal{E}_{\mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{Z_i})} \left(S_Z^{\theta}(t) \mid \theta \right) \right]. \tag{7.2.18}$$

Appealing to the moments of the Dirichlet process, the quantity inside the square brackets is

$$\mathcal{E}_{\mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{Z_i})} \left(S_Z^{\theta}(t) \mid \theta \right) = \frac{\Gamma\left(\alpha\left(R^+\right) + n\right)}{\Gamma\left(\alpha\left(R^+\right) + n + \theta\right)} \frac{\Gamma\left(\alpha(t, \infty) + N^+(Z_1) + \theta\right)}{\Gamma\left(\alpha(t, \infty) + N^+(Z_1)\right)} \cdots, \tag{7.2.19}$$

where, as before, $N^+(u) = \sum_{i=1}^{n} I[Z_i > u]$. However, explicit evaluation of this expression is difficult and therefore some approximations by expanding the ratios of gamma functions are provided. Final evaluation of the estimator $\widehat{S}_X$ is then proceeded by taking the expectation of this quantity with respect to $\text{Be}(a^*, b^*)$. Ghorai proves some asymptotic properties of this estimator, in particular, almost sure consistency and weak convergence to a Gaussian process.

7.2.8 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 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 $F \sim \mathcal{D}(\alpha)$. 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 $i = 1, \ldots, k+1$. The authors take the density of the vector $(p_1, p_2 \ldots, p_k)$ with respect to the measure $dv = \prod_{i=1}^{k} dp_i / \prod_{i=1}^{k+1} p_i$ where $p_{k+1} = 1 - \sum_{i=1}^{k} p_i$, over the simplex

$$S_k = \left\{ (p_1, p_2 \ldots, p_k) : p_i \geq 0 \text{ for } i = 1, \ldots, k, \text{ and } \sum_{i=1}^{k} p_i \leq 1 \right\}. \tag{7.2.20}$$

The prior density of $(p_1, p_2 \ldots, p_k)$ over S_k with respect to dv is proportional to $\prod_{i=1}^{k+1} p_i^{\beta_i}$. With this formulation they prove 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

$$\widehat{S}(t) = \frac{\alpha(t, \infty) + h_{j(t)}}{\alpha(R) + n} \prod_{i=1}^{j(t)} \frac{(\alpha(u_i, \infty) + h_i + \lambda_i)}{(\alpha(u_i, \infty) + h_i)}, \tag{7.2.21}$$

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 denotes 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 7.2 It is clear from all of the above results 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, as the Bayesian approach envisages. This may be construed as another attractive feature of the Dirichlet process.

7.3 Estimation of an 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, and beta-Stacy, are presented. They have similar form as for the case of the Dirichlet process prior, and in many cases the estimators are a different version of the Susarla-Van Ryzin estimator. In case of the processes neutral to the right, a posterior moment generating function (MGF) 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.

7.3.1 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 $i = 1, \ldots, n$, the pairs (Z_i, δ_i) are independent, and $X_1, \ldots, X_n \overset{iid}{\sim} S$. 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 $\hat{S}_u$ and $\hat{S}_c$ 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 $L(S., \hat{S}.) = \int_0^\infty (S. - \hat{S}.)^2 dW$:

$$\hat{S}_u(t) = \frac{\alpha^*((t, \infty), \{1\}) + \sum_{i=1}^n I[Z_i > t, \delta_i = 1]}{\alpha^*(\mathcal{R}^*) + n} \tag{7.3.1}$$

and

$$\hat{S}_c(t) = \frac{\alpha^*((t, \infty), \{0\}) + \sum_{i=1}^n I[Z_i > t, \delta_i = 0]}{\alpha^*(\mathcal{R}^*) + n}. \tag{7.3.2}$$

To unify the discrete and continuous cases of S, he follows Kalbfleisch and Prentice (1980, pp. 7–9) and defines

$$\Lambda(t) = -\int_0^{t^+} \frac{dS(u)}{S(u^-)} \tag{7.3.3}$$

and

$$\gamma(\Lambda)(t) = \lim_{k \to \infty} \prod_{i=1}^{k} \{1 - [\Lambda(u_i) - \Lambda(u_{i-1})]\}, \tag{7.3.4}$$

where $0 = u_0 < u_1 < \cdots < u_k = t$, 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,

$$S(t) = \gamma(\Lambda)(t) = \exp\left\{\oint_0^t \frac{dS(u)}{S(u^-)}\right\} \prod_{s \le t} \left(1 - \frac{\Delta S(s)}{S(s^-)}\right), \tag{7.3.5}$$

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 $\hat{S}$ of S is obtained as

$$\hat{S}(t) = \gamma\left(-\int_0^{t^+} \frac{d\hat{S}_u(u)}{\hat{S}_u(u^-) + \hat{S}_c(u^-)}\right)(t)$$
$$= \gamma\left(-\int_0^{t^+} \frac{d(\alpha^*((u,\infty),\{1\}) + \sum_{i=1}^n I[Z_i > u, \delta_i = 1])}{\alpha^*([u,\infty),\{0,1\}) + \sum_{i=1}^n I[Z_i \geq u]}\right)(t). \quad (7.3.6)$$

Then it is shown that $\hat{S}$ is the Bayes estimator of S under the loss function

$$L(S,\hat{S}) = \int_0^\infty [\gamma^{-1}(S)(t) - \gamma^{-1}(\hat{S})(t)]^2 dW(t), \quad (7.3.7)$$

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 $\hat{S}$ 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 in this chapter.

7.3.2 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. 7.2.3, Lo (1993a, Lemma 7.1), 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 was presented for a sample of size one in Sect. 4.2, Theorem 7.3.

For the general case, as mentioned there, it is easy to work with the 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. 7.2.3 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 7.3 (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*

$$M_t(\theta \mid \text{data}) = \frac{M_t(\theta + h_j(t))}{M_t(h_j(t))} \cdot \prod_{i=1}^{j(t)} \left[\frac{M_{u_i}^-(\theta + h_{i-1})}{M_{u_i}^-(h_{i-1})} \cdot \frac{C_{u_i}(\theta + h_i + \lambda_i, \delta_i)}{C_{u_i}(h_i + \lambda_i, \delta_i)} \cdot \frac{M_{u_i}(h_i)}{M_{u_i}(\theta + h_i)} \right], \tag{7.3.8}$$

where, if u is a prior fixed point of discontinuity of Y_t,

$$C_u(\alpha, \beta) = \int_0^\infty e^{-\alpha z}(1 - e^{-z})^\beta d\, G_u(z) \tag{7.3.9}$$

while, if u is not a prior fixed point of discontinuity of Y_t,

$$C_u(\alpha, \beta) = \begin{cases} \int_0^\infty e^{-\alpha z}(1 - e^{-z})^{\beta-1} d\, H_u(z) & \text{if } \beta \geq 1 \\ 1 & \text{if } \beta = 0. \end{cases} \tag{7.3.10}$$

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, $\mathcal{E}(S(t)|\text{data}) = M_t(1|\text{data})$. 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).

7.3.3 *Estimation Based on a Simple Homogeneous Process*

In the first case, let Y_t be a simple homogeneous process with the MGF

$$M_t(\theta) = \exp\left\{\gamma(t)\int_0^\infty (e^{-\theta z}-1)e^{-\tau z}(1-e^{-z})^{-1}\,dz\right\}, \tag{7.3.11}$$

where γ is assumed to be continuous. Then, with the above data scheme and notations, it is shown that

$$\mathcal{E}(S(t)|\text{data}) = e^{-\gamma(t)/(h_{j(t)}+\tau)} \times \prod_{i=1}^{j(t)}\left[\exp\left\{\gamma(u_i)\frac{h_{i-1}-h_i}{(h_{i-1}+\tau)(h_i+\tau)}\right\}\right]\cdot\frac{(h_i+\lambda_i+\tau)}{(h_i+\lambda_i+\delta_i+\tau)}, \tag{7.3.12}$$

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

$$\mathcal{E}(S(t)|\text{data}) = S_0(t)^{\tau/(h_{j(t)}+\tau)} \times\prod_{i=1}^{j(t)}\left[S_0(t)^{-\tau\frac{h_{i-1}-h_i}{(h_{i-1}+\tau)(h_i+\tau)}}\right]\cdot\frac{(h_i+\lambda_i+\tau)}{(h_i+\lambda_i+\delta_i+\tau)}. \tag{7.3.13}$$

Further if $S_0(t) > 0$ for all t, we have as $\tau \to 0$,

$$\mathcal{E}(S(t)|\text{data}) \to \begin{cases} \prod_{i=1}^{j(t)}\frac{h_i+\lambda_i}{h_i+\lambda_i+\delta_i} & \text{for } t < u_k \\ \frac{S_0(t)}{S_0(u_k)}\prod_{i=1}^{k}\frac{h_i+\lambda_i}{h_i+\lambda_i+\delta_i} & \text{for } t \geq u_k \end{cases}, \tag{7.3.14}$$

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.

7.3.4 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 simpler form

$$M_t(\theta) = \left(\frac{\tau}{\tau+\theta}\right)^{\gamma(t)} = \exp\left\{\gamma(t)\int_0^\infty (e^{-\theta z}-1)e^{-\tau z}z^{-1}dN(z)\right\}. \tag{7.3.15}$$

The posterior mean of the survival function given the above scheme of data turns out to be

$$\begin{aligned}\mathcal{E}(S(t)|\text{data}) = M_t(1|\text{data}) &= \left(\frac{h_j(t)+\tau}{h_j(t)+\tau+1}\right)^{\gamma(t)} \\ &\times\prod_{i=1}^{j(t)}\left[\left(\frac{(h_{i-1}(t)+\tau)(h_i(t)+\tau+1)}{(h_{i-1}(t)+\tau+1)(h_i(t)+\tau)}\right)^{\gamma(u_i)}\right. \\ &\left.\times\frac{\phi_G(h_i+\lambda_i+\tau+1,\delta_i)}{\phi_G(h_i+\lambda_i+\tau,\delta_i)}\right],\end{aligned} \tag{7.3.16}$$

where

$$\phi_G(\alpha,\beta) = \sum_{i=0}^{\beta-1}\binom{\beta-1}{i}(-1)^i\log\left(\frac{\alpha+i+1}{\alpha+i}\right). \tag{7.3.17}$$

Note that $\mathcal{E}(S(t)) = M_t(1) = \left(\frac{\tau}{\tau+1}\right)^{\gamma(t)}$. 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 $(n+1)$-the 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 $(n+1)$-the stage turns out to be the above estimator in which γ is replaced by

its estimator $\hat{\gamma}$ and can be simplified as

$$\widehat{S}_{n+1}(u) = \begin{cases} \left(\frac{\tau+1}{\tau+2}\right)^{\hat{\gamma}(u)} & \text{if} \quad u < Z_{n+1} \\ \left(\frac{\tau}{\tau+1}\right)^{\hat{\gamma}(u)} \left(\frac{(\tau+1)^2}{\tau(\tau+2)}\right)^{\hat{\gamma}(Z_{n+1})} & \text{if} \quad u \geq Z_{n+1}, \delta_{n+1} = 0 \\ \kappa\left(\frac{\tau}{\tau+1}\right)^{\hat{\gamma}(u)} \left(\frac{(\tau+1)^2}{\tau(\tau+2)}\right)^{\hat{\gamma}(Z_{n+1})} & \text{if} \quad u \geq Z_{n+1}, \delta_{n+1} = 1\,, \end{cases} \tag{7.3.18}$$

where $\kappa = \ln\left(\frac{\tau+2}{\tau+1}\right) \diagup \ln\left(\frac{\tau+1}{\tau}\right)$ and $\hat{\gamma}(u)$ is such that

$$\left(\frac{\tau}{\tau+1}\right)^{\hat{\gamma}(u)} = \frac{1+N^+(u)}{1+n} \prod_{j=1}^{n} \left(\frac{N^+(Z_j)+2}{N^+(Z_j)+1}\right)^{\left[\delta_j=0,\, Z_j \leq u\right]}. \tag{7.3.19}$$

Ghorai showed that the sequence of empirical Bayes estimators is asymptotically optimal with rate of convergence $O(n^{-1})$.

7.3.5 *Estimation Based on Beta Process*

As noted in Sect. 4.5, the cumulative hazard function $A(t) = \int_{[0,t]} dF(s)/F[s,\infty)$ is related to the distribution function via the correspondence $F(t) = 1 - \prod_{[0,t]} \{1 - dA(s)\}$, where Π is the product integral. Suppose A has a beta process prior with parameters $c(\cdot)$ and A_0 as defined in Sect. 4.5. 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 A 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.

$$\hat{S}(t) = \mathcal{E}(S(t)\,|\text{data}) = \prod_{[0,t]} \left\{1 - \frac{c(s)\,dA_0(s) + dN(s)}{c(s) + M(s)}\right\}, \tag{7.3.20}$$

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 $c(\cdot) \to 0$, $\hat{S}(t)$ 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, Muliere and Walker (1997) obtain the same estimator as shown below.

7.3.6 Estimation Based on Beta-Stacy Process

Assume a beta-Stacy prior (see Sect. 4.7) 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 Muliere and Walker (1997)

$$\hat{S}(t) = \mathcal{E}(S(t)|\text{data}) = \prod_{[0,t]} \left\{ 1 - \frac{c(s)\,dG(s) + dN(s)}{c(s)\,G[s,\infty) + M(s)} \right\}, \tag{7.3.21}$$

where $N(\cdot)$ and $M(t)$ are as defined above. The PL estimator is obtained if $c(\cdot) \to 0$.

7.3.7 Estimation Based on Polya Tree Priors

Muliere and Walker (1997) present the estimation of a survival curve using the posterior predictive distribution of a future observation. Assume that Π and $\mathcal{A}$, as described in Sect. 5.2 are given and $F \sim \mathcal{PT}(\Pi, \mathcal{A})$. Let $\theta_1, \theta_2, ..|F \overset{iid}{\sim} F$. The posterior predictive distribution based on exact observations was given earlier [expression (5.2.4)] as

$$\mathcal{P}\left[\theta_{n+1} \in B_{\underline{\epsilon}_m}|\text{data}\right] = \frac{\alpha_{\epsilon_1} + n_{\epsilon_1}}{\alpha_0 + \alpha_1 + n} \cdots \frac{\alpha_{\underline{\epsilon}_m} + n_{\underline{\epsilon}_m}}{\alpha_{\underline{\epsilon}_{m-1}0} + \alpha_{\underline{\epsilon}_{m-1}1} + n_{\underline{\epsilon}_{m-1}}}, \tag{7.3.22}$$

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, it is shown that the posterior predictive distribution of a future observation is given by

$$\mathcal{P}\left[\theta_{n+1} \in B_{\underline{\epsilon}_m}|\text{data}\right] = \frac{\alpha_{\epsilon_1} + n_{\epsilon_1}}{\alpha_0 + \alpha_1 + n} \cdots \frac{\alpha_{\underline{\epsilon}_m} + n_{\underline{\epsilon}_m}}{\alpha_{\underline{\epsilon}_{m-1}0} + \alpha_{\underline{\epsilon}_{m-1}1} + n_{\underline{\epsilon}_{m-1}} - \lambda_{\underline{\epsilon}_{m-1}}}, \tag{7.3.23}$$

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, $t_1 < t_2 < \cdots < t_k$, 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 remains 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.

7.3.8 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. 4.4), and $S(x) = \exp\{-\int_{[0,x)} r(t)dt\}$, then Dykstra and Laud (1981) has shown that $\hat{S}_{\alpha\beta}(t)$ 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.

$$\hat{S}_{\alpha\beta}(t) = \exp\left\{-\int_{[0,\infty)} \log(1 + \beta^*(s)(t-s)^+) d\alpha(s)\right\} \cdot \phi(\beta^{**})/\phi(\beta^*), \tag{7.3.24}$$

where

$$\phi(\beta) = \int_{[0,x_n)} \cdots \int_{[0,x_1)} \prod_{i=1}^{n} \beta(z_i) \prod_{i=1}^{n} d\left[\alpha + \sum_{j=i+1}^{n} I_{(x_j,\infty)}\right](z_i), \tag{7.3.25}$$

with $\beta^{**}(s) = \frac{\beta^*(s)}{\left[1+\beta^*(s)(t-s)^+\right]}$ and $\beta^*(t) = \beta(t) / \left[1 + \beta(t) \cdot \sum_{i=1}^{m} (x_i - t)^+\right]$.

7.3.9 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 straightforward, the expressions do not turn out to be simple. See their paper for details.

7.4 Linear Bayes Estimation of an SF

In Bayesian estimation of a survival curve S, Zehnwirth (1985) takes a different approach. In assuming the Dirichlet process or a neutral to the right process as prior, it is tacitly assumed that the hazard contributions of nonoverlapping 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, tractability and simplicity are gained but his estimator turns out to be only an approximate Bayes estimator of the survival curve. Moreover, he takes the loss function as pointwise 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 point b given that the object has survived up to point 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

$$S(u) = p(u|Z_{(l)}) \prod_{j=0}^{l-1} p(Z_{(j+1)}|Z_{(j)}). \tag{7.4.1}$$

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

$$\begin{aligned} R(S, \widehat{S}) =& \mathcal{E}\left[(a_{u,l} + b_{u,l}\frac{N^+(u)}{N(l) - \lambda_l} - p(u|Z_{(l)}))^2\right] \\ &+ \sum_{j=0}^{l-1} \mathcal{E}\left[(a_j + b_j \frac{N(j+1)}{N(j) - \lambda_j} - p(Z_{(j+1)}|Z_{(j)}))^2\right], \end{aligned} \tag{7.4.2}$$

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 $Z_{(l)} \le u \le Z_{(l+1)}$, $l = 0, 1, \ldots, m$, is

$$\widehat{S}(u) = \left\{ \frac{N^+(u) + f(u|Z_{(l)})g(u|Z_{(l)},)}{N(l) - \lambda_l + f(u|Z_{(l)})} \right\} \cdot \prod_{j=0}^{l-1} \left(\frac{N(j+1) + f(Z_{(j+1)}|Z_{(j)})g(Z_{(j+1)}|Z_{(j)})}{N(j) - \lambda_j + f(Z_{(j+1)}|Z_{(j)})} \right) \tag{7.4.3}$$

where

$$f(b|a) = \frac{\mathcal{E}[p(b|a)(1-p(b|a)]}{Var[p(b|a)]} \quad \text{and} \quad g(b|a) = \mathcal{E}[p(b|a)]. \tag{7.4.4}$$

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 satisfies $p(b|0) = p(a|0)p(b|a)$. 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

$$\begin{aligned} g(b|a) &= S_1(b^-)/S_1(a^-) \qquad \text{and} \\ f(b|a) &= \frac{S_1(b^-)/S_1(a^-) - S_2(b^-)/S_2(a^-)}{S_2(b^-)/S_2(a^-) - S_1^2(b^-)/S_1^2(a^-)}, \end{aligned} \tag{7.4.5}$$

where S_1 and S_2 are the first and second moments of S, $S_1(u) = \mathcal{E}(S(u))$, and $S_2(u) = \mathcal{E}\left(S^2(u)\right)$.

Now substituting these quantities in (7.4.3), the linear Bayes estimator for $S(u)$ reduces to

$$\widehat{S}(u) = \left\{ \frac{N^+(u) + f(u|Z_{(l)})S_1(u^-)/S_1(z_{(l)}^-)}{N(l) - \lambda_l + f(u|Z_{(l)})} \right\} \cdot \prod_{j=0}^{l-1} \left(\frac{N(j+1) + f(z_{(j+1)}|z_{(j)})S_1(z_{(j+1)}^-)/S_1(z_{(j)}^-)}{N(j) - \lambda_j + f(z_{(j+1)}|z_{(j)})} \right). \tag{7.4.6}$$

On the other hand, if $F \sim \mathcal{D}(\alpha)$, then $S(u^-)$ has a $\text{Be}(\alpha[u,\infty), \alpha[0,u))$ distribution and therefore,

$$S_1(u^-) = \frac{\alpha[u,\infty)}{\alpha(R^+)} \quad \text{and} \quad S_2(u^-) = \frac{\alpha[u,\infty)(\alpha[u,\infty)+1)}{\alpha(R^+)(\alpha(R^+)+1)}. \tag{7.4.7}$$

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.

7.5 Other Estimation Problems

In this section, we describe Bayesian solutions to some other estimation problems that have appeared in the literature.

7.5.1 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 estimating the probability $\Delta(F)$ given by

$$\Delta(F) = P(Z > X + Y) = \int_0^\infty \int_0^\infty S(x+y)dS(x)dS(y). \tag{7.5.1}$$

Assume $F \sim \mathcal{D}(\alpha)$ 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

$$\hat{\Delta}(S) = \frac{(M+n)^2}{(M+n)^{(3)}}\left[-\int_0^\infty \hat{S}_\alpha(2y)d\hat{S}_\alpha(y) + (M+n)\Delta(\hat{S}_\alpha)\right], \tag{7.5.2}$$

where $\hat{S}_\alpha$ is the Bayes estimator of S with respect to the Dirichlet process prior. When $M \to 0$, $\hat{S}_\alpha \to \hat{S}_{\text{PL}}$, the PL estimator of the survival function, therefore the estimator reduces to

$$\hat{\Delta}(S) = \frac{n^2}{n^{(3)}}\left[-\int_0^\infty \hat{S}_{\text{PL}}(2y)d\hat{S}_{\text{PL}}(y) + n\Delta(\hat{S}_{\text{PL}})\right]. \tag{7.5.3}$$

The empirical Bayes estimator is also derived using the procedure discussed in Sect. 6.2.4.

7.5.2 *Estimation of P(X ≤ Y)*

The Bayesian estimator of $\Delta = P(X \leq Y) = \int F dG$ 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 $F \sim \mathcal{D}(\alpha_1)$ and independently, $G \sim \mathcal{D}(\alpha_2)$. Based on samples $\mathbf{X}$ and $\mathbf{Y}$, he obtained the estimator as $\hat{\Delta} = \int \widehat{F}_{\alpha_1} d\widehat{G}_{\alpha_2}$, 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.

Its 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 $\hat{\Delta} = \int \mathcal{E}(G) d\mathcal{E}(F) = -\int \widehat{G}_{\alpha_2} d\widehat{F}_{\alpha_1}$ 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, respectively, 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): $\hat{\Delta}_{n+1} = -\int_{-\infty}^{M_n} \widehat{G}_{\hat{\alpha}_2} d\widehat{F}_{\hat{\alpha}_1}$ where $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are given by

$$\hat{\alpha}_1(u) = \frac{N_1^+(u)}{n} \prod_{i=1}^{n} \left[\frac{N_1^+(S_i) + 2}{N_1^+(S_i) + 1} \right]^{[\delta_i = 0, S_i \leq u]} \tag{7.5.4}$$

and

$$\hat{\alpha}_2(u) = \frac{N_2^+(u)}{n} \prod_{i=1}^{n} \left[\frac{N_2^+(T_i) + 2}{N_2^+(T_i) + 1} \right]^{[\eta_i = 0, T_i \leq u]}, \tag{7.5.5}$$

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 $\hat{\alpha}_1$ and $\hat{\alpha}_2$ 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 $0\left(n^{-1}\right)$ is derived.

7.5.3 Estimation of S in Competing Risk Models

Consider the situation in which there are two competing causes of death labeled 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 $\left(R_+^2, \mathcal{B}_+^2\right)$, 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) treated 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 $\delta_i = I[X_i \leq Y_i]$, $i = 1, 2, \ldots, n$. Let P be a Dirichlet process $\mathcal{D}(\alpha)$ defined on R_+^2 with parameter α, where α is a nonnegative finite measure on $\left(R_+^2, \mathcal{B}_+^2\right)$. The loss function used is $\int_{R_+^2} (S - \hat{S})^2 dW$, 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 $0 < z_1 < \cdots < z_k < \infty$. Further let λ_i and μ_i be the number of censored and uncensored observations at z_i, i.e., $\lambda_i = \#\left\{j | \min(X_j, Y_j) = z_i \text{ and } X_j > Y_j\right\}$, $\mu_i = \#\left\{j | \min(X_j, Y_j) = z_i \text{ and } X_j \leq Y_j\right\}$. Then the Bayes estimator of $S(s, t)$ is obtained as

$$\begin{aligned} \hat{S}(s, t) &= \mathcal{E}[S(s, t) \mid (Z, \boldsymbol{\delta})] \\ &= \frac{1}{\alpha\left(R_+^2\right) + n} \left\{\alpha((s, \infty) \times (t, \infty)) + N^+(\max(s, t)) + \sum_r \theta_r\right\}, \end{aligned} \tag{7.5.6}$$

where the summation is taken over all r such that $\min(s,t) < z_r < \max(s,t)$,

$$N^{+}(u) = \sum_{\{i:z_i>u\}} (\lambda_i + \mu_i) = \#\text{ of observations } > u, \tag{7.5.7}$$

$$\theta_r = \begin{cases} \lambda_r \alpha_s'(Z_r) & \text{if } s > t \\ \mu_r \alpha_t'(Z_r) & \text{if } s < t \\ 0 & \text{if } s = t \end{cases}, \tag{7.5.8}$$

and

$$\begin{aligned} \alpha_s'(Z_r) &= \lim_{\epsilon \searrow 0} \frac{\alpha(\{X > s, Z_r - \epsilon < Y \le Z_r\})}{\alpha(\{X > Y, Z_r - \epsilon < Y \le Z_r\})} \\ \alpha_t'(Z_r) &= \lim_{\epsilon \searrow 0} \frac{\alpha(\{Y > t, Z_r - \epsilon < X \le Z_r\})}{\alpha(\{X \le Y, Z_r - \epsilon < X \le Z_r\})}, \end{aligned} \tag{7.5.9}$$

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 survival function $S(s, 0)$.

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,

$$\alpha'(x,y) = \begin{cases} \beta(\beta+\gamma) e^{(\beta+\gamma)y} & \text{if } 0 \le x < y \\ \gamma(\beta+\gamma) e^{(\beta+\gamma)x} & \text{if } 0 \le y < x \end{cases}, \tag{7.5.10}$$

which is a special case of Freund's bivariate exponential distribution (Johnson and Kotz 1970), for $\beta, \gamma > 0$. Then by straightforward substitution and simplification, the Bayes estimator for $s > t$ is obtained as

$$\hat{S}(s,t) = \frac{1}{n+1} \left\{ e^{(\beta+\gamma)s} [1 + \gamma(s-t)] + N^{+}(s) + \sum_{t < z_r < s} \lambda_r e^{(\beta+\gamma)(z_r - s)} \right\}. \tag{7.5.11}$$

A similar expression can be obtained for $s < t$. For $t = s$, $\hat{S}(s,\infty) = (n+1)^{-1} \{\alpha((s,\infty) \times (s,\infty)) + N^{+}(s)\}$.

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 employ the special feature of the identified minima in updating the parameter α. They introduce a new random variable ξ having a distribution

$$\mathcal{P}\{\xi \in B\} = \mathcal{P}\{U \in B|Z\}, \tag{7.5.12}$$

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 $P \sim \mathcal{D}(\alpha)$, 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 two 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 caused by using Peterson (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. 7.3.1). The approach of Salinas-Torres et al. 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 $S_j(t) = P(X_j > t), j = 1, .., k$. 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, $j = 1, ..., k$. Then if $\delta = j$, $Z = X_j$ and $S(t) = P(Z > t) = \sum_{j=1}^k S_j^*(t)$. Let Δ be a nonempty 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

$$S_\Delta(t) = \varphi\left(S_\Delta^*(\cdot), S_{\Delta^c}^*(\cdot); t\right), \quad \text{for} \quad t \le \min\left(t_{S_\Delta}, t_{S_{\Delta^c}}\right), \tag{7.5.13}$$

where

$$\varphi(F(\cdot), G(\cdot); t) = \exp\left\{\oint_0^t \frac{dF(s)}{F(s) + G(s)}\right\} \prod_t \frac{F(s_+) + G(s_+)}{F(s_-) + G(s_-)}, \tag{7.5.14}$$

$t_{S_\Delta} = \sup\{t : S_\Delta(t) > 0\}$, 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 $\{s \leq t : s \text{ is a jump point of } F\}$. Let the loss function be

$$L\left(\mathbf{S}^*, \widehat{\mathbf{S}}^*\right) = \int_0^\infty \| \mathbf{S}^* - \widehat{\mathbf{S}}^* \|^2 \, dW(t), \tag{7.5.15}$$

where $\|\cdot\|$ stands for the usual norm, $\mathbf{S}^* = \left(S_1^*, \ldots, S_k^*\right)$, and $\widehat{\mathbf{S}}^* = \left(\widehat{S}_1^*, \ldots, \widehat{S}_k^*\right)$ 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. prove 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 $t \leq Z_{(m)}$,

$$\widehat{S}_\Delta(t) = \varphi\left(\widehat{S}_\Delta^*, \widehat{S}_{\Delta^c}^*; t\right) = \widehat{S}(t) \pi_k(t) \exp\left\{\frac{-1}{\alpha(R) + n}\right\} \sum_{j \in \Delta^c} \int_0^t \frac{d\alpha_j(s, \infty)}{\widehat{S}(s)} \tag{7.5.16}$$

is the Bayes estimator of $S_\Delta(t)$ under the above loss function, where $n_j = \sum_{i=1}^n I[Z_i \geq Z_{(j)}]$, $d_j = \sum_{i=1}^n I[Z_i = Z_{(j)}, \delta_i = 1], j = 1, \ldots, m$ and

$$\pi_k(t) = \prod_{i: Z_{(i)} \leq t} \frac{\sum_{j=1}^k \alpha_j\left(Z_{(i)}, \infty\right) + n_i - d_i}{\sum_{j=1}^k \alpha_j\left(Z_{(i)}, \infty\right) + n_i}. \tag{7.5.17}$$

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) equation (3.2). It is also similar to the Susarla–VanRyzin estimator derived in Sect. 7.1. Likewise, if $\alpha_j(s, \infty) \to 0$, for $j = 1, 2$, the estimator reduces to the PL estimator.

7.5.4 *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 a cumulative hazard rate obtained in Hjort (1990) by assuming a beta process prior of Sect. 4.5. 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

$$N(j) = \sum_{i=1}^{n} I\left[Z_i \leq j \text{ and } \delta_i = 1\right]; \text{ and } M(j) = \sum_{i=1}^{n} I[Z_i \geq j], \ j \geq 0. \tag{7.5.18}$$

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 $\text{Be}\{c(j)h_0(j), c(j)(1 - h_0(j))\}$. 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 $Var(h(j)) = h_0(j)(1 - h_0(j))/[c(j) + 1]$ as the prior "uncertainty." The posterior distribution of H (discrete time version of property 4 of Sect. 4.5) is

$$H|\text{data} \sim \mathcal{B}e\left\{c + M, \sum \frac{cdH_0 + dN}{c + M}\right\}, \tag{7.5.19}$$

and the nonparametric Bayesian estimator of H under the weighted quadratic loss is given by

$$\widehat{H}(j) = \mathcal{E}(H|\text{data}) = \sum_{i=0}^{n} \frac{c(i) h_0(i) + dN(i)}{c(i) + M(i)}. \tag{7.5.20}$$

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.

7.5.5 *Estimation of Hazard Rates*

Assume that the hazard rate $r(t)$ has an extended gamma process prior (see Sect. 4.4) $\Gamma(\alpha(\cdot), \beta(\cdot))$. Given a sample of n observations, the posterior distributions of $r(t)$ was given in expression (4.4.7) for censored and in expression (4.4.8) for exact observations.

The posterior mean of $r(t)$ based on m observations (censored and exact) is

$$\hat{r}(t) = \frac{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \int_{[0,t)} \prod_{i=0}^{m} \beta^*(z_i) \prod_{i=0}^{m} d\left[\alpha + \sum_{j=i+1}^{m} I_{(x_j,\infty)}\right](z_i)}{\int_{[0,x_m)} \cdots \int_{[0,x_1)} \prod_{i=1}^{m} \beta^*(z_i) \prod_{i=1}^{m} d\left[\alpha + \sum_{j=i+1}^{m} I_{(x_j,\infty)}\right](z_i)}. \tag{7.5.21}$$

Clearly, $\hat{r}(t)$ is the Bayes estimator under the loss function $\int_{[0,\infty)} (r(t) - \hat{r}(t))^2 dW(t)$ subject to the condition

$$\int_{[0,\infty)} \int_{[0,t)} \beta^2(s) d\alpha(s) dW(t) < \infty. \tag{7.5.22}$$

For the computational purposes, they show that the multi-dimensional integrals in $\hat{r}(t)$ 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. 7.3.8.

7.5.6 *Markov Chain Application*

Hjort (1990) extends the results of Sect. 7.5.4 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,

$$p_{ij}(r, s) = \mathcal{P}\{X(s) = j | X(r) = i\}, \quad 0 \le r \le s, \quad i, j = 1, \ldots, k. \tag{7.5.23}$$

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 $H_{ij}(s) = \sum_{r=1}^{s} h_{ij}(r)$, $s \ge 1$. 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. $X^{(l)}_{t(l)} = \left\{X^{(l)}(r) : r = 0, 1, \ldots, t(l)\right\}$, $l = 1, 2, \ldots, n$,

$$dN_{ij}(r) = \sum_{l=1}^{n} I[X^{(l)}(r-1) = i,\ X^{(l)}(r) = j], \tag{7.5.24}$$

and

$$M_i(r) = \sum_{l=1}^{n} I[X^{(l)}(r-1) = i,\ r \le t(l)],\ r \ge 1. \tag{7.5.25}$$

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 $(0, s]$.

Assume 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. Then, $\mathcal{E}(h_{ij}(r)) = h_{0ij}(r)$ 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

$$\hat{h}_{ij}(r) = \mathcal{E}\left(h_{ij}(r) \,|\text{data}\right) = \frac{c_i(r)h_{0i1}(r) + dN_{ij}(r)}{c_i(r) + M_i(r)}, \tag{7.5.26}$$

and for $H_{ij}(s)$ is

$$\widehat{H}_{ij}(s) = \sum_{r=1}^{s} \frac{c_i(r)h_{0i1}(r) + dN_{ij}(r)}{c_i(r) + M_i(r)},\ \ s \ge 1. \tag{7.5.27}$$

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

$$\widehat{F}_i(s) = 1 - \prod_{r=1}^{s}\left[1 - \frac{c_i(r)dH_{0,i\cdot}(r) + dN_{i\cdot}(r)}{c_i(r) + M_i(r)}\right]. \tag{7.5.28}$$

Similar analysis in the time-continuous case leads to the estimator

$$\widehat{H}_{ij}(t) = \int_0^t \frac{c_{ij}(s)dH_{0,ij}(s) + dN_{ij}(s)}{c_{ij}(s) + M_i(s)}. \tag{7.5.29}$$

The estimate of waiting time distribution

$$G_i([s,t]) = P\{X(u) \equiv i, u \in [s,t] \,|X(s) = i\} = \prod_{[s,t]} \{1 - dH_{i\cdot}(u)\} \text{ for } \le t, \tag{7.5.30}$$

is given by

$$\widehat{G}_i([s,t]) = \prod_{[s,t]} \left[1 - \frac{c_i dH_{0,i\cdot} + dN_{i\cdot}}{c_i + M_i} \right], \tag{7.5.31}$$

where $N_{i\cdot} = \sum_{j \neq i} N_{ij}$ and $H_{0,i\cdot} = \sum_{j \neq i} H_{0,ij}$.

7.5.7 *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 obtains the Bayes estimator for $S(t)$ assuming the pair (λ, F) to be independent random variables and placing a gamma $\mathcal{G}(\nu, \theta)$ and a Dirichlet process $\mathcal{D}(\alpha)$ priors on λ and F, respectively. He also 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,

$$(\lambda, F)|\underline{N}_T, \mathbf{X}_T \sim \mathcal{G}(\nu + N(T), \theta + T) \times \mathcal{D}\left(\alpha + \sum_{i=1}^{N(T)} \delta_{X_i}\right). \tag{7.5.32}$$

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 $L(S, \hat{S}) = \int_R (S(t) - \hat{S}(t))^2 dW(t)$ is obtained as

$$\hat{S}(t) = \mathcal{E}[S(t)|\underline{N}_T, \mathbf{X}_T] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathcal{E}\left[\overline{P}_k|\underline{N}_T, \mathbf{X}_T\right] \mathcal{E}[e^{-\lambda t}\lambda^k|\underline{N}_T, \mathbf{X}_T]. \tag{7.5.33}$$

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

$$\hat{S}(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{\Gamma(\nu + N(T) + k)}{\Gamma(\nu + N(T))} \left(\frac{1}{\theta + T + t}\right)^k \left(\frac{\theta + T}{\theta + T + t}\right)^{\nu + N(T)} \cdot \mathcal{E}[\overline{P}_k | \underline{N}_T, \mathbf{X}_T]. \tag{7.5.34}$$

This expression is evaluated 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.

Lo (1982) also 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) on the other hand 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 derives 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 semi-martingale approach instead of the Lévy measure approach for the independent increment processes used.

7.5.8 *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 $\mathbb{N} = \{0, 1, 2, \ldots\}$ and the loss function assumed is,

$$L((P,G),(\widehat{P},\widehat{G})) = a_1 \sum_{j=0}^{\infty} W_1(j)(p_j - \hat{p}_j)^2 + a_2 \int_0^{\infty} (G(t) - \widehat{G}(t))^2 dW_2(t), \tag{7.5.35}$$

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 $[0,T]$, $l = 0, 1, 2, \ldots$, $D_{t_1}, \ldots, D_{t_n}$ age at death of n individuals who died in $[0,T]$, and $S_{t_1}, \ldots, S_{t_m}$ 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 one would expect, they have similar expression as the Bayes estimator of survival function obtained in Sect. 7.1.

1. The Bayes estimator of P under the above loss function is given by $\widehat{P} = \{\hat{p}_j\}$, where

$$\hat{p}_j = \frac{\alpha_1(j) + N_l(T) I[j = l]}{\alpha_1(\mathbb{N}) + \sum_{l=0}^{\infty} N_l(T)}. \tag{7.5.36}$$

2. The conditional distribution of $P|N_l(T), l \in \mathbb{N}$ is $\mathcal{D}\left(\alpha_1 + \sum_{l=0}^{\infty} N_l(T)\,\delta_l\right)$.
3. Assuming α_2 to be nonatomic, the Bayes estimator of G under the above loss function is given by $\widehat{G}$, where

$$1 - \widehat{G}(t) = \frac{\alpha_2(t,\infty) + N^+(t)}{\alpha_2(R^+) + m + n} \prod_{j=1}^{k} \left(\frac{\alpha_2(S^*_{t_j},\infty) + N(S^*_{t_j})}{\alpha_2(S^*_{t_j},\infty) + N(S^*_{t_j}) - \lambda^*_j} \right)^{I\left[S_{t_j} \leq t\right]}, \tag{7.5.37}$$

where $N^+(t) = \#$ of deaths and survival times greater than t,$N(t) = \#$ of deaths and survival times greater than or equal to t, $S^*_{t_1}, \ldots, S^*_{t_k}$ are the k distinct observations among $S_{t_1}, \ldots, S_{t_m}$, and λ^*_j are their multiplicities.

Remark 7.4 As $\alpha_1(\mathbb{N}) \to 0$, $\hat{p}_j \to$ MLE of p_j. The Bayes estimator under the squared error loss of the mean M of the offspring distribution P is given by

$$\widehat{M} = \frac{\sum_{l=0}^{\infty} l\alpha_1(\{l\}) + \sum_{l=0}^{\infty} l N_l(T)}{\alpha_1(\mathbb{N}) + \sum_{l=0}^{\infty} N_l(T)}. \tag{7.5.38}$$

Remark 7.5 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 a known constant and fixed 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 $D_{t_1}, \ldots, D_{t_n}$.

The authors 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 $\hat{p}_j$ in this case not only depends on the splits n_x but also on the life-lengths and survival times.

7.6 Hypothesis Testing $H_0 : F \leq G$

Earlier in Sect. 6.9.1, hypothesis testing relative to the null hypothesis $H_0 : F \leq F_0$ against the alternative $H_1 : F \not\leq F_0$ 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 statistic. 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. 7.5.2), we want to test the hypothesis $H_0 : F \leq G$ against $H_1 : F \not\leq G$ 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. 6.9). $L((F, G), a_0) = \int (F - G)^+ dW$ and $L((F, G), a_1) = \int (F - G)^- dW$, 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 $\xi_n = P\{$ taking action $a_0 \mid$ data $\}$. Then the Bayes rule with respect to these priors is given by the test statistic

$$\xi_n = I[\psi\ (\alpha_1, \alpha_2) \leq 0], \tag{7.6.1}$$

where

$$\begin{aligned} \psi\ (\alpha_1, \alpha_2) &= \mathcal{E}\ [L((F, G), a_0) - L((F, G), a_1) \mid \text{data}] \\ &= \int (\widehat{F}_{\alpha_1}(u) - \widehat{G}_{\alpha_2}(u)) dW(u), \end{aligned} \tag{7.6.2}$$

and $\widehat{F}_{\alpha_1}$ and $\widehat{G}_{\alpha_2}$ are the usual Bayes estimators of F and G derived earlier. The minimum Bayes risk against the Dirichlet process priors is

$$r_n^*(\alpha_1, \alpha_2) = \inf_{\xi_n}(\alpha_1, \alpha_2, \xi_n) = \mathcal{E}[I[\psi\ (\alpha_1, \alpha_2) \leq 0]\psi\ (\alpha_1, \alpha_2)] + \mathcal{E}\ [L((F, G), a_1)], \tag{7.6.3}$$

which can easily be evaluated. When α_1 and α_2 are unknown, the empirical Bayes method of earlier sections was recommended. In that case the test statistics ξ_n is replaced at the $(n+1)$-stage by $\hat{\xi}_{n+1}$ with ψ replaced by $\hat{\psi}_{n+1}$ given by $\hat{\psi}_{n+1}(\cdot) = \int(\widehat{F}_{\hat{\alpha}_1}(u) - \widehat{G}_{\hat{\alpha}_2}(u))dW(u)$, where as before, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ given in (7.5.4) may be used. It is shown that the above test statistic is asymptotically optimal with the rate of convergence $0(n^{\frac{1}{2}})$.

Damien and Walker (2002) present a different approach for comparing two treatment effects in which the Bayes Factor is used.

7.7 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. 4, Kalbfleisch considered these problems in a Bayesian framework. 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 $Var(H(t)) = H_0(t)/c$. 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 $\mathcal{E}\left(e^{-\Lambda(t)}|\text{observations}\right)$ 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 nonrandom, 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 7.6 (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 $s < t \leq x$ 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 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 7.7 (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*

$$M_t(\theta \mid \text{data}) = \frac{M_t(\theta + h_{j(t)})}{M_t(h_{j(t)})} \cdot \prod_{i=1}^{j(t)} \left[\frac{M_{u_i}^-(\theta + h_{i-1})}{M_{u_i}^-(h_{i-1})} \cdot \frac{C_{u_i}(\theta + h_i + \lambda_i, d_i)}{C_{u_i}(h_i + \lambda_i, d_i)} \cdot \frac{M_{u_i}(h_i)}{M_{u_i}(\theta + h_i)} \right], \tag{7.7.1}$$

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

$$C_{u_i}(\alpha, d_i) = \int_0^\infty e^{-\alpha z} \prod_{l\in D(i)} \left(1 - \exp\left(-ze^{\boldsymbol{\beta}\mathbf{W}_l}\right)\right) d\, G_{u_i}(z), \tag{7.7.2}$$

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

$$C_{u_i}(\alpha, d_i) = \begin{cases} 1 & \text{if } d_i = 0 \\ \int_0^\infty e^{-\alpha z} d\, H_{u_i}(z) & \text{if } d_i = 1 \\ \int_0^\infty e^{-\alpha z} \prod \{1 - \exp\left(-ze^{\boldsymbol{\beta}\mathbf{W}_l}\right)\} d\, H_u(z) & \text{if } d_i > 1 \end{cases}, \tag{7.7.3}$$

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

$$\frac{C_{u_i}(\alpha+1,d_i)}{C_{u_i}(\alpha,d_i)}=\begin{cases}1 & \text{if } d_i=0\\ \log\frac{\alpha+1+\exp(\boldsymbol{\beta}\mathbf{W})}{\alpha+1}\Big/\log\frac{\alpha+\exp(\boldsymbol{\beta}\mathbf{W})}{\alpha} & \text{if } d_i=1\\ \frac{\int_0^\infty e^{-(\alpha+1)z}\prod_{l\in D(i)}\{1-\exp(-ze^{\boldsymbol{\beta}\mathbf{W}_l})\}z^{-1}dz}{\int_0^\infty e^{-\alpha z}\prod_{l\in D(i)}\{1-\exp(-ze^{\boldsymbol{\beta}\mathbf{W}_l})\}z^{-1}dz} & \text{if } d_i>0.\end{cases} \tag{7.7.4}$$

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.

$$\begin{aligned}E(1-F(t)|\text{data})=M_t(1|\text{data})&=\left(\frac{\mathrm{h_{j(t)}+fi}}{\mathrm{h_{j(t)}+fi+1}}\right)^{,(t)}\\ &\times\prod_{i=1}^{j(t)}\left[\left(\frac{(h_{i-1}+\tau)(h_i+\tau+1)}{(h_{i-1}+\tau+1)(h_i+\tau)}\right)^{\gamma(u_i)}\right.\\ &\left.\times\frac{C_{u_i}(h_i+\lambda_i+\tau+1,d_i)}{C_{u_i}(h_i+\lambda_i+\tau,d_i)}\right].\end{aligned} \tag{7.7.5}$$

The difference is in the quantities

$$h_i=\sum_{l\in R(i)}\exp(\boldsymbol{\beta}\mathbf{W}_l),\qquad \lambda_i=\sum_{l\in C(i)}\exp(\boldsymbol{\beta}\mathbf{W}_l), \tag{7.7.6}$$

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 $1 - dA_i(s) = \{1 - dA(s)\}^{\exp(\boldsymbol{\beta}\mathbf{w}_i)}$, where A is the cumulative hazard function and $dA_i(s)$ is the hazard function of the i-th individual. By assuming $\boldsymbol{\beta}$ known and $A \sim \mathcal{B}e\{c(\cdot), A_0(s)\}$, he shows that the posterior distribution of A 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 A 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 Bayesian nonparametric problems. *Annals of Statistics, 2*, 1152–1174.

Barlow, R. E., Barthalomew, 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, 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, 1*, 356–358.

Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.

Binder, D. A. (1982). Nonparametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society B, 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, 9* , 803–811.

Blei, D. M., & Frazier, P. I. (2011). Distant dependent Chinese restaurant processes. *Journal of Machine Learning Research, 12* , 2461–2488.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3* , 993–1022.

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.

Bondesson, L. (1982). On simulation from infinitely divisible distributions. *Advances in Applied Probability, 14*, 855–869.

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.

Broderick, T., Jordan, M. L., & Pitman, J. (2012). Beta processes, Stick-breaking and power laws. *Bayesian Analysis, 7*, 439–476.

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1

Broderick, T., Jordan, M. L., & Pitman, J. (2013). Cluster and feature modeling from combinatorial stochastic processes. *Statistical Science, 28*(3), 289–312.

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 B, 43*, 65–75.

Campbell, G., & Hollander, M. (1978). Rank order estimation with the Dirichlet prior. *Annals of Statistics, 6*(1), 142–153.

Caron, F., Davy, M., & Doucet, A. (2007). Generalized Polya urn for time-varying Dirichlet process mixtures. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (Vol. 23).

Christensen, R., Hanson, T., & Jara, A. (2008). Parametric nonparametric statistics: An introduction to mixtures of finite Polya trees. *Annals of Statistics, 62*, 296–306.

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. (1979). Considerazioni generali sull'impostazione bayesiana di problemi non parametrici, *Rivista di matematica per le Scienze Economicje e Socialli, 2*, Part I 39–52, Part II 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, 15*, 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.

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 & Methods, 12*(8), 947–963.

Damien, P., Laud, P. W., & Smith, A. F. M. (1995). Random variate generation form infinitely divisible distributions with applications to Bayesian inference. *Journal of the Royal Statistical Society B, 57*, 547–64.

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.

Damien, P., & Walker, S. (2002). A Bayesian nonparametric comparison of two treatments. *Scandinavian Journal of Statistics, 29*, 51–56.

DeIorio, M., Müller, P., Rosner, G .L., & MacEachern, S. N. (2004). An anova model for dependent random measures. *Journal of the American Statistical Association, 99*, 205–215.

Dey, J., Erickson, R. V., & Ramamoorthi, R. V. (2003). Some aspects of neutral to right priors. *International Statistical Review, 71*(2), 383–401.

Dey, D., Müller, P., & Sinha, D. (Eds.). (1998). *Practical nonparametric and semiparametric Bayesian statistics.* Lecture notes in statistics. New York: Springer.

Diaconis, P., & Freedman, D. A. (1986). On inconsistent of Bayes estimates of location. *Annals of Statistics, 14*, 68–87.

Doksum, K. A. (1972). Decision theory for some nonparametric models. *Proceedings of the Sixth Berkeley symposium on Mathematical Statistics and Probability, Vol. I: 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.

Dråghici, 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. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 2, pp. 183–214).

Dunson, D. B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics, 7*(4), 551–568.

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.

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 and Hall.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology, 3*, 87–112.

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.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 831–853).

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).

Favaro, S., & Teh, Y. W. (2013). MCMC for normalized random measure mixture models. *Statistical Science, 28*, 335–359.

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.

Ferguson, T. S., & Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Annals of Mathematical Statistics, 43* , 1634–1643.

Ferguson, T. S., & Phadia, E. G. (1 979). 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.). *Current issues in statistical inference: Essays in honor of D. Basu*. IMS lecture notes-monograph series (Vol. 17, pp. 127–150).

Foti, N. J., Futoma, J. D., Rockmore, D. N., & Williamson, S. (2012). A unifying representation for a class of dependent random measures. arXiv:1211.475v1[stat.ML]

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.). *Survival analysis*. IMS lecture notes-monograph series (Vol. 2, pp. 26–40).

Gardiner, J. C., & Susarla, V. (1983). Weak convergence of a Bayesian nonparametric estimator of the survival function under progressive censoring. *Statistics and Decision, 1*, 257–263.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika, 52*, 203–213.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.

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.

Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Bayesian nonparametric* .Springer series in statistics. 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 & Methods, 14*(9), 2081–2094.

Ghosh, M., Lahiri, P., & Tiwari, R. C. (1989). Nonparametric empirical Bayes estimation of the distribution and the mean. *Communications in Statistics - Theory & Methods , 18*(1), 121–146.

Ghorai, J. K. (1981). Empirical Bayes estimation of a distribution function with a gamma process prior. *Communications in Statistics - Theory & Methods, A10*(12), 1239–1248.

Ghorai, J. K. (1989). Nonparametric Bayesian estimation of a survival function under the proportional hazard model. *Communications in Statistics - Theory & Methods, A18*(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 (pp. 101–114)*. Dordrecht: D. Reidel Publishing Company.

Gnedin, A., & Pitman, J. (2007). Poisson representation of a Ewens fragmentation process. *Combinatorics, Probability and Computing, 16*, 819–827.

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). Allele frequencies in multidimensional Wright-Fisher models with a general symmetric mutation structure. *Theoretical Population Biology, 17*(1), 51–70.

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, MA: MIT.

Ghahramani, Z., Griffiths, T. L., & Sollich, P. (2007). Bayesian nonparametric latent feature models (with discussion and rejoinder). In J. M. Bernado, et al. (Eds.). *Bayesian statistics* (Vol. 8). Oxford, UK: Oxford University Press.

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.

Hannah, L. A., Blei, D. M., & Powell, W. B. (2011). Dirichlet process mixtures of general linear models. *Journal of Machine Learning Research, 12*, 1923–1953.

Hannum, R. C., & Hollander, M. (1983). Robustness of Ferguson's Bayes estimator of a distribution function. *Annals of Statistics, 11*, 632–639, 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.

Hansen, B., & Pitman, J. (2000). Prediction rules for exchangeable sequences related to species sampling. *Statistics & Probability Letters, 46*, 251–256.

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 statistcs in quality and reliability* (pp. 1385–1390). New York: Wiley.

Hanson, T. E., Branscum, A., & Gardner, I. (2008). Multivariate mixtures of Polya trees for modelling ROC data. *Statistical Modelling, 8*, 81–96.

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.

Hjort, N. L. (1985). *Bayesian Nonparametric Bootstrap Confidence Intervals*. NSF-and LCS-Technical Report, Department of Statistics, Stanford University.

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). *Bayesian nonparametrics*. Cambridge series in statistical and probabilistic mathematics. 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 & Methods, A5*(14), 1369–1383.

Hollander, M., & Korwar, R. M. (1982). Nonparametric Bayesian estimation of the horizontal distance between two populations. In *Nonparametric statistical inference* (Vol. 1). New York: North Holland.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for Stick-breaking priors. *Journal of the American Statistical Association, 96* , 161–173.

Ishwaran , H., & James, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica, 13*, 1211–1235.

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. (2003). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics, 30* , 269–283.

Ishwaran , H., & Zarepour, M. (2003). *Random probability measures via Polya sequences: Revisiting the Blackwell-MacQueen Urn scheme*. airXiv:Math/0309041v1.

Ibrahim, J. L., Chen, M., & Sinha, D. (2001). *Bayesian survival analysis* . New York: Springer.

Jain, S., & Neal, R. (2004). A split-merge Markov chain monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics, 13*, 158–182.

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* (Chap. 41, 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.

Jordan, M. I. (2010). Hierarchical models, nested models and completely random measures. In M.-H. Chen, D. Dey, P. Mueller, D. Sun, & K. Ye (Eds.), *Frontiers of statistical decision making and Bayesian analysis: In honor of James O. Berger*. New York: Springer.

Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival data. *Journal of the Royal Statistical Society B, 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 B, 75*, 1–22.

Kingman, J. F. C. (1993). *Poisson processes*. Oxford: Clarendon Press.

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 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.). *Bayesian nonparametrics.* Cambridge series in statistical and probabilistic mathematics (pp. 80–136).

Lin, D., Grimson, E., & Fisher, J. (2010). Construction of dependent Dirichlet processes based on Poisson processes. In *Neural Information Processing Systems*.

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. (1983). Weak convergence for Dirichlet processes. *Sankhya, 45*, 105–111.

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: American Statistical Association.

MacEachern, S. N., & Müller, P. (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics, 7*, 223–238.

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., & Petrone, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models. *Journal of the Italian Statistical Society, 2*, 349–364.

Muliere, P., & Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics, 26*, 283–297.

Muliere, P., & Walker, S. (1997). A Bayesian non-parametric approach to survival analysis using Polya trees. *Scandinavian Journal of Statistics, 24*, 331–340.

Müller, P., & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science, 19*, 95–110.

Müller, P., Quintana, F. A., Jara, A., & Hanson, T. (2015). *Bayesian nonparametric data analysis*. New York: Springer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics, 9* , 249–265.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics, 31* , 705–767.

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. (2003). Polya tree distributions for statistical modeling of censored data. *Journal of Applied Mathematics and Decision Sciences, 7*(3), 175–186.

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.

Nieto-Barajas, L. E., Prunster, I., & Walker, S. G. (2004). Normalized random measures driven by increasing additive processes. *Annals of Statistics, 32* , 2343–2360.

Nieto-Barajas, L. E., Müller, P., Ji, Y., Lu, Y., & Mills, G. B. (2012). A time-series DDP for functional proteomics profiles. *Biometrics, 68*, 859–868.

Ongaro, A., & Cattaneo, C. (2004). Discrete random probability measures: A general framework for nonparametric Bayesian inference. *Statistics & Probability Letters, 67*, 33–45.

Paddock, S., Ruggeri, F., Lavine, M., & West, M. (2003). Randomised Polya tree models for nonparametric Bayesian inference. *Statistica Sinica, 13*, 443–460.

Padgett, W. J., & Wei, L. J. (1981). A Bayesian nonparametric estimator of survival probability assuming increasing failure rate. *Communications in Statistics - Theory & Methods, A10*(1), 49–63.

Paisley, J., Blei, D. M., & Jordan, M. I. (2012). Stick-breaking beta processes and the Poisson process. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, La Palma, Canary Islands.

Paisley, J., Zaas, A., Woods, C. W., Ginsburg, G. S., & Carin, L. (2010). A stick-breaking construction of the beta process. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa.

Papaspiliopoulos, O., & Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika, 95* , 169–186.

Patil, G. P., & Taillie, C. (1977). Diversity as a concept and its implications for random communities. *Bulletin International Statistical Institute, 47*, 497–515.

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.

Perman, M., Pitman, J., & Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields, 92*, 21–39.

Petrone, S. (1999). Random Bernstein polynomials. *Scandinavian Journal of Statistics, 26*, 373–393.

Petrone, S., Guindani, M., & Gelfand, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society B, 71*, 75–782.

Phadia, E. G. (1971). *Minimax Estimation of a Cumulative Distribution Function*. Technical Report 71-1, Division of Statistics, The Ohio State University.

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 (2007) (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 & Methods, A8*(13), 1327–1351.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields, 102*, 145–158.

Pitman, J. (1996a). 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, CA: IMS.

Pitman, J. (1996b). Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability, 28*, 525–539.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability, 25*, 855–900.

Pruitt, R. C. (1992). An inconsistent Bayes estimate in bivariate survival curve analysis. *Statistics and Probability Letters, 15*(3), 177–180.

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.

Rao, V., & Teh, Y. W. (2009). Spatial normalized gamma processes. In *Neural Information Processing Systems, 2009*.

Regazzini, E., Lijoi, A. , & Prunster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics, 31*, 560–585.

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.

Ren, L., Dunson, D., & Carin, L. (2008). Dynamic hierarchical Dirichlet process. In *Proceedings of the International Conference on Machine Learning*, Helsinki.

Ren, L., Wang, Y., Dunson, D. , & Carin, L. (2011). The kernel Beta process. In *Neural Information Processing Systems, 2011*.

Rodriguez, A. , & Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis, 6*(1), 145–177.

Rodriguez, A., Dunson, D. B. , & Gelfand, A. E (2008). The nested Dirichlet process. *Journal of the American Statistical Association, 103*, 1131–1154.

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. *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.

Savitsky, T. D. , & Paddock, S. M. (2013). Bayesian nonparametric hierarchical modeling for multiple membership data in grouped attendance interventions. *Annals of Applied Statistics, 7*, 1074–1094.

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).

Shahbaba, B., & Neal, R. M. (2009). Non-linear models using Dirichlet process mixtures. *Journal of Machine Learning Research, 10*, 1829–1850.

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 & Methods, A5*(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. (1978c). 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. (2009). Indian buffet processes with power-law behavior. In *Advances in neural information processing systems* (Vol. 22).

Teh, Y. W., Gorur, D., & Ghahramani, Z. (2007). Sick-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. 556–63). Brookline, MA: Microtone.

Teh, Y. W. , & Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In N. L. Hjort, et al. (Eds.). *Bayesian nonparametrics*. Cambridge series in statistical and probabilistic mathematics.

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, MA: 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.

Thibaux, R. (2008). *Nonparametric Bayesian Models for Machine Learning*. Ph.D. dissertation, Department of Statistics, University of California, Berkeley.

Thibaux, R. , & Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In M. Meila and X. Shen (Eds.). *Proceedings of the International Conference on Artificial Intelligence and Statistics* (Vol. 11, pp. 564–571). Brookline, MA: Microtone.

Titsias, M. K. (2008). The infinite Gamma-Poisson feature model. *Advances in Neural Information Processing Systems, 20*, 97–104.

Tiwari, R. C. (1981). *A Mathematical Study of the Dirichlet Process*. Ph.D. dissertation, Department of Statistics, Florida State University.

Tiwari, R. C. (1988). Convergence of the Dirichlet Invariant measures and the limits of Bayes estimates. *Communications in Statistics - Theory & Methods, 17*(2), 375–393.

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 & Methods, A17*(10), 3591–3606.

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 & 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.

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.

Wade, S., Dunson, D. B., Petrone, S. , & Trippa, L. (2014). Improving prediction from Dirichlet process mixtures via enrichment. *Journal of Machine Learning Research, 15*, 1041–1071.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation, 36*, 45–54.

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., Damien, P., Laud, P., & Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society B, 61*, 485–527.

Walker, S. G. , & Mallick, B. K. (1997). A note on the scale parameter of the Dirichlet Process. *Canadian Journal of Statistics, 25*, 473–479.

Walker, S. G. , & Mallick, B. K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society B, 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.

West, M. (1992). Modelling with mixtures (with discussion). In J. M. Bernardo, J. 0. Berger, A. P. Dawid, & A. F. M. Smith (Eds.). *Bayesian statistics* (Vol. 4, pp. 503–524). Oxford: Oxford University Press.

West, M., Müller, P. , & Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In A. F. M. Smith & P. R. Freeman (Eds.). *Aspects of uncertainty: A tribute to D.V. Lindley* (pp. 363–386). New York: Wiley.

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 & Methods, A6*, 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 & 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 & 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.

Zabel, S. L. (1982). W. E. Johnson's "sufficientness" postulate. *Annals of Statistics, 10*, 1091–1099.

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 & Methods, 15*(10), 3079–3101.

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 & 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 & Methods, 14*(8), 1769–1778.

Author Index

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1

Subject Index

E.G. Phadia, *Prior Processes and Their Applications*, Springer Series in Statistics,
DOI 10.1007/978-3-319-32789-1