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, $\text{Be}(a, b)$ 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 $\mathcal{DGI}(\alpha + \sum_{i=1}^{n} \delta_{X_i}^{g})$, where $\delta_{X_i}^{g} = (1/k) \sum_{i=1}^{k} \delta_{gX_i}$. It is found to be useful in solving 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., $q_j = \left(F(t_j) - F(t_{j-1})\right) / (1 - F(t_{j-1}))$. 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 $\nu\left(d\theta, dp\right) = cp^{-1}e^{-cp}dpG_0\left(d\theta\right),\ p > 0$. 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 $\alpha = G_0\left(\Omega\right)$ 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 $A_1,\ A_2 \in \Omega,\ P\left(A_1\right)$ and $P\left(A_1\right)$ 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 $\nu\left(d\theta, dp\right) = cp^{-1}\left(1 - p\right)^{c-1} dpB_0\left(d\theta\right),\ 0 < p < 1$.

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