Springer Series in Statistics

Peter Müller
Fernando Andrés Quintana
Alejandro Jara
Tim Hanson

Bayesian Nonparametric Data Analysis

Springer

Springer Series in Statistics

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

Peter Müller • Fernando Andrés Quintana •
Alejandro Jara • Tim Hanson

Bayesian Nonparametric Data Analysis

Springer

Peter Müller
Department of Mathematics
University of Texas at Austin
Austin, TX
USA

Fernando Andrés Quintana
Departamento de Estadística
Pontificia Universidad Católica
Santiago, Chile

Alejandro Jara
Departamento de Estadística
Pontificia Universidad Católica
Santiago, Chile

Tim Hanson
Department of Statistics
University of South Carolina
Columbia, SC
USA

ISSN 0172-7397 ISSN 2197-568X (electronic)
Springer Series in Statistics
ISBN 978-3-319-18967-3 ISBN 978-3-319-18968-0 (eBook)
DOI 10.1007/978-3-319-18968-0

Library of Congress Control Number: 2015943065

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

In this book, we review nonparametric Bayesian methods and models. The organization of the book follows a data analysis perspective. Rather than focusing on specific models, chapters are organized by traditional data analysis problems. For each problem, we introduce suitable nonparametric Bayesian models and show how they are used to implement inference in the given data analysis problem. In selecting specific nonparametric models, we favor simpler and traditional models over specialized ones. The organization by inferential problem leads to some repetition in the discussion of specific models when the same nonparametric prior is used in different contexts.

Historically, Bayesian nonparametrics and indeed Bayesian statistics in general remained largely theoretical except for very simple models. The "discovery" and subsequent widespread use of Markov chain Monte Carlo and other Monte Carlo methods in the 1990s has made Bayesian nonparametric models an attractive and computationally viable possibility only in the last 20 years. Thus, a review of Bayesian nonparametric data analysis would be incomplete without a discussion of posterior simulation methods. We include pointers to available software, in particular public domain R packages. R code for some of the examples is available at a software page for the book at

https://www.ma.utexas.edu/users/pmueller/bnp/.

In the text, references to the software page are labeled as "**Software note**."

Chapter 1 introduces the framework for nonparametric and semiparametric inference and discusses the distinction between Bayesian and classical nonparametric inference. Chapters 2 and 3 start with a discussion of density estimation problems. Density estimation is one of the simplest statistical inference problems, and has traditionally been a popular application for nonparametric Bayesian methods. The emphasis is on the Dirichlet process, Polya trees, and related models. Chapter 4 is about nonparametric regression, including nonparametric priors on residual distributions, nonparametric mean functions, and fully nonparametric regression. The latter is also known as density regression. Chapter 5 introduces methods for categorical data, including contingency tables for multivariate categorical data and methods specifically for ordinal data. Chapter 6 discusses applications to survival

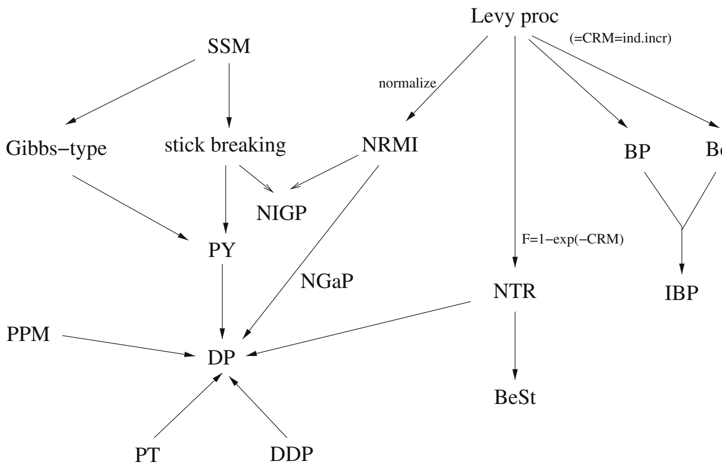

Fig. 1 Overview of some popular Bayesian nonparametric models for random probability measures. An *arrow* from model M1 to model M2 indicates that M2 is a special case of M1. For the case of NRMI and NTR, the *arrow* indicates that the descendant model is defined through a transformation of the CRM. Notice the central role of the Dirichlet process (DP) model

analysis. Probability models for a hazard function and random probability models for event times are a traditional use of nonparametric Bayesian methods. Perhaps this is the case because for event times it is natural to focus on details of the unknown distribution, beyond just the mean. Chapter 7 considers the use of random probability models in hierarchical models. Nonparametric priors for random effects distributions are some of the most successful and widespread applications of nonparametric Bayesian inference. In Chap. 8, we discuss models for random clustering and for feature allocation problems. Finally, in Chap. 9, we conclude with a brief discussion of some more problems. In the Appendix, we include a brief introduction to `DPpackage`, a public domain R package that implements inference for many of the models that are discussed in this book.

Figure 1 gives an idea of how popular nonparametric Bayesian models relate to each other. The Dirichlet process (DP), Polya tree (PT), Pitman Yor process (PY), normalized random measures (NRMI), and stick-breaking priors are discussed as priors for random probability measures in Chaps. 2 and 3. The dependent DP (DDP) is used to define a fully nonparametric regression model in Chap. 4, and then also features again in Chap. 6 for survival regression, and again in Chap. 7 to define a prior for a family of dependent random probability measures. Neutral to the right (NTR) processes come up in Chap. 6. The product partition model (PPM) and Gibbs-type priors are introduced as priors for random partitions, that is, cluster arrangement, in Chap. 8. The Indian buffet process (IBP) is introduced as a feature allocation model, also in Chap. 8.

The selection and focus is necessarily tainted by subjective choices and preferences. Finally, we recognize that the outlined classification of data analysis problems is arbitrary. Meaningful alternative organizations could have focused on the probability models, or on application areas.

Austin, TX, USA — Peter Müller
Santiago, Chile — Fernando Andrés Quintana
Santiago, Chile — Alejandro Jara
Columbia, SC, USA — Tim Hanson

Acronyms

We use the following acronyms. When applicable we list a corresponding section number in parentheses. We omit acronyms that are only used within specific examples.

AH	Accelerated hazards (Sect. 6.2.4)
AFT	Accelerated failure time (Sect. 6.2.2)
ANOVA DDP	DDP with AN(C)OVA model on $\{m_{hx},\ x \in X\}$ (Sect. 4.4.2)
BART	Bayesian additive regression trees (Sect. 4.3.4)
BNP	Bayesian nonparametric (model, inference)
CALGB	Cancer and Leukemia Group B
CART	Classification and regression tree (Sect. 4.3.4)
c.d.f.	Cumulative distribution function
CI	Credible interval
CPO	Conditional predictive ordinate (Chap. 9)
CRM	Completely random measure (Sect. 3.5.2)
CSDP	Centrally standardized Dirichlet process (Sect. 5.2.1)
DDP	Dependent Dirichlet process (Sect. 4.4.1)
DP_K	Finite Dirichlet process (Sect. 2.4.6)
DPM	Dirichlet process mixture (Sect. 2.2)
DPT	Dependent Polya tree (Sect. 4.4.3)
DP	Dirichlet process (Sect. 2.1)
FFNN	Feed-forward neural network (Sect. 4.3.1)
FPT	Finite Polya tree (Sect. 3.2.2)
GLM	Generalized linear model (Sect. 5.2.)
GLMM	Generalized linear mixed model (Sect. 5.2.2)
GP	Gaussian process (Sect. 4.3.3)
HDPM	Hierarchical Dirichlet process mixture (Sect. 7.3.1)
IBP	Indian buffet process (Sect. 8.5.2)
LDTFP	Linear dependent tail-free process (Sect. 4.4.3)
LPML	Log pseudo marginal likelihood (Chap. 9)
MAD	MAP-based asymptotic derivation

MAP	Maximum a posterior estimate
MCMC	Markov chain Monte Carlo (posterior simulation)
m.l.e.	Maximum likelihood estimator
MPT	Mixture of Polya tree (Sect. 3.2.1)
NRMI	Normalized random measure with independent increments (Sect. 3.5.2)
NTR	Neutral to the right (Sect. 6.1.1)
p.d.f.	Probability density function
PD	Pharmacodynamics (Sect. 7.2)
PH	Proportional hazards (Sect. 6.2.1)
PK	Pharmacokinetics (Sect. 7.2)
PO	Proportional odds (Sect. 6.2.3)
PPM	Product partition model (Sect. 8.3)
PPMx	Product partition model with regression on covariates (Sect. 8.4.1)
PT	Polya tree (Sect. 3.1)
PY	Pitman-Yor process (Sect. 2.5.2)
SSM	Species sampling model (Sect. 2.5.2)
TF	Tail free (Sect. 2.5.1)
WBC	White blood cell counts
WDDP	Weight dependent Dirichlet process (Sect. 4.4.4).

We use the following notational conventions. *Probability measures:* We use $p(\cdot)$ to generically indicate probability measures. The use of the arguments in $p(\cdot)$ or the context clarifies which probability measure is meant. Only when needed we use subindices, such as $p_X(\cdot)$, or introduce specific names, such as $q(\cdot)$. We use $\pi(\cdot)$ to indicate (upper level) prior probability models, usually BNP priors on a random probability measure, for example, $\pi(G)$ for a random probability measure G. When the use is clear from the context, we use $p(\cdot)$, etc. to also refer to the p.d.f, and introduce separate notation only when we wish to highlight something. We use $f_\theta(\cdot)$ for kernels and $f_G(\cdot) = \int f_\theta(\cdot)\, dG(\theta)$ for a mixture. We use notation like $\mathsf{Be}(\epsilon \mid a, b)$ to indicate that the random variable ϵ follows a $\mathsf{Be}(a, b)$ distribution. *Variables:* We use y_i for observed outcomes, x_i for known covariates, boldface ($\boldsymbol{y}$) for vectors, and uppercase symbols (A), or boldface uppercase ($\boldsymbol{C}$) when needed for distinction or emphasis, for matrices. *Clusters:* Many models include a notion of clusters. We use * to mark quantities that are cluster-specific, such as $\boldsymbol{y}_j^\star$, $\theta_j^\star$, etc.

Contents

Chapter 1
Introduction

Abstract We introduce the setup of nonparametric and semiparametric Bayesian models and inference.

Statistical problems are described using probability models. That is, data are envisioned as realizations of a collection of random variables $y_1, \ldots, y_n$, where y_i itself could be a vector of random variables corresponding to data that are collected on the i-th experimental unit in a sample of n units from some population of interest. A common assumption is that the y_i are drawn independently from some underlying probability distribution G. The statistical problem begins when there exists uncertainty about G. Let g denote the probability density function (p.d.f.) of G. A statistical model arises when g is known to be a member g_θ from a family $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$, labeled by a set of parameters θ from an index set Θ.

Models that are described through a vector θ of a finite number of, typically, real values are referred to as finite-dimensional or *parametric models*. Parametric models can be described as $\mathcal{G} = \{g_\theta : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p\}$. The aim of the analysis is then to use the observed sample to report a plausible value for $\boldsymbol{\theta}$, or at least to determine a subset of $\boldsymbol{\Theta}$ which plausibly contains $\boldsymbol{\theta}$. In many situations, however, constraining inference to a specific parametric form may limit the scope and type of inferences that can be drawn from such models. Therefore, we would like to relax parametric assumptions to allow greater modeling flexibility and robustness against mis-specification of a parametric statistical model. In these cases, we may want to consider models where the class of densities is so large that it can no longer be indexed by a finite dimensional parameter $\boldsymbol{\theta}$, and we therefore require parameters $\boldsymbol{\theta}$ in an infinite dimensional space.

Example 1 (Density Estimation) Consider a simple random sample $y_i \mid G \overset{\text{iid}}{\sim} G$, $i = 1, \ldots, n$, from some unknown distribution G. One could now proceed by restricting G to a normal location family, say $\mathcal{G} = \{\mathsf{N}(\theta, 1) : \theta \in \mathbb{R}\}$. Figure 1.1a shows the resulting inference conditional on an assumed random sample $y_1, \ldots, y_n$. Naturally, inference about the unknown G is restricted to the assumed normal location family and does not allow for multi-modality or skewness. In contrast, a nonparametric model would proceed with a prior probability model π for the unknown distribution G. For example, later we will introduce the Dirichlet process mixture prior for G.

P. Mueller et al., *Bayesian Nonparametric Data Analysis*, Springer Series in Statistics, DOI 10.1007/978-3-319-18968-0_1

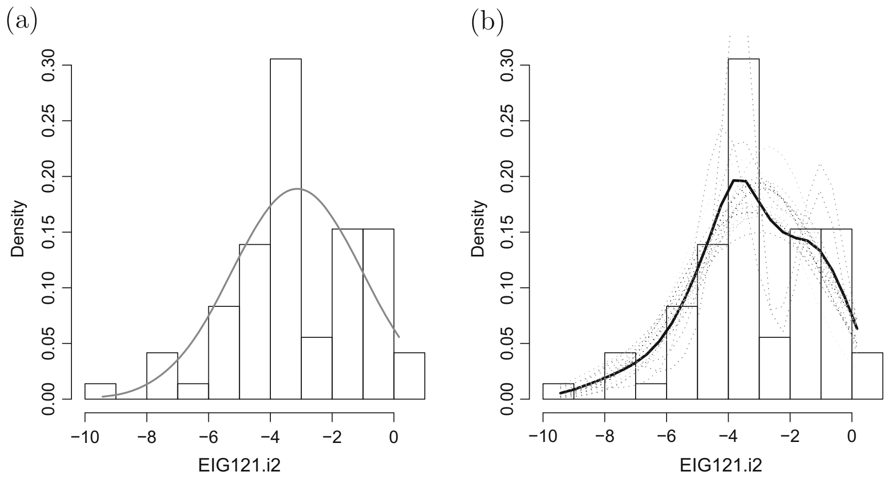

Fig. 1.1 Example 1. Inference on the unknown distribution G under a parametric model (panel **a**) and nonparametric model (panel **b**). The histogram of the observed data is also displayed. The *dotted lines* in panel (**b**) correspond to posterior draws

Figure 1.1b contrasts the parametric inference with the flexible BNP inference under a Dirichlet process mixture prior.

In Example 1 the unknown, infinite dimensional, parameter is the distribution G itself. Another example of an infinite-dimensional parameter space is the space of continuous functions defined on the real line, $\mathcal{S} = \{m(z) : z \in \mathbb{R},\ m(\cdot) \text{ is a continuous function}\}$. This could arise, for example, in a regression model with unknown mean function $m(z)$. Models with infinite-dimensional parameters are referred to as *nonparametric models* (Ghosh and Ramamoorthi 2003; Tsiatis 2006). In some other cases, it is useful to write the infinite-dimensional parameter $\boldsymbol{\theta}$ as $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ is a q-dimensional parameter and $\boldsymbol{\theta}_2$ is an infinite-dimensional parameter. These models are referred to as *semiparametric models* because both a parametric component $\boldsymbol{\theta}_1$ and a nonparametric component $\boldsymbol{\theta}_2$ describe the model (see e.g., Tsiatis 2006). As an example of a semiparametric model, consider the proportional hazards model that is commonly used in modeling a survival time T as a function of a vector of covariates z. The model was first introduced by Cox (1972). Let

$$\lambda(t \mid z) = \lim_{h \to 0} \left\{ \frac{p(t \leq T < t + h \mid T \geq t, z)}{h} \right\} \tag{1.1}$$

denote the conditional hazard rate, conditional on some covariates z. The proportional hazards model assumes

$$\lambda(t \mid z) = \lambda_0(t) \exp(z'\boldsymbol{\beta}). \tag{1.2}$$

where $\lambda_0(\cdot)$ is the underlying or baseline hazard function and $\boldsymbol{\beta}$ is a q-dimensional vector of regression coefficients. In the classical semiparametric version of the model, the underlying hazard function is left unspecified. Since this function can be any positive function in t, subject to some regularity conditions, it is an infinite-dimensional parameter. The parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda_0)$ completely characterize the data generating mechanism. In fact, the density f_T of the survival time T is related to the hazard function through

$$f_T(t \mid z) = \lambda_0(t) \exp(z'\boldsymbol{\beta}) \exp\left\{-\exp(z'\boldsymbol{\beta}) \int_0^t \lambda_0(u)du\right\}.$$

The parameters of interest can be written as $\boldsymbol{\theta}_1 = \boldsymbol{\beta}$ and $\boldsymbol{\theta}_2 = \lambda_0$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \boldsymbol{\Theta} = \mathbb{R}^q \times \mathcal{S}$ and $\mathcal{S}$ is the infinite-dimensional space of all nonnegative functions on $\mathbb{R}^+$ with infinite integral over $[0, \infty)$.

Example 2 (Oral Cancer) We use a dataset from Klein and Moeschberger (2003, Sect. 1.11). The data report survival times for 80 patients with cancers of the mouth. Samples are recorded as aneuploid (abnormal number of chromosomes) versus diploid (two copies of each chromosome) tumors. We define $z_i \in \{0, 1\}$ as an indicator for aneuploid tumors and carry out inference under model (1.2) with a BNP prior on λ_0. Figure 1.2 shows the estimated hazard curves under $z = 0$ and

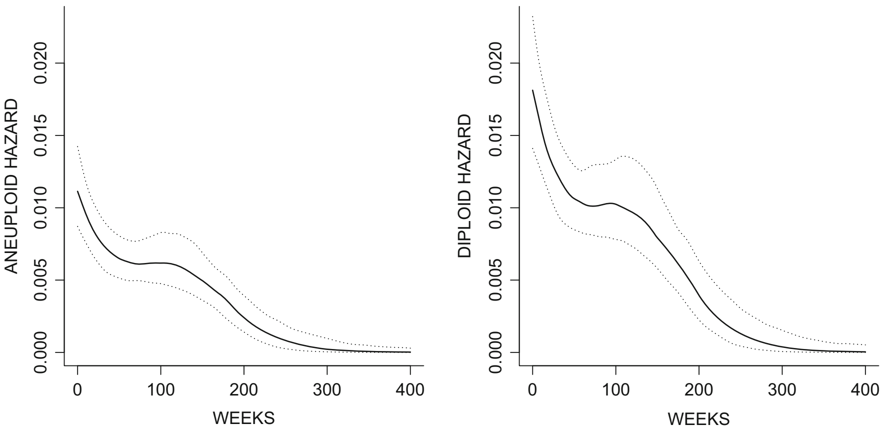

Fig. 1.2 Example 2. Hazard curves for aneuploid and diploid groups under the proportional hazard model with point-wise 50 % CIs

$z = 1$. The prior on $\log\{\lambda_0(\cdot)\}$ is a penalized B-spline, fit through the R package `R2BayesX`.

To proceed with Bayesian inference in a nonparametric model we need to complete the probability model with a prior on the infinite-dimensional parameter. Such priors are known as Bayesian nonparametric (BNP) priors. We take this as a definition of BNP models. That is, we define BNP priors as probability models for infinite-dimensional parameters and refer to the entire inference model as a BNP model. This highlights the similarities and distinctions of Bayesian versus classical nonparametric inference. Under both BNP and classical nonparametric approaches an infinite-dimensional parameter characterizes the family of sampling probability models. The major difference between Bayesian and classical nonparametrics is that Bayesian inference completes the model with a prior on the infinite-dimensional parameter (random probability measure, regression function, etc.). As a result, inference includes a full probabilistic description of all relevant uncertainties. Classical nonparametrics, on the other hand, treats infinite-dimensional parameters as nuisance parameters and derives procedures where they are left unspecified to make inferences on finite-dimensional parameters of interest.

Infinite-dimensional parameters of interest are usually functions. Functions of common interest include probability distributions, or conditional trends, e.g. mean or median regression functions. Consideration of probability distributions requires the definition of probability measures on a collection of distribution functions. Such probability measures are generically referred to as random probability measures.

While our main focus is on data analysis and how to build models in some important special cases, it is important to know that there is a solid body of theory supporting the use of nonparametric models. In the upcoming discussion we will briefly state some of the important results and the particular effect they have on models. But we stop short of an exhaustive list of BNP prior models. An excellent recent review of a large number of BNP models appears in Phadia (2013). See also Figure 1.1 in Phadia (2013), which is an interesting variation of Fig. 1 in the preface of the current text. Other recent discussions of BNP priors include Hjort et al. (2010), including an excellent and concise review of BNP models beyond the Dirichlet process in Lijoi and Prünster (2010), Hjort (2003), Müller and Rodríguez (2013), Müller and Quintana (2004), Walker et al. (1999), and Walker (2013). Gelman et al. (2014, Part V) includes a discussion of nonparametric Bayesian data analysis. A mathematically rigorous discussion, with an emphasis on asymptotic properties can be found in the forthcoming book by Ghoshal and van der Vaart (2015).

In this text we will not prove any new results and therefore never need to refer to measure theoretic niceties. We refer interested readers to Phadia (2013), who discusses all the same models that also feature in this text. See also Ghosh and Ramamoorthi (2003), Ghoshal (2010), and Ghoshal and van der Vaart (2015) for a mathematically more rigorous discussion. Briefly summarized, assume an underlying probability space $(\Omega, \mathcal{A}, \mu)$ and let S be a complete and separable metric space equipped with the Borel σ-algebra $\mathcal{B}$. Denote by $M(S)$ the space of probability measures on S endowed with the topology of weak convergence which makes it a complete and separable space. Most BNP priors that we introduce in the following discussion are distributions over $M(S)$ or, in other terms, laws of random probability measures i.e. random elements defined on Ω and taking values in $M(S)$.

References

Cox DR (1972) Regression models and life-tables (with discussion). J R Stat Soc Ser B 34:187–220

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014) Bayesian data analysis. CRC Press, Boca Raton

Ghosh JK, Ramamoorthi RV (2003) Bayesian nonparametrics. Springer, New York

Ghoshal S (2010) The Dirichlet process, related priors and posterior asymptotics. In: Hjort NL, Holmes C, Müller P, Walker SG (eds) Bayesian nonparametrics. Cambridge University Press, Cambridge, pp 22–34

Ghoshal S, van der Vaart A (2015) Fundamentals of nonparametric Bayesian inference. Cambridge University Press, Cambridge

Hjort NL (2003) Topics in nonparametric Bayesian statistics. In: Green P, Hjort N, Richardson S (eds) Highly structured stochastic systems. Oxford University Press, Oxford, pp 455–487

Hjort NL, Holmes C, Müller P, Walker SG (eds) (2010) Bayesian nonparametrics. Cambridge University Press, Cambridge

Klein JP, Moeschberger ML (2003) Survival analysis: techniques for censored and truncated data. Springer, New York

Lijoi A, Prünster I (2010) Models beyond the Dirichlet process. In: Hjort NL, Holmes C, Müller P, Walker SG (eds) Bayesian nonparametrics. Cambridge University Press, Cambridge, pp 80–136

Müller P, Quintana FA (2004) Nonparametric Bayesian data analysis. Stat Sci 19(1):95–110

Müller P, Rodríguez A (2013) Nonparametric Bayesian inference. IMS-CBMS Lecture notes. IMS, Beachwood

Phadia EG (2013) Prior processes and their applications. Springer, New York

Tsiatis AA (2006) Semiparametric theory and missing data. Springer, New York

Walker S (2013) Bayesian nonparametrics. In: Damien P, Dellaportas P, Polson NG, Stephens DA (eds) Bayesian theory and applications. Oxford University Press, Oxford, pp 249–270

Walker SG, Damien P, Laud PW, Smith AFM (1999) Bayesian nonparametric inference for random distributions and related functions (with discussion). J R Stat Soc B 61:485–527

Chapter 2
Density Estimation: DP Models

Abstract We discuss the use of nonparametric Bayesian models in density estimation, arguably one of the most basic statistical inference problems. In this chapter we introduce the Dirichlet process prior and variations of it that are the by far most commonly used nonparametric Bayesian models used in this context. Variations include the Dirichlet process mixture and the finite Dirichlet process. One critical reason for the extensive use of these models is the availability of computation efficient methods for posterior simulation. We discuss several such methods.

Density estimation is concerned with inference about an unknown distribution G on the basis of an observed i.i.d. sample,

$$y_i \mid G \stackrel{\text{iid}}{\sim} G, \quad i = 1, \ldots, n. \tag{2.1}$$

If we wish to proceed with Bayesian inference, we need to complete the model with a prior probability model π for the unknown parameter G. Assuming a prior model on G requires the specification of a probability model for an infinite-dimensional parameter, that is, a BNP prior.

2.1 Dirichlet Process

2.1.1 Definition

One of the most popular BNP models is the Dirichlet process (DP) prior. The DP model was introduced by Ferguson (1973) as a prior on the space of probability measures.

Definition 1 (Dirichlet Process—DP) Let $M > 0$ and G_0 be a probability measure defined on S. A DP with parameters (M, G_0) is a random probability measure G defined on S which assigns probability $G(B)$ to every (measurable) set B such that for each (measurable) finite partition $\{B_1, \ldots, B_k\}$ of S, the joint distribution of the

P. Mueller et al., *Bayesian Nonparametric Data Analysis*, Springer Series in Statistics, DOI 10.1007/978-3-319-18968-0_2