A Gentle Introduction to Bayesian Nonparametrics

Pietro Lesci

15 January 2019

Contents

Preface

In this book we will review some basic concepts regarding Bayesian Nonparametric statistics. In the Chapter 1, we will provide the core theoretical ideas behind the nonparametric approach and how it is implemented from a Bayesian perspective. In Chapter 2, we will review the definition, construction and properties of the Dirichlet Process: the "normal distribution of Bayesian nonparametrics" (Ghosal and Vaart 2017). Finally, in Chapter 3, we will focus on regression problems and how they are approached from a Bayesian nonparametric perspective.

Chapter 1

Introduction

In this chapeter we will provide the core theoretical ideas behind the nonparametric approach and how it is implemented from a Bayesian perspective. We will review basic definitions and set-out the notation that will be used in the book. We will describe how the challenging task of constructing priors on infinite-dimensional objects has been tackeld.

Bayesian nonparametrics concerns *Bayesian inference* methods for *nonparametric* models. A nonparametric model involves at least one infinite-dimensional parameter and hence may also be referred to as an "infinite-dimensional model". Indeed, the nomenclature "nonparametric" is misleading since it gives the impression that there are no parameters in the model, while in reality there are infinitely many unknown quantities. Examples of infinite-dimensional parameter are functions or measures. The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible. Usually, this means using statistical models that are infinite-dimensional.

Where does Bayesian nonparametrics fit? We can distinguish four sectors of statistical methodologies, defined by the interplay two factors (Hjort et al. 2010) whose "cartesian product" defines four "boxes":

- Parametric vs Nonparametric

- Frequentist vs Bayesian

The *Frequentist parametrics* box contains methods, developed from 1920s onwards, like maximum likelihood, optimum tests, with calculation of p-values, optimal estimators, confidence intervals, and so forth. Some of the procedures stem from exact probability calculations while others (a lot of them) relate to the application of large-sample techniques.

The *Bayesian parametrics* box comprises classic methodology for prior and posterior distributions in models with a finite number of parameters. Such methods, starting from the premise that uncertainty about model parameters may somehow be represented in terms of probability distributions, have been around for more than a hundred years (the "Bayes theorem" dates back to the 1763), but they were limited to a to simple statistical models and priors. The applicability of Bayesian parametrics widened significantly with the advent of modern computers and the development of further numerical methods and software packages pertaining to numerical integration and Markov chain Monte Carlo (MCMC) simulations. Asymptotics is also useful for Bayesian parametrics.

The *Frequentist nonparametrics* box contains mixed objects. The term has historically been associated with test procedures that are "distribution free" leading also to nonparametric confidence intervals and bands. Furthermore, still belong to this box methodologies for the estimation of probability densities and regression functions without parametric assumptions; and also specific computational techniques such as the bootstrap.

Finally, the *Bayesian nonparametrics* box comprises models and methods characterized by big (infinite) parameter spaces and construction of probability measures over these spaces. Typical examples include Bayesian setups for density estimation, nonparametric regression with or without a fixed error distribution, survival function estimation for survival analysis, etc. In the Bayesian nonparametric paradigm, a prior distribution is assigned to all relevant unknown quantities, whether finite or infinite dimensional. The posterior distribution is the conditional distribution of these quantities, given the data, and is the basis for all inference – as in any Bayesian inference, except that the unknown quantities or

parameters may be infinite dimensional in this case. A model completely specifies the conditional distribution of all observed, given all unobserved quantities, while a prior distribution specifies the distribution of all unobservables. The posterior distribution involves an inversion of the order of conditioning and gives the distribution of the unobservables given the observables. Latent variables are unobservables and are treated in the same way as the unknown parameters used to describe the model.

Why adopt the nonparametric Bayesian approach for inference? Nonparametric models can allow one to avoid the arbitrary and possibly unverifiable assumptions inherent in parametric models. While Bayesian procedures may be desirable for philosophical or practical reasons.

Let's now look at the matter from another point of view: let's try to interpret Bayesian parametric models from a Bayesian nonparametric standpoint. Parametric models make restrictive assumptions about the data generating mechanism, which may cause serious bias in inference. In the Bayesian framework a parametric model assumption can be viewed as an extremely strong prior opinion. Indeed, a parametric model specification $X|\theta \sim p_\theta$, for $\theta \in \Theta \subset \mathbb{R}^d$, with a prior $\theta \sim \Pi$, may be considered within a nonparametric Bayesian framework as $X|p \sim p$, for $p \in \mathcal{P}$ with $\mathcal{P}$ a set of densities equipped with a prior $p \sim \Pi$ with the property that $\Pi(\{p_\theta : \theta \in \Theta\}) = 1$. Thus parametric modelling is equivalent to insisting on a prior that assigns probability one to a thin subset of all densities.

1.1 The problem

A Bayesian analysis cannot proceed without a prior distribution on all parameters. A prior ideally expresses a quantification of pre-experiment and subjective knowledge. A prior on a function requires knowing many aspects of the function – e.g. for a probability density function that it integrates to one – including the ability to quantify the information in the form of a *probability measure*.

This poses an apparent conceptual contradiction, as expressed in Ghosal and Vaart (2017): "A nonparametric Bayesian approach is pursued to minimize restrictive

parametric assumptions, but at the same time requires specification of the minute details of a prior on an infinite-dimensional parameter".

It is usually thought that inference must be based on an *objective prior*. This is vaguely understood as a prior that is proposed by some automatic mechanism that is not in favour of any particular parameter values, and has low information content compared to the data – also called *default priors*. Some of the earliest statistical analyses in this way suggested to use a uniform prior. Uniform priors were strongly criticised for lacking invariance (e.g. to nonlinear trasformations). However, invariance-friendly methods such as Jeffreys' priors and reference analysis arised, although most of these ideas are restricted to finite-dimensional parametric problems. An objective prior should be automatically constructed using a default mechanism. It need not be non-informative, but should be spread all over the parameter space (Ghosal and Vaart 2017). Unlike in parametric situations, where non-informative priors are often *improper*, default priors considered in nonparametric Bayesian inference are almost always *proper*. Large support of the prior means that the prior is not too concentrated in some particular region. This generally causes that the prior is subdued gradually by the data if the sample size increases, so that eventually the data overcome the prior.

A prior should also have some "good" properties such as **robustness**: Bayesian robustness means that the choice of the prior does not influence the posterior distribution too much. Another way to formulate the problem is to study the *asymptotic properties* of the prior, as the information in the data increases indefinitely, such as *posterior consistency* – which means that the posterior probability eventually concentraes in a (any) small neighborhood of the actual value of the parameter – and *rate of convergence* or functional limit of the prior.

Another "good" property a prior must have is a "nice" structure. This is important in the context of computations: we cannot directly simulate from the posterior distribution with infinitely many parameters in a finite time. Therefore, unless it is parameterized by finitely many parameters the problem is infeasible. We must break up the function of interest into more elementary finite-dimensional objects, and simulate from their posterior distribution. For this reason the structure of the prior is important. Useful, or nice, structures may come from conjugacy or

approximation. Loosely speaking, what we usually try to do is to integrate out the infinite-dimensional parameter given the latent variables.

1.2 Intermezzo: basic concepts

Probability space. A probability space is a triple $(\Omega, \mathcal{A}, P)$ consisting of:

- The *sample space* Ω: an arbitrary non-empty set
- The *σ-algebra* $\mathcal{A} \subseteq 2^{\Omega}$ (also called σ-field): a set of subsets of Ω, called events, such that:
  - It contains the sample space, $\Omega \in \mathcal{A}$
  - It is closed under complements: if $A \in \mathcal{A}$, then also $(\Omega \setminus A) \in \mathcal{A}$
  - It is closed under countable unions: if $A_i \in \mathcal{A}$ for $i = 1, 2, \ldots$, then also $(\bigcup_{i=1}^{\infty} A_i) \in \mathcal{A}$; by the De Morgan's law it also holds that $\mathcal{A}$ is also closed under countable intersections: if $A_i \in \mathcal{A}$ for $i = 1, 2, \ldots$, then also $(\bigcap_{i=1}^{\infty} A_i) \in \mathcal{A}$
- The *probability measure* $P : \mathcal{A} \to [0, 1]$: a function on $\mathcal{A}$ such that:
  - It is countably additive (also called σ-additive): if $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{A}$ is a countable collection of pairwise disjoint sets, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
  - The measure of entire sample space is equal to one: $P(\Omega) = 1$

Measure. Let Ω be a set and $\mathcal{A}$ a σ-algebra over Ω. A function $\mu : \Omega \to \mathbb{R}$, the extended real line, is called a *measure* if it satisfies the following properties[1]:

- *Non-negativity*: For all A in $\mathcal{A} : \mu(A) \geq 0$
- *Null empty set*: $\mu(\varnothing) = 0$.
- *Countable additivity (or σ-additivity)*: For all countable collections $\{A_i\}_{i=1}^{\infty}$ of pairwise disjoint sets in $\Omega : \mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k)$

Measurable space: A measurable space is a pair $(\Omega, \mathcal{A})$ consisting of a set Ω and a σ-algebra $\mathcal{A}$ of subsets of Ω. Also, $A \in \mathcal{A}$ are called *measurable sets*.

[1]Note that the greatest difference with a *probability measure* P is that $\mu(\Omega)$ is not bounded in $[0, 1]$, that is they differ in their support.

Measurable function: Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A map $f : X \to Y$ is called measurable if $f^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$.

Random variable: If P is a probability measure on $(\Omega, \mathcal{A})$, a random variable X taking values in X is simply a measurable function $X : \Omega \to \mathrm{X}$. Intuition: think of the probability space $(\Omega, \mathcal{A}, P)$ as a black-box random number generator, and X as a function taking random samples in Ω and producing random samples in X.

Markov Kernel. Let $(X, \mathcal{A}), (Y, \mathcal{B})$ be measurable spaces. A Markov kernel with source $(X, \mathcal{A})$ and target $(Y, \mathcal{B})$ is a map $\kappa : X \times \mathcal{B} \to [0, 1]$ with the following properties:

- The map $x \mapsto \kappa(x, B)$ is $\mathcal{A}$-measurable for every $B \in \mathcal{B}$
- The map $B \mapsto \kappa(x, B)$ is a probability measure on $(Y, \mathcal{B})$ for every $x \in X$

In other words it associates to each point $x \in X$ a probability measure $\kappa(x, \cdot)$ on $(Y, \mathcal{B})$ such that, for every measurable set $B \in \mathcal{B}$, the map $x \mapsto \kappa(x, B)$ is measurable with respect to the σ-algebra $\mathcal{A}$.

Polish space. A topological space is called Polish if its topology is generated by a metric that makes it complete and separable. A metric space $\mathbb{M}$ is called *complete* if every Cauchy sequence is convergent. A topological space is called *separable* if there exists a sequence $\{x_n\}_{n=1}^{\infty}$ of elements of the space such that every nonempty open subset of the space contains at least one element of the sequence.

Borel set. A Borel set is any set in a topological space X that can be formed from open sets through the operations of countable union, countable intersection, and relative complement. The collection of all Borel sets on X forms a σ-algebra, known as the Borel σ-algebra, which is the smallest σ-algebra containing all open sets.

Stochastic process. A stochastic process indexed by a set $\mathbb{I}$ is a collection $W = (W(i) : i \in \mathbb{I})$ of random variables defined on a probability space. A sample path $i \to W(i)$, given an $i \in \mathbb{I}$, of W is a *random function* and hence the law of W, the *law of the set of sample paths*, is a prior on a space of functions $f : \mathbb{I} \to \mathbb{R}$.

Regular conditional probability.[2] Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $X : \Omega \to \mathrm{X}$ be a random variable, defined as a Borel-measurable function from Ω to its state space $(\mathrm{X}, \mathcal{B})$. Then a regular conditional probability is defined as a function $\nu : \mathrm{X} \times \mathcal{A} \to [0, 1]$, i.e. a Markov kernel, such that for all $A \in \mathcal{A}$ and all $B \in \mathcal{B}$, $P(A \cap X^{-1}(B)) = \int_B \nu(x, A)\ P(X^{-1}(dx))$, or equivalently, $P(A|X = x) = \nu(x, A)$.

1.3 Priors on spaces of probability measures: Definition

Now that we have review some basic definitions, we can proceed analysing the construction of priors properly. Let's repeat, once, that in the Bayesian framework the data X follows a distribution determined by a parameter θ, which is itself considered to be generated from a prior distribution Π. The corresponding posterior distribution is the conditional distribution of θ given X. This framework is identical in parametric and nonparametric Bayesian statistics, the only difference being the dimensionality of the parameter. Obviously, proper definitions of priors and (conditional) distributions require more care in the nonparametric case. Let's formalize the problem.

In the nonparametric setting it is natural to placing a prior distribution directly on the model. However, this is complicated. To make the task easier, usually it is assumed that the sample space $(\mathrm{X}, \mathcal{A})$ is a Polish space with its Borel σ-field (the smallest σ-field containing all open sets). This ensures some nice mathematical properties. The prior is defined on the collection $\mathrm{M} = \mathrm{M}(\mathrm{X})$ of all probability

[2]Motivation: Normally we define the conditional probability of an event A given an event B as: $P(A|B) = \frac{P(A \cap B)}{P(B)}$. The difficulty with this arises when the event B is too small to have a non-zero probability.

measures on $(\mathrm{X}, \mathcal{A})$ are consided. Simmetrically, we could define a prior on $\mathrm{M} = \mathrm{M}(\Theta)$ the set of all probability measures on $(\Theta, \mathcal{B})$ the parameter space.

A prior Π on M is a probabiity measure on a σ-field of subsets of M. We can look at this prior from another point of view though, perhaps more intuitive. A prior on M can be viewed as the *law of a random measure* P and can be identified with the collection of "random probabilities" $P(A)$ of sets $A \in \mathcal{A}$. Usually the measurability structure of M is choosen such that each of these $P(A)$ is a random variable, and thus $(P(A) : A \in \mathcal{A})$ is a stochastic process on the underlying probability space: given a specific A, $P(A)$ is random because P is random and it is distributed according to this prior we want to assign. It is a strange (complicated) stochastic process: the index set is composed by the elements of the σ-algebra $\mathcal{A}$!

Define $\mathcal{M}$ the smallest σ-field that makes all maps in M, $P : \mathcal{A} \to \mathbb{R}$, measurable for $A \in \mathcal{A}$. Thus, the priors Π's are measures on $(\mathrm{M}, \mathcal{M})$. Now we can restate the problem of finding a prior on all probability measures defined on X more clearly.

The parameter θ that indexes the statistical model $(P_\theta : \theta \in \Theta)$ can be taken equal to the distribution P itself, with M as the parameter set, giving a model of the form $(P : P \in \mathrm{M})$. In this way, the P's are Markov kernels from $(\mathrm{M}, \mathcal{M})$ to $(\mathrm{X}, \mathcal{A})$.

Let's recap what we have stated so far. A statistical model on a sample space X is a subset of probability measures, $M \subset \mathrm{M}(\mathrm{X})$, on X. The elements of the subset M are indexed by a parameter θ with values in a parameter space Θ, that is,

$$M = \{P_\theta | \theta \in \Theta\}$$

We call a model parametric if Θ has finite dimension, which usually means $\Theta \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$. If Θ has infinite dimension, M is called a nonparametric model: there is an infinite number of probability measures to choose from. Therefore, defining a nonparametric Bayesian model means defining a prior distribution on an infinite-dimensional space.

1.3.1 Example

As usual we formulate a statistical problem assuming that n observations $x_1, \ldots, x_n$ with values in X are collected. We model them as random variables $X_1, \ldots, X_n$. In classical statistics, we assume that these random variables are generated i.i.d. from a measure, P_θ, in the model, M, that is

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta, \qquad \theta \in \Theta$$

The objective of statistical inference is then to draw conclusions about the value of θ and, hence, about the distribution P_θ – which is indexed by θ – from the observations.

In Bayesian statistics, we model the parameter as a random variables: a basic principle of Bayesian statistics is that all forms of uncertainty should be expressed as randomness. We therefore have to interpret θ as a random variable with values in Θ. Thus, the parameter set Θ is equipped with a σ-field $\mathcal{B}$, thus $(\Theta, \mathcal{B})$ is a measurable space. The prior, $\Pi : \mathcal{B} \to [0, 1]$, is a probability measure on this measurable space; thus $(\Theta, \mathcal{B}, \Pi)$ is a probability space.

Then, we assume that the distribution P_θ of X given θ is a Markov kernel from $(\mathrm{X}, \mathcal{A})$ to $(\Theta, \mathcal{B})$, that is $P_\theta : \mathrm{X} \times \mathcal{B} \to [0, 1]$, i.e. a *regular conditional distribution* on the measurable space $(\mathrm{X}, \mathcal{A})$ such that $P_\theta(A)$ is a probability measure for every $\theta \in \Theta$ and is measurable for every $A \in \mathcal{A}$.

Then the pair (X, θ) has a well defined joint distribution on the product space $(\mathrm{X} \times \Theta, \mathcal{A} \times \mathcal{B})$, given by

$$\Pr(X \in A, \theta \in B) = \int_B P_\theta(A) d\Pi(\theta)$$

This gives rise to the marginal distribution of X, defined by

$$\Pr(X \in A) = \int P_\theta(A) d\Pi(\theta)$$

By Kolmogorov's definition of conditional probabilities, $\Pr(\theta \in B|X)$ for $B \in \mathcal{B}$ is always well defined, as measurable function of X.

The task to compute the **posterior distribution** is then equivalent to the task of finding a **Markov kernel** from $(\Theta, \mathcal{B})$ to $(\mathrm{X}, \mathcal{A})$, that is $P_X : \Theta \times \mathcal{A} \to [0, 1]$, i.e. a *regular conditional distribution* on the measurable space $(\Theta, \mathcal{B})$. A sufficient condition for the existence of such Markov kernel is that Θ is a *Polish space* and $\mathcal{B}$ its Borel σ-algebra.

Hence, the model P_θ, from a Bayesian perspective, consists of a model $M = \{P_\theta : \theta \in \Theta\}$ as above, called the observation model (or likelihood), and a prior Π. The data are, thus, generated in two stages (capital letters are used to define random variables)

$$\begin{aligned} \theta &\sim \Pi \\ X_1, X_2, \ldots | \theta &\overset{i.i.d.}{\sim} P_\theta \end{aligned}$$

The data are *conditionally* i.i.d. rather than i.i.d. Our objective is then to determine the posterior distribution, i.e. the conditional distribution of θ given the data

$$\Pi\left(\theta \in B | X_1 = x_1, \ldots, X_n = x_n\right)$$

The value of the parameter remains uncertain given a finite number of observations, and Bayesian statistics uses the posterior distribution to express this uncertainty.

Now we have identified what we are looking for, but how do we practically construct such priors on the space of all probability measures?

1.4 Priors on spaces of probability measures: Construction

The exist few methods to construct priors on spaces of probability measure, we will only review the most known (and simple).

1.4.1 Construction through a stochastic process

The best way to familiarize with this approach is to think that a distribution on an infinite-dimensional space Θ is a stochastic process with paths in Θ (Orbanz 2014). Using random processes, we are merely constructing random density functions with unrestricted shapes. The prior, in this case, becomes the law governing the stochastic process.

One general method of constructing a random measure is to start with the stochastic process $(P(A) : A \in \mathcal{A})$, constructed using Kolmogorov's consistency theorem: this theorem guarantees that a suitably "consistent" collection of finite-dimensional distributions will define a stochastic process. Next step is to show that this process can be realized within M, viewed as a subset of $R^{\mathcal{A}}$, the space of all functions that from $\mathcal{A}$ go in $\mathbb{R}$.

The details are as follows. For every *finite* collection $A_1, \ldots, A_k$ of Borel sets in X, the vector $(P(A_1), \ldots, P(A_k))$ of probabilities obtained from a random measure P is an ordinary random vector in $\mathbb{R}^k$ – given P, it becomes simply a vector of numbers between 0 and 1. The construction of P may start with the specification of the distributions of all vectors of this type. For any consistent specification of the distributions, Kolmogorov's theorem allows us to construct, on a suitable probability space $(\Omega, \mathcal{A}, P)$, a stochastic process $(P(A) : A \in \mathcal{A})$ with the given finite-dimensional distributions.

Example. A simple and important example is to specify the distributions of each one of these vectors as Dirichlet distributions with parameter vector $(\mu(A_1), \ldots, \mu(A_k))$, for a given Borel measure μ – a Borel measure is any measure μ defined on a Borel σ-algebra. Intuitively μ returns the probability of the partition A_k.

1.4.2 Construction in Countable Sample Spaces

A probability distribution on a countable sample space, say, Θ, equipped with the σ-field $\mathcal{B}$, can be represented as an infinite-length probability vector $s = (s_1, s_2, \ldots)$, where each s_k gives the probability of a partition B_k in the σ-field.

Basically we are saying that each vector s is a probability measure on Θ. As any probability measure, we would like each component of the vector to be positive and the infinite sum of the components be equal to one.

The set M in this case is the space of all these vectors s. A prior on M can therefore be identified with the distribution of a random element, a vector s_n, with values in the countable-dimensional unit **simplex** (that can be imagined as a multidimensional triangle), and we denote it as

$$\Delta := \left\{ (s_n)_{n\in\mathbb{N}} \ : \ s_n \geq 0 \quad \text{and} \quad \sum_{n=1}^{\infty} s_n = 1 \right\}$$

That is, it is the space of sequences s_n that respect the two properties stated: each element greater than zero and the sum of the elements equal to 1. We write $\mathrm{M} \stackrel{def}{=} \Delta$. The σ-field $\mathcal{M}$ on Δ is assumed to be generated by the coordinate map $i \mapsto s_i$ for $i \in \mathbb{N}$.

What we want to assign a probability measure to each of one of these sequences, that is we want to assign our prior.

A map π from some probability space into Δ is a random element if and only if every coordinate variable p_i is a random variable. Hence a prior simply corresponds to an infinite sequence of nonnegative random variables $\pi_1, \pi_2, \ldots$ that adds up to 1. Constructing priors using Kolmogorov's theorem applies, but can be simplified by ordering the coordinates: it suffices to construct consistent marginal distributions for $(\pi_1, \ldots, \pi_k)$, for every $k = 1, 2, \cdots \in \mathbb{N}$. The way we construct these consistent marginal distribution is by **normalization** or **stick-breaking**.

1.4.2.1 Construction through normalization

Given nonnegative random variables $Y_1, Y_2, \ldots$ such that $\sum_{i=1}^{\infty} Y_i$ is positive and converges *almost surely*, we can define a prior on Δ by putting

$$\pi_k = \frac{Y_k}{\sum_{i=1}^{\infty} Y_i}, \qquad k \in \mathbb{N}$$

A simple, sufficient condition for the convergence of the random series is that $\sum_{i=1}^{\infty} \mathrm{E}(Y_i) < \infty$. Usually, for convenience, Y_i are assumed independent.

1.4.2.2 Construction through Stick-Breaking

Stick-breaking is a technique to construct a prior directly on Δ (Hjort et al. 2010, Ghosal and Vaart (2017)). The problem at hand is to distribute the total mass 1, which we identify with a stick of length 1, randomly to each element of $\mathbb{N}$. We first break the stick at a point given by the realization of a random variable $0 \leq V_1 \leq 1$ and assign mass V_1 to the point $1 \in \mathbb{N}$. Of course we have to choose to sample the V's from a suitable distribution: usually the $\mathrm{Beta}(a_1, b_1)$ is used. We think of the remaining mass $1 - V_1$ as a new stick, and break it into two pieces of relative lengths V_2 and $1 - V_2$ according to the realized value of another random variable $0 \leq V_2 \leq 1$. We assign mass $(1 - V_1)V_2$ to the point $2 \in \mathbb{N}$, and are left with a new stick of length $(1 - V_1)(1 - V_2)$. Continuing in this way, we assign mass to the point $k \in \mathbb{N}$ equal to

$$\pi_k = \left(\prod_{i=1}^{k-1} (1 - V_i) \right) V_k$$

Clearly, by continuing to infinity, this scheme will attach a random subprobability distribution to $\mathbb{N}$ for any sequence of random variables $V_1, V_2, \ldots$ with values in $[0, 1]$. Under mild conditions, the probabilities π_k will sum to one.

1.4.3 Construction through a Randomly Selected Discrete Set

A complementary approach is to construct priors through **structural definitions**, that is collect priors on measures on a general Polish space that are defined explicitly from mathematical theory. On example is the construction of a prior using randomly selected discrete sets in $\mathbb{N}$ (Ghosal and Vaart 2017).

Given an integer $k \in \mathbb{N} \cup \{\infty\}$, nonnegative random variables $\pi_1, \ldots, \pi_k$ with $\sum_{i=1}^{k} \pi_i = 1$ and random variables $\theta_1, \ldots, \theta_k$ taking their values in $(\Theta, \mathcal{B})$, we can

define a random probability measure by

$$P = \sum_{i=1}^{k} \pi_i \delta_{\theta_k}$$

The realizations of this prior are discrete with finitely or countably many support points, which may be different for each realization. Given the number, n, of support points, their "weights" $\pi_1, \ldots, \pi_k$ and their "locations" $\theta_1, \ldots, \theta_k$ are often chosen independent. An important special case is obtained by choosing $k = \infty$, yielding a prior of the form

$$P = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$$

Further specializations are to choose $\theta_1, \theta_2, \ldots$ an i.i.d. sequence in Θ, and to choose the weights $\pi_1, \pi_2, \ldots$ independently by the stick-breaking algorithm.

There are, of course, other ways to construct such priors. We stop here to explore in more detail how these methods are actually implemented in applied cases. In the next chapter we will explore a very common example of a prior on the space of probability measures: the *Dirichlet Process* starting with a very important use-case: density estimation.

Chapter 2

Dirichlet Processes

In this chapter we will review the definition, construction and properties of the Dirichlet Process: the "normal distribution of Bayesian nonparametrics" (Ghosal and Vaart, 2017). We will review the role of the Dirichlet process in density estimation problems and its role as nonparametric Bayesian prior.

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

$$y_i|G \sim G$$

To use the machineries of Bayesian inference, we need to complete the model with a prior on probability model, say Π – that as we will see is indeed the Dirichlet Process – for the unknown parameter G. A prior model on G requires the specification of a probability measure on an infinite-dimensional parameter, that is, the specification of a BNP prior.

2.1 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. Each draw from a Dirichlet process is itself a distribution. It is called a Dirichlet process because it has Dirichlet distributed finite dimensional marginal distributions

$$\text{Dir}(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}\prod_{i=1}^{K}x_i^{\alpha_i-1} \qquad x_i \in (0,1), \quad \sum_{i=1}^{K}x_i = 1, \quad \alpha_i > 0, \quad K \geq 2$$

Distributions drawn from a Dirichlet process are discrete *almost surely*, but cannot be described using a finite number of parameters.

The Dirichlet process (DP) is a stochastic process whose sample paths are **probability measures** with probability one, i.e. it is a distribution over probability measures in Θ. Thus draws from a DP can be interpreted as random distributions. For a distribution over probability measures to be a DP, its marginal distributions have to take on a specific form. In fact, for a random distribution G to be distributed according to a DP, its marginal distributions have to be Dirichlet distributed. Therefore, we can give the formal definition as follows

> **Dirichlet Process.** Let $\alpha > 0$ be a positive real number and G_0 be a probability measure (distribution) over Θ. A DP with parameters (α, G_0) is a random probability measure G defined on Θ which assigns probability $G(A)$ to every (measurable) set A such that for each (measurable) finite partition $\{A_1, \ldots, A_k\}$ of Θ, the joint distribution of the the vector $(G(A_1), \ldots, G(A_k))$ is the Dirichlet distribution with parameters $(\alpha G(A_1), \ldots, \alpha G(A_k))$.

Therefore, we say G is Dirichlet process distributed with base distribution G_0 and concentration parameter α, written $G \sim \text{DP}(\alpha, G_0)$ if

$$(G(A_1), \ldots, G(A_k)) \sim \text{Dir}\left(\alpha G_0(A_1), \ldots, \alpha G_0(A_k)\right)$$

for every finite measurable partition $A_1, \ldots, A_k$ of Θ. Using Kolmogorov's consistency theorem (Kolmogorov 1960), Ferguson (1973) showed that such a process exists. Furthermore, there are a number of approaches to establish existence that make use of powerful and general mathematical results to establish existence, and often require regularity assumptions on G_0 and Θ. One direct and elegant construction of the DP which need not impose such regularity assumptions is the **stick-breaking construction** (Sethuraman 1994) which we explore in the following chapters.

The interpretaion of the parameter is the following. The base distribution can be seen as the mean of the DP: for any measurable set $A \in \Theta$, $\mathrm{E}[G(A)] = G_0(A)$. On the other hand, the concentration parameter can be understood as an inverse variance: $V[G(A)] = G_0(A)(1 - G_0(A))/(1 + \alpha)$. The larger α, the smaller the variance, i.e. the DP will concentrate more of its mass around the mean.

2.1.1 Properties

An important property of the DP is the discrete nature of G. As a discrete random probability measure we can always write G as a weighted sum of point masses $G(\cdot) = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}(\cdot)$, where $\pi_1, \ldots, \pi_k$ are probability weights and $\delta_\theta(\cdot)$ denotes the Dirac measure at θ. One concern can be the coverage of DP within the class of all distributions over Θ. Samples from the DP are discrete, thus the set of distributions with positive probability under the DP is small. However it can be shown that if the topological support – the smallest closed set S in Θ with $G_0(S) = 1$ – of G_0 is Θ itself, then any distribution over Θ can be approximated arbitrarily accurately by a sequence of draws from $\mathrm{DP}(\alpha, G_0)$. In other words, another important property of the DP is its large weak support, which means that under mild conditions, any distribution with the same support as G_0 can be well approximated weakly by a DP random probability measure.

2.2 Stick-breaking construction

We have stated that draws from a DP can be written as a weighted sum of point masses, given its discrete nature. Sethuraman (1994) made this precise by providing a constructive definition of the DP as such, called the **stick-breaking** construction. This construction is also significantly more straightforward and general than mathematical proofs of the existence of DPs. It is simply given as follows: the locations θ_i are i.i.d. draws from the centering distribution G_0. Starting with a stick of length 1, each weight π_k is defined as a fraction of $(1 - \sum_{i<k} \pi_i)$ that is, a fraction of what is left of the stick after the preceding $k-1$ portions (point masses) are deducted. Formally,

$$\pi_k = v_k \prod_{i<k} (1 - v_i)$$

with $v_k \overset{i.i.d.}{\sim} \text{Beta}(1, \alpha)$, $\pi_1 = v_1$, and $\theta_k \overset{i.i.d.}{\sim} G_0$, where $\{v_k\}$ and $\{\theta_k\}$ are independent. Then

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

defines a $\text{DP}(\alpha, G_0)$. The stick-breaking distribution over $\pi = \pi_1, \pi_2, \ldots$ is sometimes written $\pi \sim \text{GEM}(\alpha)$, where the letters stand for Griffiths, Engen and McCloskey (Pitman 2002). A consequence of this representation is that if $G \sim \text{DP}(\alpha, G_0)$, $\theta \sim G_0$, and $\pi \sim \text{Beta}(1, \alpha)$, and all of them are independent, then $\pi\delta_\theta(\cdot) + (1 - \pi)G(\cdot)$ follows again $\text{DP}(\alpha, G_0)$.

Finally, the DP has an important conditioning property that can be shown to following immediately from the definition. If A is a (measurable) set with $G_0(A) > 0$ (which implies that $G(A) > 0$ a.s.), then the random measure $G|_A$, i.e. the restriction of G to A defined by $G|_A(B) = G(B|A)$ is also a DP with parameters α and $G_0|_A$, and is independent of $G(A)$. The argument can be extended to more than one set. Thus the DP *locally* splits into numerous independent DP's.

2.2.1 Posterior Distribution

Let $G \sim \mathrm{DP}(\alpha, G_0)$. Since G is a (random) distribution, we can in turn draw samples from G itself. Let $\theta_1, \ldots, \theta_n$ be a sequence of independent draws from G. Note that the θ_i's take values in Θ since G is a distribution over Θ. We are interested in the posterior distribution of G given observed values of $\theta_1, \ldots, \theta_n$. Let $A_1, \ldots, A_k$ be a finite measurable partition of Θ, and let n_k be the observed number of θ_i's in partition A_k. By the conjugacy between the Dirichlet and the multinomial distributions, we have:

$$G(A_1), \ldots, G(A_k) \mid \theta_1, \ldots, \theta_n \sim \mathrm{Dir}\Big(\alpha G_0(A_1) + n_1 \ , \ \ldots, \ \alpha G_0(A_k) + n_k\Big)$$

Since the above is true for all finite measurable partitions, the posterior distribution over G must be a DP as well. A little algebra shows that the posterior DP has updated concentration parameter $\alpha + n$ and base distribution $\frac{\alpha G_0 + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}$, where δ_i is a point mass located at θ_i and $n_k = \sum_{i=1}^n \delta_i(A_k)$. In other words, the DP provides a conjugate family of priors over distributions that is closed under posterior updates given observations. Rewriting the posterior DP, we have:

$$G|\theta_1, \ldots, \theta_n \sim \mathrm{DP}\left(\alpha + n, \ \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} \frac{\sum_{i=1}^n \delta_{\theta_i}}{n}\right)$$

Notice that the posterior base distribution is a weighted average between the prior base distribution G_0 and the empirical distribution $\frac{\sum_{i=1}^n \delta_{\theta_i}}{n}$. The weight associated with the prior base distribution is proportional to α, while the empirical distribution has weight proportional to the number of observations n. Thus we can interpret α as the strength or mass associated with the prior. In other words, under the sampling model $\theta_i | G \overset{i.i.d.}{\sim} G$ with a DP on G, the posterior distribution for G is again a DP. The base measure of the posterior DP adds a point mass to the prior base measure at each observed data point θ_i: the centering measure of the posterior DP is a weighted average of G_0 and the empirical distribution $\sum_{i=1}^n \delta_{\theta_i}/n$ and the posterior total mass parameter is incremented to $\alpha + n$ (Ferguson 1973).

Taking $\alpha \to 0$, the prior becomes non-informative in the sense that the predictive distribution is just given by the empirical distribution. On the other hand, as the

amount of observations grows large, $n \gg \alpha$, the posterior is simply dominated by the empirical distribution which is in turn a close approximation of the true underlying distribution. This gives a consistency property of the DP: the posterior DP approaches the true underlying distribution.

2.2.2 Predictive Distribution

A key property, as said, of the DP prior is its a.s. discreteness. Consider a random sample, $\theta_i|G \sim G$, i.i.d., $i = 1, \ldots, n$. The discreteness of G implies a positive probability of ties among the θ_i. This is at the heart of the **Polya urn representation** of Blackwell and MacQueen (1973). In other words, the predictive distribution of the observations is given by the Polya urn scheme. The name *Polya urn* stems from a metaphor useful in interpreting the marginal distribution. Specifically, each value in Θ is a unique color, and draws $\theta \sim G$ are balls with the drawn value being the color of the ball. In addition we have an urn containing previously seen balls. In the beginning there are no balls in the urn, and we pick a color drawn from G_0, i.e. draw $\theta_1 \sim G_0$, paint a ball with that color, and drop it into the urn. In subsequent steps, say the $n+1$st, we will either, with probability $\frac{\alpha}{\alpha+n}$, pick a new color – that is draw $\theta_{n+1} \sim G_0$ – paint a ball with that color and drop the ball into the urn, or, with probability $\frac{n}{\alpha+n}$, reach into the urn to pick a random ball out – draw θ_{n+1} from the empirical distribution – paint a new ball with the same color and drop both balls back into the urn. This scheme has been used to show the existence of the DP (Blackwell and MacQueen 1973).

Denote $\theta_1^\star, \ldots, \theta_{k_n}^\star$ the k_n unique values among the n observations in the sample $\theta_1, \ldots, \theta_n$ generated as $\theta|G \overset{i.i.d.}{\sim} G$ and $G \sim \mathrm{DP}(\alpha, G_0)$. Then

$$\theta_1 \sim G_0$$

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha+n} G_0 + \frac{1}{\alpha+n} \sum_{i=1}^{k_n} n_i \, \delta_{\theta_i^\star}$$

where $n_i = \sum_{i=1}^{n} \mathbf{1}(\theta_i = \theta_i^\star)$ is the number of observations that are equal to the j-th unique value.

Since the values of draws are repeated, let $\theta_1^\star, \ldots, \theta_m^\star$ be the unique values among

$\theta_1, \ldots, \theta_n$, and n_k be the number of repeats of θ_k. The unique values of $\theta_1, \ldots, \theta_n$ – say, $n_1 = 3$ means that in the sample there are 3 $\theta_i = \theta_1^\star$ – induce a partitioning of the set $[n] = \{1, \ldots, n\}$ into clusters such that within each cluster, say cluster k, the θ_i's take on the same value $\theta_k^\star$. Given that $\theta_1, \ldots, \theta_n$ are random, this induces a random partition of $[n]$. The distribution over partitions is called the Chinese restaurant process (CRP). In this metaphor we have a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers. The first customer enters the restaurant and sits at the first table. The second customer enters and decides either to sit with the first customer, or by herself at a new table. In general, the n + 1st customer either joins an already occupied table k with probability proportional to the number n_k of customers already sitting there, or sits at a new table with probability proportional to α. Identifying customers with integers $1, 2, \ldots$ and tables as clusters, after n customers have sat down the tables define a partition of $[n]$ with the distribution over partitions being the same as the one above. The fact that most Chinese restaurants have round tables is an important aspect of the CRP. This is because it does not just define a distribution over partitions of $[n]$, it also defines a distribution over permutations of $[n]$ with each table corresponding to a cycle of the permutation (P. Muller and Quintana 2004).

Therefore, we can rewrite the Polya urn sampling scheme in a different way. The observations $(\theta_1, \ldots, \theta_n)$ can be equivalently parametrized in terms of the independent vectors $(s_1, \ldots, s_n)$ and $\theta_1^\star, \ldots, \theta_m^\star$ where

$$s_1 \sim \delta_1$$

meaning that $p(s_1) = 1$ if $s_1 = 1$ or 0 otherwise.

$$s_{n+1} | s_1, \ldots, s_n \sim \frac{\alpha}{\alpha + n} \delta_{n+1} + \frac{1}{\alpha + n} \sum_{k=1}^{m} n_k \, \delta_k$$

$$\theta_k^\star \overset{i.i.d.}{\sim} G_0, \qquad k = 1, \ldots, m$$

and $\theta_i = \theta_k^\star$ if $s_i = k$.

2.3 Applications

DPs are used across a wide variety of applications of Bayesian analysis in both statistics and machine learning. The main 3 examples are and most prevalent applications include:

2.3.1 Bayesian model validation

How does one validate that a model gives a good fit to some observed data? The Bayesian approach would usually involve computing the marginal probability of the observed data under the model, and comparing this marginal probability to that for other models. If the marginal probability of the model of interest is highest we may conclude that we have a good fit. The choice of models to compare against is an issue in this approach, since it is desirable to compare against as large a class of models as possible. The Bayesian nonparametric approach gives an answer to this question: use the space of all possible distributions as our comparison class, with a prior over distributions. The DP is a popular choice for this prior, due to its simplicity, wide coverage of the class of all distributions, and recent advances in computationally efficient inference in DP models. The approach is usually to use the given parametric model as the base distribution of the DP, with the DP serving as a nonparametric relaxation around this parametric model. If the parametric model performs as well or better than the DP relaxed model, we have convincing evidence of the validity of the model.

2.3.2 Density estimation

Another application of DPs is in density estimation. Here we are interested in modeling the density from which a given set of observations is drawn. To avoid limiting ourselves to any parametric class, we may again use a nonparametric prior over all densities. Here again DPs are a popular. However The DP generates distributions that are discrete with probability one, making it awkward for continuous density estimation. This limitation can be fixed by *convolving* its trajectories with some *continuous kernel*, or more generally, by using a DP random

measure as the mixing measure in a mixture over some simple parametric forms. Such an approach was introduced by Ferguson (1973). The mixture model is the object of the next Chapter.

Chapter 3

Nonparametric Regression

In this chapter we will focus on regression problems and how they are approached from a Bayesian nonparametric perspective. In particular, we will provide the definition of mixture models and their Bayesian treatment, with and without dependence on covariates. In particular, we will discuss the Dependent Dirichlet Process prior. Finally, we will give the intuition behind the use of Dirichlet Process mixture models in settings where it is reasonable to introduce dependence among different samples and the ideas related to "borrowing of strenght".

Consider a generic regression problem with dependent variable y_i, covariates x_i, $i = 1, \ldots, n$, and an assumed model $y_i = f(x_i) + \varepsilon_i$ with $\varepsilon_i \sim p_\varepsilon(\varepsilon_i)$. As long as both, the regression function $f(\cdot)$ and the residual distribution $p(\cdot)$, are indexed by finitely many parameters, inference reduces to a traditional parametric regression problem. The problem becomes a nonparametric regression when the investigator wants to relax the parametric assumptions of either of the two model elements. This characterization of nonparametric regression allows for three cases.

Nonparametric Residuals. The model can be generalized by going nonparametric on the residual distribution, assuming $\varepsilon|G \overset{i.i.d.}{\sim} G$ with a nonparametric prior $p(G)$ on G, while keeping the regression mean function parametric as $f(\cdot) = f_\theta(\cdot)$ indexed by a (finite-dimensional) parameter vector θ with prior π.

We refer to this case as a *nonparametric error model*. Essentially this becomes density estimation for the residual error. Of course the residuals ε_i are not usually observable. Hence, the problem reduces to one of density estimation conditional on assumed values for the parameters θ. In principle, any model that was used for density estimation could be used. However, there is a minor complication. To maintain the interpretation of ε as residuals and to avoid identifiability concerns, it is desireable to center the random G at zero, for example, with $E(G) = 0$ or median 0.

Nonparametric Mean Function. One could, instead, relax the parametric assumption on the mean function and complete the model with a nonparametric prior $f(\cdot) \sim p(f)$. We refer to this as a *nonparametric regression mean function*. Popular choices for $p(f)$ are Gaussian process priors or priors based on basis expansions, such as wavelet based priors or neural network models. This approach, however, is limited in the sense that it only allows for flexibility in the mean. Many datasets present non normality or multi-modality of the errors, degrees of skewness, or tail behavior in different regions of the covariate space. To capture such behavior, a flexible approach for modeling the conditional density that allows both the mean and error distribution to evolve flexibly with the covariates is required.

Fully Nonparametric Regression. One could go nonparametric on both assumptions. We refer to this as a *fully nonparametric regression*. The sampling model becomes $p(y_i|x_i) = G_x$, with a prior on the family of conditional Random Probability Measures, $p(G_x, x \in \mathrm{X})$. Many commonly used BNP priors for $\mathcal{G} = \{G_x\}$ are variations of dependent DP priors. In the next section we will review mixture model, a useful way to face this task.

3.1 Mixture models

For independent and identically distributed data, **mixture models** are an extremely useful tool for flexible density estimation due to their ability to approximate a large class of densities and their attractive balance between smoothness and flexibility in modeling local features. The canonical form of a mixture model

is the following

$$p(y|G) = \int K(y;\theta)\, dG(\theta)$$

where G is a probability measure on the parameter space Θ, Y is the sample space, and $K(y;\theta)$ is a *kernel* on $\mathrm{Y} \times \Theta$. The kernel, $K(y;\theta)$, is defined by

- $K(\cdot;\theta)$ is a density on Y with respect to the Lebesgue measure and
- $K(y;\theta)$ is a measurable function of θ, where $Theta$ is assumed to be a complete and separable metric space and equipped with its Borel σ-algebra

That is K is a Markov Kernel (see Chapter 1). In a Bayesian setting, this model is completed with a prior distribution on the mixing measure G taking values in $\mathrm{M}(\Theta)$ where $\mathrm{M}(\Theta)$ denotes the set of probability measures on Θ. A common prior choice takes G as a discrete random measure with probability one. In this case, G has the following representation almost surely

$$G = \sum_{k=j}^{J} \pi_j \delta_{\theta_j}$$

for some random atoms θ_j taking values in Θ and weights π_j such that $\pi_j \geq 0$ and $\sum_j pi_j = 1$ almost surely. The mixture model can then be, thus, expressed as a convex combination of kernels

$$p(y|G) = \sum_{k=j}^{J} \pi_j K(y;\theta_j)$$

Now we would like to make this problem of density estimation covariate-dependent. In general, the model may be extended in one of two ways (Wade 2013).

The first approach is closely related to classical kernel regression methods and involves augmenting the observed data to include the covari-ates. The joint density is modelled as

$$p(y,x|G) = \sum_{k=j}^{J} \pi_j K(y,x;\theta_j)$$

and the conditional density estimate is obtained as, from Wade (2013),

$$p(y|x, G) = \frac{\sum_{j=1}^{J} \pi_j K(y, x; \theta_j)}{\sum_{j'=1}^{J} \pi_{j'} K(x; \theta_{j'})}$$

However, this approach requires the modelling of the marginal of X, and usually we take X as deterministic and thus we do not need to model it.

The second approach overcomes this by directly modelling the covariate-dependent density. In this case the mixture model is extended by allowing the mixing distribution G to depend on x. Hence, for every $x \in \mathrm{X}$

$$p(y, x|G_x) = \int K(y, x; \theta)\ G_x(\theta)$$

the Bayesian model is completed by assigning a prior distribution on the family $\mathcal{G} = \{G_x, x \in \mathrm{X}\}$ of covariate-dependent mixing probability measures. The realizations of $\mathcal{G}$ are in $\mathrm{M}(\Theta)^{\mathrm{X}}$. If the prior gives probability one to the set of discrete probability measures, then (a.s.)

$$G_x = \sum_{j=1}^{J} \pi_j(x)\ \delta_{\theta_j(x)}$$

and

$$p(y, x|G_x) = \sum_{j=1}^{J} \pi_j(x)\ K\Big(y, x; \theta_j(x)\Big)$$

where $\theta_j(x)$ takes values in Θ and the weights $\pi_j(x)$ are such that $\pi_j(x) \geq 0$ and $\sum_j \pi_j(x) = 1$ (a.s.) for all $x \in \mathrm{X}$. The number of mixture components, J, plays a key role in the flexibility of the model. Finite mixtures are defined with $J < \infty$ and they are known as *mixture of experts* in machine learning literature.

However, they require either the choice of J, which in practice is chosen through post-processing techniques, or, in Bayesian setting, a prior on J, which requires posterior sampling of J. Instead, nonparametric mixtures define $J = \infty$. The general models described above, substituting with $J = \infty$ are the starting point for Bayesian nonparametric mixture models for regression. The models are completed with a definition of the kernel and a prior choice for the weights and atoms.

The choice of an appropriate kernel, $K(\cdot; \cdot)$, depends on the underlying sample

space (P. Muller and Quintana 2004). If the underlying density function is defined on the entire real line, a location-scale kernel is appropriate. On the unit interval, beta distributions is a flexible choice. On the positive half line, mixtures of gamma is sensible. The use of a uniform kernel leads to random histograms. Petrone and Veronese (2002) motivated a canonical way of viewing the choice of a kernel through the notion of a Feller sampling scheme, and called the resulting prior a Feller prior.

3.2 Dirichlet Process Mixture Models

The Dirichlet process is commonly chosen as the prior for the mixing measure. The mixture model above, together with a DP prior on the mixing measure G, can equivalently be written as a hierarchical model. Here the nonparametric nature of the Dirichlet process translates to mixture models with a countably infinite number of components. We model a set of observations $\{y_1, \ldots, y_n\}$ using a set of *latent* parameters $\{\theta_1, \ldots, \theta_n\}$. Each θ_i is drawn independently and identically from G, while each y_i has distribution $p(y_i|\theta_i)$ parametrized by θ_i. Assume $y_i|G \overset{i.i.d.}{\sim} p(y|G)$ as above, then the equivalent hierarchical model is

$$y_i|\theta_i \overset{ind}{\sim} p(y_i|\theta_i)$$

$$\theta_i|G \overset{i.i.d.}{\sim} G$$

$$G|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$$

The hierarchical model introduces new latent variables θ_i specific to each observation. Integrating out the $(\theta_1, \ldots, \theta_n)$, we have that given G, the y_i are independent with density

$$p(y|G) = \int_\Theta K(y;\theta)\, dG(\theta) = \sum_{j=1}^{\infty} \pi_j\, K(y;\theta_j)$$

where $K(\cdot;\theta)$ is the density of $p(\cdot|\theta)$. Under this hierarchical model, the posterior distribution on G, $p(G|y)$ is a mixture of DP's, mixing with respect to θ_i, that is

$$G|y \sim \int \mathrm{DP}\left(\alpha+n,\ \frac{\alpha}{\alpha+n}G_0 + \frac{n}{\alpha+n}\frac{\sum_{i=1}^n \delta_{\theta_i}}{n}\right) dG(\theta|y)$$

where $\theta = (\theta_1, \ldots, \theta_n)$ and $y = (y_1, \ldots, y_n)$. Therefore, marginalizing with respect to θ, the posterior given y becomes a mixture over the posterior DP (given θ) with respect to the posterior distribution on θ.

3.3 Dependent Dirichlet Process

Many applications call for more than one random probability measure G. The generic regression problem of predicting an outcome y conditional on a covariate x could be described as inference for the conditional distributions $G_x(\cdot) = p(y_i|x_i = x)$ for $x \in \mathrm{X}$ – when $p(y_i|x_i = x)$ is indexed by finitely many parameters we are back to parametric. In our case, the problem becomes one of inference for a family of random probability measures $\mathcal{G} = \{G_x, x \in \mathrm{X}\}$, indexed by the covariates x. We thus need a BNP prior $p(\mathcal{G}) \stackrel{def}{=} p(G_x; x \in \mathrm{X})$ for the entire family. In the application to nonparametric regression as well as many other applications it is natural to require that G_x be dependent across x. Surely we would not expect G_x to change substantially for minor changes of x.

Perhaps the most popular prior model for a *family* of random probability measures is the dependent DP (DDP). It was originally introduced by maceachern99, with many variations defined in later papers. The basic idea is simple. We say that $\mathcal{G}$ is a dependent DP (DDP) if, for for every $x \in \mathrm{X}$ we can write the following.

3.3.1 Covariate-dependent atoms

Start with the stick breaking construction of a DP random probability measure

$$G_x = \sum_{j=0}^{\infty} \pi_j\ \delta_{\theta_j(x)}$$

with point masses at locations $\theta_j(x) \sim G_x$ and weights $\pi_j = v_j \prod_{i<j}(1 - v_i)$ for i.i.d. beta fractions $v_j \overset{i.i.d.}{\sim} \text{Beta}(1, \alpha)$. By the following construction, the model can be generalized to a joint prior for $\mathcal{G}$, keeping a DP prior as the marginal for G_x, for every $x \in \text{X}$, but introducing the desired dependence across x.

To ensure the marginal DP prior we have to keep the i.i.d. prior on $\theta_j(x)$ across j. But we are free to introduce dependence of $\theta_j(x)$ across x (for every j). The simple, yet powerful idea of the DDP construction is to introduce dependence over x, i.e., to link the G_x through *dependent locations* of the point masses: let $\theta_j = (\theta_j(x),\ x \in \text{X})$ denote the family of random variables $\theta_j(x)$ for fixed j that are mutually independent, that is given j, θ_j is a stochastic process indexed by x.

Implicit in the notation used in the formula above is the definition of weights π_j that are common across x – this variation of the DDP model is sometimes referred to as "common weight" or "single π" DDP which features only **covariate dependent atoms**. The regression model can be, thus, written as

$$p(y|x, G_x) = \sum_{j=1}^{\infty} \pi_j \ K\Big(y, x; \theta_j(x)\Big)$$

$$G_x = \sum_{j=1}^{\infty} \pi_j \ \delta_{\theta_j(x)}$$

For continuous covariates and a continuous response, one of the most popular choices for the kernel of the single-p DDP model is the Gaussian distribution

$$p(y|x, G_x) = \sum_{j=1}^{J} \pi_j \ \mathcal{N}\left(y; \mu_j(x), \sigma_j^2\right)$$

where $\mu(\cdot)$ are independent Gaussian processes with a mean function of $m(\cdot)$ and covariance function of $c(\cdot, \cdot)$, denoted by $\text{GP}(m, c)$. For a heteroskedastic model, also the variance can be made dependent on x.

3.3.2 Covariate-dependent weights

In general, the weights could have an additional x index, defining

$$G_x = \sum \pi_j(x)\ \delta_{\theta_j(x)}(x)$$

We write that $\mathcal{G}$ is distributed according to a dependent Dirichlet Process as follows

$$\mathcal{G} \sim \text{DDP}(\alpha, S)$$

where S is the law of the stochastic process governing the θ_j's given j. For example, S could be a Gaussian process with index set X.

The main constraint in this case is given by the need to specify a prior such that $\sum_j \pi_j(x) = 1$ for all $x \in \mathcal{X}$. The technique used to explicitly define $\pi_j(x)$ and satisfy this constraint is based on the stick-breaking representation:

$$\pi_1(x) = v_1(x)$$

$$\pi_j = v_j(x) \prod_{i<j} \Big(1 - v_i(x)\Big)$$

where $0 \leq v_j(x) \leq 1$ a.s. for all j and x.

3.3.3 Example

Suppose that there are finitely many dependent random probability measures $\mathcal{G} = \{G_j, j = 1, \ldots, J\}$ that are judged to be **exchangeable**, i.e., the prior model $p(\mathcal{G})$ should be invariant with respect to any permutation of the indices. This case could arise, for example, as a prior model for unknown random effects distributions G_j in related studies, $j = 1, \ldots, J$. In words, the DDP permits to define a prior probability model $p(\mathcal{G})$ that allows us to *borrow strength* across the J studies: e.g. patients under study j_1 should inform inference about patients enrolled in another related study $j_2 \neq j_1$. Two extreme modeling choices would be

1. To pool all patients and assume one common random effects distribution: $G_j \stackrel{def}{=} G, j = 1, \ldots, J$ with a prior $p(\mathcal{G})$

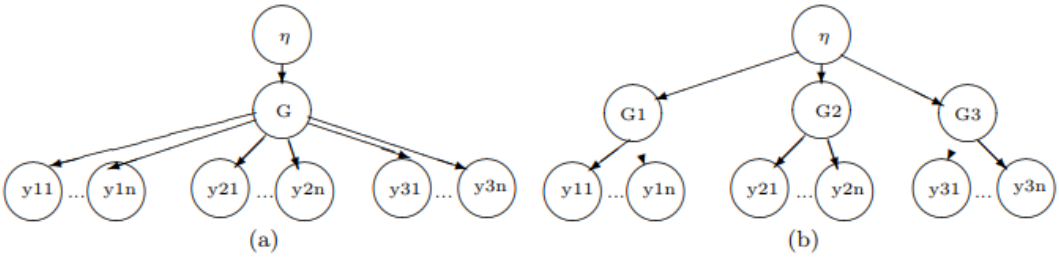

Figure 3.1: One common RPM G (panel a) versus distinct RPMs Gj , independent across studies (panel b).

2. To assume J distinct random effects distributions with independent priors: $G_j \sim p(G_j)$, independently, $j = 1, \ldots, J$

These two choices are opposite since the first choice implies maximum borrowing of strengths, and the other choice implies no borrowing of strength. In most applications, the desired level of borrowing strength is somewhere in-between these two extremes, exemplified in the following figure.

Note that in the figure[1] we added a hyperparameter η to index the prior model $p(G_j|\eta)$ and $p(G|\eta)$, which was implicitly assumed fixed. The use of a random hyperparameter η allows for *some* borrowing of strength even in the case of conditionally independent $p(G_j|\eta)$. Learning across studies can happen through learning about the hyperparameter η. This kind of construction can be found in Muliere and Petrone (1993) and Mira and Petrone (1997). However, the nature of the learning across studies is determined by the parametric form of η.

[1]Taken from *NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 9, Institute of Mathematical Statistics and American Statistical Assocation, 2013.*

References

Blackwell, D., and J. B. MacQueen. 1973. "Ferguson Distributions via Polya Urn Schemes." *Ann. Statist.* 1: 353–55.

Ferguson, Thomas S. 1973. "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics* 1 (2): 209–30.

Ghosal, Subhashis, and Aad van der Vaart. 2017. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Hjort, Nils Lid, Chris Holmes, Peter Muller, and Stephen G. Walker. 2010. *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Kolmogorov, Andrey N. 1960. *Foundations of the Theory of Probability*. 2nd ed. Chelsea Pub Co.

Mira, Antonietta, and Sonia Petrone. 1997. "Bayesian Hierarchical Nonparametric Inference for Change-Point Problems." *Bayesian Statistics* 5 (March).

Muliere, Pietro, and Sonia Petrone. 1993. "A Bayesian Predictive Approach to Sequential Search for an Optimal Dose: Parametric and Nonparametric Models." *Journal of the Italian Statistical Society* 2 (3): 349–64.

Muller, P., and F. A. Quintana. 2004. "Nonparametric Bayesian Data Analysis." *Statistical Science* 19: 95–110.

Orbanz, Peter. 2014. *Lecture Notes on Bayesian Nonparametrics*.

Petrone, Sonia, and Piero Veronese. 2002. "Non Parametric Mixture Priors Based on an Exponential Random Scheme." *Statistical Methods and Applications* 11 (1): 1–20.

Pitman, Jim. 2002. "Combinatorial Stochastic Processes."

Sethuraman, Jayaram. 1994. "A Constructive Definition of Dirichlet Priors."

Statistica Sinica 4: 639–50.

Wade, Sara Kathryn. 2013. "Bayesian Nonparametric Regression Through Mixture Models."