Chapter 3

Bayesian Nonparametrics

In this chapter we review general notions of Bayesian nonparametric procedures. Rather than assuming an inferential perspective, we put attention to the manipulation of marginal and conditional probabilities involving random probability measures. The content of this chapter has been extracted from Ferguson (1973, 1974), Walker, Damien, Laud, and Smith (1999) and Lijoi and Prünster (2010), among other sources.

3.1 Introduction

Bayesian nonparametric methods extend the scope of traditional Bayesian methods by dealing with the problem of placing priors over function (typically infinite dimensional) spaces. An attractive feature of nonparametric procedures, from the inferential viewpoint, is that they constitute a robust statistical procedure that overcome many limitations implicitly associated with parametric alternatives, in terms of inference and prediction.

In this section we review a justification of Bayesian nonparametric procedures arising from the exchangeable framework. For that, let us assume that there is an (hypothetical) infinite sequence of (observable) exchangeable random variables, $\{Y_j\}_{j=1}^{\infty}$, and a probability measure Π associated with the whole sequence, such that the joint probability law

of any finite collection of random variables, $Y_1, \ldots, Y_n$, can be decomposed as

$$\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n) = \int \prod_{j=1}^{n} F(A_j)\Pi(\mathrm{d}F), \tag{3.1}$$

for any measurable sets $(A_1, \ldots, A_n)$. In the above representation F can be understood as an unknown and random element connecting the Y_i's and Π as it associated probability law. The above representation of the probability law for an (infinite) sequence of exchangeable random variables is due to de Finetti (1937) and Hewitt and Savage (1955).

In the above representation, the probability measure Π can be seen in two different ways. From an objective point of view Π represents some empirical measure associated with F. That point of view has been the one deriving the original result for $0-1$ sequences of exchangeable random variables.

From a subjective point of view, Π can be seen as one's prior beliefs concerning F (see, de Finetti, 1937). So, the above representation has served as a justification to implement Bayesian inferential procedures under the assumption of exchangeability. Two particular frameworks emerge from such a conception: parametric and nonparametric procedures. Parametric procedures are characterized by manifesting one's prior beliefs around a specific parametric functional form for F. Accordingly, the probability measure Π will put probability mass one to such a family of distribution functions and one's prior beliefs concerning F will be reduced to place a prior distribution on the space of indexing parameters associated with the chosen parametric family.

A suitable approach to randomize F, which is the one we shall consider throughout this thesis, consists in associating a stochastic process, say Z, such that after a suitable transformation, its sample paths satisfy the requirements to be a probability distribution function. Hence F can be defined as the transformed Z. In that case, Π represents the probability law driven the process Z. See Ferguson (1974), Doksum (1974) and Walker, Damien, Laud, and Smith (1999). Under this approach, F is considered as a random probability measure and $\mathcal{F}$ corresponds to the space of probability measures where F

may take values.

In Section 3.2 we stablish some notation and review the notion of random probability measures. Section 3.3 revises the definition and some properties of Dirichlet processes, which up to date is the most frequently random probability measure used in the Bayesian framework. Sections 3.6 and 3.5 review some properties of extended families of random distributions functions, termed neutral-to-the-right and stick-breaking processes, respectively. In Section 3.7 we review some notions of nonparametric mixture models and summarize some of their properties. Section 3.8 concludes with a brief discussion.

3.2 Random probability measures

In this section we review the notion of random probability measures. For that, let us introduce some notation and convention. Let us consider a general probability space, $(\Omega, \mathscr{B}_\Omega, \mathbb{P})$. Notice that all the definitions we are about to review are defined on that probability space. Let $(\mathcal{Y}, \mathscr{B}_\mathcal{Y})$ be an arbitrary (Polish) measurable space of observable quantities. [A Polish space is a separable completely metrizable topological space. Refer to Crauel (2002) for some probabilistic insights concerning random probability measures.]. Let $\mathcal{F}$ be a space of probability measures on $(\mathcal{Y}, \mathscr{B}_\mathcal{Y})$. As mentioned above, the essential assumption in Bayesian nonparametric statistics is that any element of $\mathcal{F}$, let us say F, is random. Here is where the notion of random probability measures becomes relevant for Bayesian statistics.

Definition 3.2.1. (Aldous, 1985) Let F be a measure on $(\mathcal{Y}, \mathscr{B}_\mathcal{Y})$, it is said that F is a random measure if there is random variable, α, such that for each $\omega \in \Omega$, F assigns a measure $\alpha(\omega, A)$ to any subset $A \in \mathscr{B}_\mathcal{Y}$. In addition, if F is such that

i) $F(\cdot) = \alpha(\omega, \cdot)$ is a probability measure, for any $\omega \in \Omega$,

ii) $F(A) = \alpha(\cdot, A)$ is a $[0, 1]$-valued random variable, for any $A \in \mathscr{B}_\mathcal{Y}$,

then it is said that F is a random probability measure on $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$. Therefore, it is natural to conceive the class $\mathcal{F}$ as the collection of $[0,1]$-valued random variables

$$\mathcal{F} = \{F(A) = \alpha(\omega, A) : A \in \mathscr{B}_{\mathcal{Y}}\}.$$

Also, we can equip the class $\mathcal{F}$ with its corresponding σ-field, hereafter denoted by $\mathscr{B}_{\mathcal{F}}$, which is the one generated by the maps

$$F \to F(A); \qquad \text{for any } A \in \mathscr{B}_{\mathcal{Y}}, \tag{3.2}$$

such that $F(A) = \alpha(\omega, A)$ for some $\omega \in \Omega$. That means, $\mathscr{B}_{\mathcal{F}}$ is the collection of all measurable mappings from $\mathscr{B}_{\mathcal{Y}}$ into the unit interval.

The reader should notice the resemblance between the above definition of random probability measures and probability kernels stated in Section 2.3. Anticipating a proper definition, let us assume that there is a probability measure Π defined on the space $\mathcal{F}$. Combining the idea of seeing F as a probability kernel and assuming the existence of Π gives us the intuition to extend the definition of a random probability measure to the product space $\mathcal{Y} \times \mathcal{F}$ with a probability measure given by

$$\mathbb{P}(B) = \int_{\mathcal{Y}\times\mathcal{F}} \mathbb{1}_B(y, F) F(\mathrm{d}y)\Pi(\mathrm{d}F), \tag{3.3}$$

for some probability measure for any B measurable subset of $\mathcal{Y}\times\mathcal{F}$. Notice that assuming that both spaces $\mathcal{Y}$ and $\mathcal{F}$ are Polish we obtain $\mathscr{B}_{\mathcal{Y}\times\mathcal{F}} = \mathscr{B}_{\mathcal{Y}} \otimes \mathscr{B}_{\mathcal{F}}$.

According to the above definition, a random probability measure can be specified by a system of all finite dimensional distributions defined for the collection of disjoint sets $(A_1, \ldots, A_n)$ of $\mathscr{B}_{\mathcal{Y}}$. In order to do so, some additional conditions are required (see, e.g. Ferguson, 1973; Balan, 2004):

1. **Finitely additive property.** For any collection of disjoint sets $(A_j)_{j=1}^n$ of elements in $\mathscr{B}_{\mathcal{Y}}$ and for any $1 \leq i_1 < \cdots, i_m \leq k$, F should be that

$$\left\{F\left(\cup_{j=1}^{i_1} A_j\right), ..., F\left(\cup_{j=i_m}^{n} A_j\right)\right\} \stackrel{d}{=} \left\{\sum_{j=1}^{i_1} F(A_j), ..., \sum_{j=i_m}^{n} F(A_j)\right\},$$

2. **Total mass property.** $F(\mathcal{Y}) = 1$, a. s.[$\mathbb{P}$],

3. **Countably additivity property.** For every decreasing sequence $\{A_n\}_{n=1}^{\infty}$ in $\mathscr{B}_{\mathcal{X}}$ such that $\bigcap_{n=1}^{\infty} A_n = \emptyset$, follows that

$$\lim_{n\to\infty} F(A_n) = 0, \text{ a. s.}[\mathbb{P}].$$

According to the above definition and properties, a random probability measure can be completely specified by the system of all finite dimensional random entities of the form $\{F(A_1), \ldots, F(A_k)\}$ for any unordered collection of disjoint sets $(A_1, \ldots, A_k)$ contained in $\mathscr{B}_{\mathcal{Y}}$. Even more, (Ferguson, 1973) developed F not through any collection of disjoint sets contained in $\mathscr{B}_Y$, but through the collection of all measurable partitions of $\mathcal{Y}$. Such a construction gives rise to the well known Dirichlet process, which we shall briefly describe in the next section where the discussion focuses on the role of random probability measures with (a. s.) discrete paths.

3.2.1 Discrete probability measures

Before proceeding, let us anticipate that the first notion of random probability measure on $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$ we shall be dealing with takes the form of a discrete probability measure (other types of probability measures will be explored later) admitting the following representation

$$F(A) = \sum_{j=1}^{\infty} W_j \, \delta_{Z_j}(A), \tag{3.4}$$

for any $A \in \mathscr{B}_{\mathcal{Y}}$; where $\{W_j\}_{j=1}^{\infty}$ is a sequence of probability weights and $\{Z_j\}_{j=1}^{\infty}$ is a sequence locations taking values in $\mathcal{Y}$. The random version of F would consider the sequences $\{W_j\}_{j=1}^{\infty}$ and $\{Z_j\}_{j=1}^{\infty}$ to be random, with the additional assumption of being stochastically independent, and such that

$$0 \leq W_j \leq 1 \quad \text{and} \quad \sum_{j=1}^{\infty} W_j = 1, \text{ a. s.}. \tag{3.5}$$

It is commonly assumed that the random locations (Z_j) are i. i. d. according to a given distribution function, say Q. What remains as a challenge in the above specification is a proper definition of the random weights (W_j) in order to satisfy the summability condition (3.5). The Dirichlet process and related processes we shall review in the next sections define random probability measures satisfying such a summability condition.

But, let us notice that the notion of random probability measures can also be extended to absolute continuous distributions, by means of nonparametric mixture distributions, as we shall review in Section 3.7.

Let us for now proceed with the definition of the Dirichlet process.

3.3 Dirichlet process

Since its appearance in the literature, the Dirichlet process has played the role of the archetype of random probability measure used in the Bayesian nonparametric framework. Its importance is derived from the simplicity of its formulation and tractability to implement updating inferential procedures. This section is devoted to review some of its most important properties.

Definition 3.3.1. (Ferguson, 1973) Let $\mathcal{Y}$ be a Polish space endowed with its corresponding Borel σ-field $\mathscr{B}_{\mathcal{Y}}$, and let γ be a non-null finite measure on $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$. A random probability measure F on $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$ is called a Dirichlet process with parameter α if for every measurable finite partition of $\mathcal{Y}$, namely $A_1, \ldots, A_n$, the joint distribution of $F(A_1), \ldots, F(A_n)$ is a n-dimensional Dirichlet distribution with parameters $\gamma(A_1), \ldots, \gamma(A_n)$.

Ferguson (1973) showed the existence of such a measure and its consistency with respect to the properties described in the previous section.

3.3.1 Some properties

From Definition 3.3.1 it follows straightforwardly that for any measurable set A of $\mathcal{X}$,

$$\mathbb{E}_{\mathrm{DP}}[F(A)] = \frac{\gamma(A)}{\gamma(\mathcal{Y})}.$$

From the above relationship it is common to find a Dirichlet process parameterized in terms of its total probability mass, $\alpha = \gamma(\mathcal{Y})$, and a (non-atomic) probability measure G_0 on $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$, with $G_0(A) = \gamma(A)/\alpha$. In such case, the probability law of the Dirichlet process is denoted by $\mathrm{DP}(\alpha, G_0)$, with $\alpha > 0$. The probability distribution G_0 is commonly known as the baseline probability measure, as $\mathbb{E}_{\mathrm{DP}}[F(A)] = G_0(A)$ for any measurable set A. Whereas the total mass parameter α is usually referred as the precision parameter, as it regulates the concentration of F around its mean/baseline probability measure, since

$$\mathrm{var}_{\mathrm{DP}}[F(A)] = \frac{G_0(A)[1 - G_0(A)]}{\alpha + 1}.$$

So, the larger the α the closer the random probability measure F will be to the mean G_0.

3.3.2 Conjugacy and prediction

Notice that under the assumption of exchangeability (3.1), the joint probability measure of n random variables $(Y_1, \ldots, Y_n)$ defined on the common measurable space $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$ having a common random distribution function F, and the probability measure for F, can be written as

$$\mathbb{P}(\mathrm{d}y_1, \ldots, \mathrm{d}y_n, \mathrm{d}F) = \left\{ \prod_{i=1}^{n} F(\mathrm{d}y_i) \right\} \cdot \mathrm{DP}(\mathrm{d}F; \alpha, G_0). \tag{3.6}$$

Perhaps one of the most important properties of the Dirichlet process, which makes it entirely compatible with the reasoning of Bayesian thinking, is that F, conditional on $(Y_1 = y_1, \ldots, Y_n = y_n)$, is also a Dirichlet process, $\mathrm{DP}(\alpha_n, G_n)$, with updated parameters

(see Ferguson, 1973)

$$\begin{aligned} \alpha_n &= \alpha + n, \\ G_n &= \frac{\alpha}{\alpha+n} G_0 + \frac{n}{\alpha+n} \cdot \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}, \end{aligned} \tag{3.7}$$

where δ_y is the probability mass function at y.

It also follows that the conditional mean of F given $Y_1 = y_1, \ldots, Y_n = y_n$, known as the predictive distribution of Y_{n+1} given $(y_1, \ldots, y_n)$, is given by

$$\mathbb{E}_{\mathrm{DP}}[F(A)|Y_1 = y_1, \ldots, Y_n = y_n] = \frac{\alpha}{\alpha+n} G_0(A) + \frac{n}{\alpha+n} \cdot \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}(A), \tag{3.8}$$

for any measurable set A. So, we can see that the conditional mean of F given a set of observations is a weighted mixture of the baseline distribution and the empirical distribution defined at the observed points. It is evident that the above mixture is tuned by the precision parameter α. So, the larger the α the closer F will be to its baseline measure after conditioning.

From equation (3.8) one can naturally notice the discreteness feature involved with the Dirichlet process. Next section will review some issues related to such a discreteness.

3.3.3 Discreteness

One of the most relevant properties that the Dirichlet process has is related to the discreteness of its sample paths, i.e. the Dirichlet process places probability 1 on the space of discrete distribution functions with support in the measurable space $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$. This result was exhibited by Ferguson (1973) and Blackwell (1973). There are several ways of interpreting this property. One of them comes by looking at the series representation of the Dirichlet process (see, Sethuraman, 1994). A more intuitive derivation of that result comes with the Pólya-urn representation of the sequence of the predictive distributions associated with the Dirichlet process (see, Blackwell and MacQueen, 1973), in which it is explicitly exhibited the discreteness of the sampling scheme selection from random distributions derived from such a process. The later is briefly described below.

Looking closely at (3.8) it is easy to see that when conditioning on $(Y_1, \ldots, Y_n)$ it is possible to conceive the possibility of observing repeated values among them. So, the set of distinct values (ties) among $(Y_1, \ldots, Y_n)$ is denoted by $(\tilde{Y}_1, \ldots, \tilde{Y}_{K_n})$, where K_n denotes the number of distinct observations among $(Y_1, \ldots, Y_n)$. Considering the predictive distribution (3.8) one can see that each tie will have associated a frequency $(n_1, \ldots, n_{K_n})$. And that frequency is such that $n = \sum_{j=1}^{K_n} n_j$. Therefore, the predictive distribution for a future random variable, say Y_{n+1}, conditionally on $(Y_1, \ldots, Y_n)$, can be rewritten as

$$Y_{n+1}|Y_1, \ldots, Y_n \sim \begin{cases} \delta_{\tilde{Y}_j} & \text{, with probability } \frac{n_j}{\alpha+n}, \text{ for } j = 1, \ldots, K_n, \\ G_0 & \text{, with probability } \frac{\alpha}{\alpha+n}. \end{cases} \tag{3.9}$$

An important remark on the above result is that the probability measure for a countable sequence of exchangeable variables with a marginal Dirichlet process distribution can be completely characterized by the collection of one-step conditional distributions (3.9). Moreover, due to exchangeability, such a representation is not restricted to the order chosen to characterize these one-step conditional distributions. The sequential formulation of this procedure gives rise to what is known as the generalized Pólya-urn scheme. See Blackwell and MacQueen (1973) for further details.

Let us anticipate that the general form of the predictive rule (3.9) is useful when considering mixtures of Dirichlet process, as we shall discuss in Section 3.7.

Before that, let us mention that the Dirichlet process has a number of alternative representations, which result in a variety of extended random probability measures. In the upcoming section we shall review some of those extension. Specifically, we revise species sampling models (Pitman, 1996), which are derived from the Pólya-urn representation of the Dirichlet process. We will also review the stick-breaking process, which is derived from a constructive representation of the Dirichlet process (Sethuraman, 1994). And we shall continue revising notions of neutral-to-the-right processes (Doksum, 1974; Ferguson, 1974), which are defined as a suitable transformation of increasing independent

increment processes. Other representation has been left aside, but will be mentioned in the discussion of this chapter.

3.4 Species sampling models

Let us notice that there exists an extended version to the prediction rule (3.9) due to Pitman (1996), which gives rise to what is known as *species sampling models*. As with the Dirichlet process, species sampling models characterize the distribution of a given infinite sequence of exchangeable random variables through the system of prediction rules given by

$$Y_{n+1}|Y_1,\ldots,Y_n \sim \begin{cases} \delta_{\tilde{Y}_j} & \text{, with probability } p_{j,n}, \text{ for } j=1,\ldots,K_n, \\ G_0 & \text{, with probability } q_n, \end{cases} \tag{3.10}$$

for some non-negative measurable functions of $(Y_1,\ldots,Y_n)$. See Hansen and Pitman (2000). Clearly, the Dirichlet process is a particular case of species sampling models with $p_{j,n}=\frac{n_j}{\alpha+n}$ and $q_n=\frac{\alpha}{\alpha+n}$. The required assumption regarding G_0 is to be a diffuse probability measure on the measurable space $(\mathcal{Y},\mathscr{B}_{\mathcal{Y}})$ (i.e. a non-atomic probability measure, which means that $G_0(\{y\})=0$ for all $y\in\mathcal{Y}$).

3.5 Stick-breaking processes

Sethuraman (1994) described a constructive representation of the Dirichlet process, which exhibits the discreteness of their sample paths with the form (3.4). The key component in such a representation is the specification of random weights (W_j) satisfying the summability condition (3.5). Such a constructive representation has been extended by Ishwaran and James (2001) to define a general class of random probability measures termed stick-breaking process. The name makes reference to the way the random weights (W_j) are specified.

As before, let us assume that $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$ is an arbitrary measurable (Polish) space and that F is a discrete random probability measure on $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$ of the form (3.4). So,

$$F(\cdot) = \sum_{k=1}^{\infty} W_k \delta_{Z_k}(\cdot),$$

where $\{W_k\}_{k=1}^{\infty}$ and $\{Z_k\}_{k=1}^{\infty}$ are two random sequences mutually independent. It is worth to mention here that the stick breaking representation accommodates finite versions of the above random probability measure; such versions are useful when dealing with arbitrary or stochastic truncations of discrete random probability measures for computational purposes (see, Ishwaran and James, 2001). For the moment, let us focus on the description of stick-breaking processes involving infinite series expansions.

Let us recall that the specification of the random weights (W_k) should satisfy the summability condition (3.5). For that, Sethuraman (1994) developed a specification of (W_k) through additional independent latent variables (V_k). Such a specification gives rise to the general stick-breaking specification of (W_k) as follows:

$$W_1 = V_1 \quad \text{and} \quad W_k = V_k \prod_{l=1}^{k-1} (1 - V_l), \tag{3.11}$$

for $k = 2, \ldots,$ with the (V_j) being stochastically independent and $V_k \sim \text{Be}(a_k, b_k)$ for some $a_k > 0$ and $b_k > 0$. Ishwaran and James (2001) showed that the (W_k) defined as above satisfy the summability condition iff

$$\sum_{k=1}^{\infty} \mathbb{E}\left(\log(1 - V_k)\right) = -\infty, \tag{3.12}$$

or equivalently if

$$\sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty. \tag{3.13}$$

The specification of stick breaking processes is completed by assuming that the random location (Z_k) are such that

$$Z_k \overset{\text{i.i.d.}}{\sim} G_0,$$

with G_0 being a diffuse probability measure on $(\mathcal{Y}, \mathscr{B}_{\mathcal{Y}})$. Also, the sequences $\{Z_k\}$ and $\{W_k\}$ taking to be mutually independent (so do $\{V_k\}$ and $\{Z_k\}$).

Under the above representation of F, we obtain that the probability measure on $(\mathcal{F}, \mathscr{B}_{\mathcal{F}})$, reduces to the joint probability law of the sequence $\{(W_k, Z_k)\}$ (or equivalently of $\{(V_k, Z_k)\}$). So, when a random discrete probability measure F has a stick-breaking representation, it is denoted by

$$F \sim \text{SB}(\boldsymbol{a}, \boldsymbol{b}, G_0) \tag{3.14}$$

where $\boldsymbol{a} = \{a_k\}_{k=1}^{\infty}$, $\boldsymbol{b} = \{b_k\}_{k=1}^{\infty}$ and G_0 is a diffuse probability distribution. Recall that G_0 plays the role of the baseline probability measure, i.e. $\mathbb{E}_{\text{SB}}[F] = G_0$.

Two particular examples of stick-breaking processes are the Dirichlet process (Sethuraman, 1994) and the two-parameter Poisson-Dirichlet processes (Pitman and Yor, 1997). Those processes only differ in the way the random weights (W_j) are specified.

3.5.1 Dirichlet process

Let F be a Dirichlet process with parameters $\alpha > 0$ and G_0, i.e. $F \sim \text{DP}(\alpha, G_0)$, with G_0 being the baseline probability measure. The stick breaking representation of F due to Sethuraman (1994) consists on taking the (V_k) in (3.11) as

$$V_k \overset{\text{i.i.d.}}{\sim} \text{Be}(1, \alpha), \tag{3.15}$$

for $k = 1, \ldots$ and $Z_j \overset{\text{i.i.d.}}{\sim} G_0$.

3.5.2 Two-parameter Poisson-Dirichlet process

The two-parameter Poisson-Dirichlet process (also known as the Pitman-Yor process) was introduced by Pitman and Yor (1997) and constitutes a flexible extension to Dirichlet process. The construction of such a process was originally conceived in a different way, as a normalized version of random measures. Its alternative representation as a stick-breaking process is due to Pitman (1996).

So, it is said that a discrete random probability measure F is a two-parameter

Poisson-Dirichlet process if

$$V_k \overset{\text{i.i.d.}}{\sim} \text{Be}(1-\alpha, \beta + k\alpha), \tag{3.16}$$

for $k = 1, \ldots$ and $Z_k \overset{\text{i.i.d.}}{\sim} G_0$, where $0 \leq \alpha < 1$ and $\beta > -\alpha$. And it is denoted by

$$F \sim \text{PDP}(\alpha, \beta, G_0). \tag{3.17}$$

It is worth noticing that two-parameter Poisson-Dirichlet process encompasses the Dirichlet process with parameter $\alpha' > 0$ as a particular case, by simply taking $\alpha = 0$ and $\beta = \alpha'$ in the former case.

3.6 Neutral-to-the-right processes

Neutral-to-the-right processes characterise random distributions functions on the positive real line as suitable transformation of processes with nondecreasing independent increments (see Subsection 2.4.2.1). These processes were originally introduced to Bayesian nonparametrics by Doksum (1974) for modeling cumulative hazard functions in survival analysis. For these processes, the probability law governing the evolution of the underlying process with nondecreasing independent increments serves as a prior measure on the space of cumulative distributions (or equivalently, on the space of cumulative hazards) on the positive real line. As the probability law of the underlying process is characterised by a unique Lévy measure, every computation in a Bayesian setting involving these processes takes place at the level of characteristic functions (via the Lévy-Kitchine representation). So, posterior distributions will typically be characterised by updated versions of the Lévy measure and some additional jump components.

It is worth noticing that it is possible to extend the notion of neutral-to-the-right process to the whole real line. This can be carried out by means of introducing an additional Lévy measure for the negative real line, the counterpart to the one governing the evolution of the process on the positive real line. All the remaining implementations/-calculations can be conducted by symmetry arguments on both Lévy measures. For the

moment, for the ease of exposition, let us focus on neutral-to-the-right process on the positive real line.

Definition 3.6.1. A random distribution function $F(t)$ defined on the measure space $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$ is said to be neutral-to-the-right (hereafter abbreviated NTR) if for every n and $0 < t_1 < \ldots < t_n$ there exist independent random variables $V_1, \ldots, V_n$ such that

$$(1 - F(y_1), \ldots, 1 - F(y_n)) \stackrel{d}{=} \left(V_1, V_1 V_2, \ldots, \prod_{j=1}^{n} V_j \right), \tag{3.18}$$

with respect to some probability law $\mathbb{P}$.

So, it is equivalent to say that a random distribution function is NTR if the normalised increments for a given partition of the positive real line are mutually independent, i.e. if for any n and $0 < t_1 < \ldots < t_n$ it follows that the random variables

$$F(t_1), \frac{F(t_2) - F(t_1)}{1 - F(t_1)}, \ldots, \frac{F(t_n) - F(t_{n-1})}{1 - F(t_{n-1})} \tag{3.19}$$

are mutually independent. Notice the above definition of NTR process it is possible to observe their resemblance to stochastic process with nondecreasing increments processes. In fact, Doksum (1974) exhibited such a correspondence by specifically taking $\{Z(t) : t \geq 0\}$ to be a process with nondecreasing independent increments (see Subsection 2.4.2.1), such that:

- $Z(t)$ is non-decreasing almost surely a. s.,
- $Z(t)$ is right-continuous a. s.,
- $Z(0) = 0$ a.s.,
- $\lim_{t \to \infty} Z(t) = +\infty$ a. s..

Then, if F is a separable random distribution function it is said that F is NTR iff

$$F(t) \stackrel{d}{=} 1 - \exp\{-Z(t)\}, \tag{3.20}$$

for a given process $\{Z(t) : t \geq 0\}$ with the above properties. See Theorem 3.1 in Doksum (1974).

For the moment, let us assume that the process $\{Z(t)\}$ does not have fixed points of discontinuities. Therefore, by means of the Lévy-Kintchine representation of $Z(t)$ one obtains

$$\begin{aligned} \mathbb{E}_{\text{NTR}}[F(t)] &= 1 - \mathbb{E}\left[e^{-Z(t)}\right] \\ &= 1 - \exp\left\{-\int_0^t \int_0^\infty (1 - e^{-s})\nu(\mathrm{d}u, \mathrm{d}s)\right\}, \end{aligned} \tag{3.21}$$

for any $t > 0$ and a given Lévy measure ν such that :

i) $\int_0^t \int_B \nu(\mathrm{d}u, \mathrm{d}s)$ is non-decreasing and continuous for any B in $\mathscr{B}_{\mathbb{R}_+}$,

ii) $\nu_t(\cdot) = \int_0^t \nu(\mathrm{d}u, \cdot)$ is a measure on $\mathbb{R}_+$ which must satisfy,

$$\int_0^\infty \frac{s}{1+s}\nu_t(\mathrm{d}s) < \infty, \tag{3.22}$$

for any $t > 0$.

Let us notice that in a more general setting the process $\{Z(t) :\geq 0\}$ can be defined as the sum of a (stochastically) 'continuous' component, Z_c, and a 'deterministic' jump component, Z_d, with at most countably many fixed points of discontinuity at fixed points $\{t_k\}_{k=1}^\infty$ with associated random jumps $\{S_k\}_{k=1}^\infty$. In addition, it is assumed that Z_c and Z_d are mutually independent. Therefore, by means of the Lévy-Kintchine representation for Z_c, it follows that

$$\begin{aligned} \mathbb{E}[F(t)] &= 1 - \mathbb{E}\left[e^{-Z(t)}\right] \\ &= 1 - \mathbb{E}\left[e^{-Z_d(t)}\right] \cdot \mathbb{E}\left[e^{-Z_c(t)}\right] \\ &= 1 - \left\{\prod_{t_k \leq t} \mathbb{E}\left[e^{-S_k}\right]\right\} \cdot \exp\left\{-\int_0^t \int_0^\infty (1 - e^{-s})\nu(\mathrm{d}u, \mathrm{d}s)\right\}, \end{aligned} \tag{3.23}$$

for any $t > 0$. This result will be used in the next subsection when analysing the conjugacy of NTR processes.

Let us now present a particular example of NTR process: the Dirichlet process. A more general case, called the beta-Stacy process, is being treated later in more detail, as it comprises the mentioned example as a particular case.

Example 3.1. Dirichlet process. The Dirichlet process described in Section 3.3 with support on the positive real line can be seeing as a particular case of NTR processes, with Lévy measure:

$$\nu(\mathrm{d}u, \mathrm{d}s) = \frac{e^{-s\alpha[u,\infty)}}{1-e^{-s}}\mathrm{d}s\ \alpha(\mathrm{d}u), \tag{3.24}$$

where α is a measure parameter on $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$. Notice that the measure parameter α corresponds to the baseline measure of the Dirichlet process. See Ferguson (1974) for further details.

∘

3.6.1 Conjugacy

Here, let us consider a broad general setting by considering a collection of exchangeable positive real-valued random variables $\{Y_i\}_{i=1}^{\infty}$ such that for any fixed $n \geq 1$ there is F a NTR process such that $Y_1, \ldots, Y_n | F \overset{\text{ind}}{\sim} F$ and $F \sim$ NTR. A distinctive feature of NTR processes is their conjugacy under updating (conditioning) with exact and right-censored observations, in the sense that the F given $(Y_1, \ldots, Y_n)$ is also a NTR process, with updated stochastic components. This was shown by Ferguson (1974) when considering exact observation; whilst the conjugacy of NTR under right-censored observations was developed by (see, Ferguson and Phadia, 1979). Here we shall focus only in the conjugacy of NTR under updating with exact observations. Accordingly, the probability law for $(Y_1, \ldots, Y_n, F)$ is given by

$$\mathbb{P}(\mathrm{d}y_1, \ldots, \mathrm{d}y_n, \mathrm{d}F) = \prod_{i=1}^{n} F(\mathrm{d}y_i)\, \mathrm{NTR}(\mathrm{d}F). \tag{3.25}$$

So, marginally F is distributed NTR. It is assumed that marginally F does not have fixed points of discontinuity. Therefore, it is marginally specified by a given Lévy measure ν.

Computing the conditional measure for F given $(y_1, \ldots, y_n)$ can be carried out by repeated applications of the following theorem due to Ferguson (1974).

Theorem 3.6.1. Let F be a NTR process with no fixed points of discontinuity, characterised by a given Lévy measure ν. Let Y_1 be a positive real-valued random variable such that $Y_1|F \sim F$. Then, $F|Y_1 = t_1$ is also NTR characterised by an updated process $Z^*(t)$ with a fixed-jump component and a continuous component, where:

- Z_d^* is the fixed-jump component with a single point of discontinuity at t_1 and a random jump-size $S_1|t_1 \sim F_{t_1}$ with $F_{t_1}(\mathrm{d}s) \propto (1 - e^{-s})\nu(\mathrm{d}s, t_1)$, and
- Z_c^* is the continuous component, which is characterised by the updated Lévy measure $\nu^*(\mathrm{d}u, \mathrm{d}s) = \exp\{-s\,\mathbb{1}(t_1 \geq u)\}\nu(\mathrm{d}u, \mathrm{d}s)$.

So, the conditional measure for F given $Y = t_1$ is characterized by F_{t_1} and ν^*. That is to say that $F|Y_1 = t_1 \sim \mathrm{NTR}(F_{t_1}, \nu^*)$.

Let us now consider an additional random variable, Y_2, such that $Y_2|F \sim F$. So, the distribution of Y_2 given $Y_1 = t_1$ can be computed by means of the Lévy-Kintchine representation of Z^* which is given by

$$\begin{aligned} G_1(t_2|t_1) &= \mathbb{E}_{\mathrm{NTR}}[F(t_2)|t_1] \\ &= 1 - \mathbb{E}\left[e^{-Z^*(t_2)}\right] \\ &= 1 - \prod_{t_1 \leq t_2} \mathbb{E}\left[e^{-S_{t_2}}\right] \cdot \exp\left\{-\int_0^{t_2}\int_0^{\infty} \left(1 - e^{-s}\right) \nu^*(\mathrm{d}u, \mathrm{d}s)\right\}, \end{aligned} \tag{3.26}$$

with S_{t_1} and ν^* given as in Theorem 3.6.1.

3.6.2 Beta-Stacy process

The beta-Stacy process (hereafter abbreviated BS) was introduced by Walker and Muliere (1997) and is a particular parameterization of the NTR process whose continuous component is driven by a log-beta process with nonhomogeneous Lévy measure of the form

$$\nu(\mathrm{d}u, \mathrm{d}s) = \frac{1}{1 - e^{-s}} e^{-\,s\beta(u)} \mathrm{d}s\ \alpha(\mathrm{d}u), \tag{3.27}$$

for $u > 0$ and $s \geq 0$, with α being a continuous finite measure on $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$ and β being an nonnegative function defined on the positive real line, such that

$$\lim_{t \to +\infty} \int_0^t \frac{\alpha(\mathrm{d}u)}{\beta(u)} = +\infty. \tag{3.28}$$

So, if F is beta-Stacy with parameters α and β, we shall write $F \sim \mathrm{BS}(\alpha, \beta)$. In particular, if α is continuous (meaning that marginally F does not have fixed point of discontinuity), one can get that the marginal distribution for $Y_1|F \sim F$ becomes

$$G_0(t_1) = \mathbb{E}_{\mathrm{BS}}[F(t_1)] = 1 - \exp\left\{ - \int_0^{t_1} \frac{\alpha(\mathrm{d}u)}{\beta(u)} \right\}, \tag{3.29}$$

for any $t_1 \geq 0$.

Remark 3.1. Walker and Muliere (1997) exposed the connection of beta-Stacy process with the extended gamma process of Dykstra and Laud (1981) and the beta process of Hjort (1990). Such a connection takes place in their corresponding Lévy measures, due to suitable transformation highlighted by Walker and Muliere. However, the later process are not cases of NTR process due to the transformation used to induce their corresponding induced probability measure; the extended gamma process is intended to model a hazard function as a kernel mixture with respect to a Lévy process, whereas the beta process aims at modelling directly the cumulative hazard function.

3.6.2.1 Conditioning and prediction

Let us explore now the conditioning structure of beta-Stacy process under exact observations. So, if Y_1 is such that $Y_1|F \sim F$ and $F \sim \mathrm{BS}(\alpha, \beta)$, with α being a continuous measure on the positive real line. Then, $F|Y_1 = t_1$ is also beta-Stacy with (see, Walker and Muliere, 1997, Corollary 2):

- a jump component at t_1 with random jump-size S_1 such that

$$1 - \exp\{-S_{t_1}\}|t_1 \sim \mathrm{Be}(1, \beta(t_1)) \tag{3.30}$$

- a continuous component with updated Lévy measure ν^* given by

$$\nu^*(\mathrm{d}u, \mathrm{d}s) = \frac{1}{1-e^{-s}} e^{-s[\beta(u)+\mathbb{1}(t_1 \geq u)]} \mathrm{d}s\alpha(\mathrm{d}u). \tag{3.31}$$

Accordingly, if Y_2 is another positive real-valued random variable such that it is conditionally independent of Y_1 and $Y_2|F \sim F$, then the conditional distribution for Y_2 given $Y_1 = t_1$ is given by

$$\begin{aligned} G_1(t_2|t_1) &= \mathbb{E}_{\mathrm{BS}}[F(t_2)|t_2] \\ &= 1 - \mathbb{E}[e^{-Z^*(t_2)}] \\ &= 1 - \frac{\beta(t_1)}{\beta(t_1)+\mathbb{1}(t_1 \leq t_2)} \cdot \exp\left\{ -\int_0^{t_2} \frac{\alpha(\mathrm{d}u)}{\beta(u)+\mathbb{1}(t_1 \geq u)} \right\}. \end{aligned} \tag{3.32}$$

This is of great importance to our future work on time series modelling.

As mentioned before, conjugacy of NTR process under right-censored observations is also a very important feature, which makes this type of processes suitable to analyse survival data. However, such a property is not relevant for our purpose in this thesis, hence details have been omitted. But details concerning this property can be found in, e.g., Ferguson and Phadia (1979) and Doksum (1974).

3.7 Nonparametric mixture models

All the stochastic processes described above share the key peculiarity of sampling discrete paths with probability one. As a consequence, their probability law may have a limited usage as a prior distribution when the purpose of the analysis is of density estimation of continuous data. However, the notion of random probability measures can be easily extended to notions of absolutely continuous random distributions in a variety of ways [see Walker, Damien, Laud, and Smith (1999) and Walker (2005), and the references therein, for a review on this matter]. The approach we shall consider in this thesis consists in randomizing a density function through nonparametric mixture models; as it was first explored by Antoniak (1974) and Lo (1984).

Lo's model considers nonparametric mixture densities, with a given parametric kernel mixed with respect to a random probability measure. The way those mixtures are defined are fundamentally different. Let us clarify, Antoniak (1974) considers densities generated as mixtures of nonparametric process (i.e. random distribution functions) mixed with respect to some underlying characteristics of its corresponding baseline measure. On the other hand; Lo (1984) represents a random density as a kernel model mixed with respect to a given random probability distribution. It is worth noticing that Lo's model has been the most popular approach, due that it is easier to work with from a computationally and theoretical point of view. In short, the random density functions we shall consider here take the form

$$p_F(y) = p(y|F) = \int K(y|\lambda)F(\mathrm{d}\lambda), \tag{3.33}$$

where λ is an auxiliary mixing random variable taking values in a measurable (Polish) space $(\mathcal{L}, \mathscr{B}_{\mathcal{L}})$, $K(\cdot|\cdot)$ is a given probability kernel on $\mathscr{B}_{\mathcal{Y}} \times \mathcal{L}$ (see, Definition 2.3.1), and F is a random probability measure on the measurable space $(\mathcal{L}, \mathscr{B}_{\mathcal{L}})$. In the sequel we shall be assuming that the kernel, $K(\cdot|\cdot)$, and the way the mixing variable, λ, enter into the model are fixed.

So, the space of random distribution (or density) functions we shall be working here becomes

$$\mathcal{P} = \{P_F : P_F = P(\cdot|F) \text{ is a distribution function on } (\mathcal{Y}, \mathscr{B}_{\mathcal{Y}}), \text{ with density } p_F\}, \tag{3.34}$$

and, as usual, we shall assume that $\mathcal{P}$ is endowed with its corresponding Borel σ-field $\mathscr{B}_{\mathcal{P}}$. We shall also assume that F is defined in the measurable (Polish) space $(\mathcal{F}, \mathscr{B}_{\mathcal{F}})$.

Notice that for a given kernel the distribution function P_F and its density p_F are uniquely specified by the random measure F. Hence, a probability measure $\widetilde{\Pi}$ on the space $(\mathcal{P}, \mathscr{B}_{\mathcal{P}})$ it is equivalent to the probability measure Π on the space $(\mathcal{F}, \mathscr{B}_{\mathcal{F}})$. Consequently, for a given fixed kernel K, it follows that $\widetilde{\Pi}(p_F \in A) = \Pi(F \in B)$, for

certain compatible sets $A \in \mathscr{B}_{\mathcal{P}}$ and $B \in \mathscr{B}_{\mathcal{F}}$, and

$$\mathbb{E}_{\widetilde{\Pi}}[p_F(y)] = \int K(y|\lambda)\, \mathbb{E}_{\Pi}[F(\mathrm{d}\lambda)],$$

due to Fubini's theorem, for any $y \in \mathcal{Y}$.

Accordingly, any random distribution function, F, can be used to specify a different nonparametric version of density functions like p_F, and their law would serve as the prior distribution on the space $\mathcal{P}$, indexed by F. In such a representation, things become relatively simple when F is a discrete random measure, as we will discuss below. But it is worth to say that more general representations, considering more general random measures, are current under study.

For the moment, let us say that as with the previous random probability measures we have described above, the pair of density, p_F, and probability law, Π, define a joint probability measure on the space $(\mathcal{Y} \times \mathcal{F}, \mathscr{B}_{\mathcal{Y}} \otimes \mathscr{B}_{\mathcal{F}})$, with

$$\mathbb{P}(\mathrm{d}y, \mathrm{d}F) \quad = \quad p(y)\mathrm{d}y\Pi(\mathrm{d}F) \tag{3.35}$$

An exchangeable sequence of random variables can be also generated by (3.35) after assuming that any finite sequence (Y_j) are i. i. d. F, conditionally on F. As with the other models, we would be interested in calculating conditional probabilities for F given past observations of (Y_j), and its corresponding predictive versions. Such a conditional and predictive versions of (3.35) will serve two purposes: as a stochastic component for the model construction, and as a procedure for statistical inference.

Let us anticipate that Π loses the conjugacy in other random probability measures. That is basically due to the data augmentation scheme required to compute such a conditional distribution, which replaces the exchangeable dependence among (Y_j) with a notion of partial exchangeable dependence. So, due to the mixture component intrinsically involved in p_F it is no longer possible to find closed analytical forms for the conditional distribution of F given past observations of (Y_j). In Subsection 3.7.1 we shall describe a general data augmentation scheme used to compute such a conditional

distribution, with the aim of recalling some basic properties for model construction. In Subsection 3.7.2 we shall also address some sampling schemes to approximate such a conditional for general discrete nonparametric mixture models, via Markov chain Monte Carlo (MCMC) methods [see, Robert and Casella (1999), for an annotated review of this matter]. Some issues regarding more general nonparametric mixture models are discussed in Subsection 3.7.3.

3.7.1 Conditioning and prediction

Computing the conditional distribution for F, or equivalently for p_F, given a collection of random variables $(Y_1 = y_1, \ldots, Y_n = y_n)$ generated by (3.41), involves augmenting the model with missing (latent) variables, by means of associating with each Y_j a latent variable Λ_j. By doing so, it is possible to rewrite the model in the following extended hierarchical form,

$$\begin{aligned} Y_i|\lambda_i &\overset{\text{ind}}{\sim} K(y_i|\lambda_i) \\ \lambda_i|F &\overset{\text{i.i.d.}}{\sim} F(\lambda_i) \\ F &\sim \Pi, \end{aligned} \tag{3.36}$$

for $i = 1, \ldots, n$ and any $n \geq 1$. Notice that by extending the model with the above data augmentation scheme breaks down the exchangeable assumption among (Y_i) given (λ_i) and F. However, the (λ_i) remain exchangeable with respect to F and (Y_i) share a partial exchangeable dependence.

By making use of the extended hierarchical structure of the model, and considering the conditional independence structure involved, the probability for F conditionally on $(y_1, \ldots, y_n)$ can be written in terms of the conditional distribution of F given the latent variables $(\lambda_1, \ldots, \lambda_n)$ (i.e. the posterior distribution of F in the Bayesian framework), which is computed under the assumption of exchangeability among (λ_i); and the marginal conditional distribution for $(\lambda_1, \ldots, \lambda_n)$ given $(y_1, \ldots, y_n)$, computed after integrating

out F from the hierarchical model (see, Lo, 1984; Ferguson, 1974, for further details). That is,

$$\begin{aligned}\widetilde{\Pi}(\mathrm{d}p_F|y_1,\ldots,y_n) &= \Pi(\mathrm{d}F|y_1,\ldots,y_n) \\ &= \int\cdots\int \Pi(\mathrm{d}F|\lambda_1,\ldots,\lambda_n)P(\mathrm{d}\lambda_1,\ldots,\mathrm{d}\lambda_n|y_1,\ldots,y_n), \end{aligned} \tag{3.37}$$

where $\Pi(\mathrm{d}F|\lambda_1,\ldots,\lambda_n)$ denotes the probability measure for F conditionally on the latent variables $(\lambda_1,\ldots,\lambda_n)$, and $P(\mathrm{d}\lambda_1,\ldots,\mathrm{d}\lambda_n|y_1,\ldots,y_n)$ represents the conditional distribution of $(\lambda_1,\ldots,\lambda_n)$ given $(y_1,\ldots,y_n)$. Notice that $\Pi(\mathrm{d}F|\lambda_1,\ldots,\lambda_n)$ is relatively easy to derive, due that the (λ_i) are i. i. d. conditionally on F. Such a derivation obeys the exchangeability assumption among (λ_j), or equivalently the i. i. d. assumption for (λ_j) conditionally on F. However, $P(\mathrm{d}\lambda_1,\ldots,\mathrm{d}\lambda_n|y_1,\ldots,y_n)$ deserves special attention, as it is computed after marginalizing F in the model. Assuming that F and Π are such that $\mathbb{E}_\Pi[F(\lambda)] = G_0(\lambda)$, for any $\lambda \in \mathcal{L}$ and a given diffuse probability measure G_0 on $(\mathcal{L}, \mathscr{B}_\mathcal{L})$, with density g_0, it follows that probability $P(\mathrm{d}\lambda_1,\ldots,\mathrm{d}\lambda_n|y_1,\ldots,y_n)$ is determined by

$$p(\lambda_1,\ldots,\lambda_n|y_1,\ldots,y_n) \propto K(y_1|\lambda_1)p(\lambda_1)\cdot\prod_{i=2}^{n} K(y_i|\lambda_i)p(\lambda_i|\lambda_{i-1},\ldots,\lambda_1), \tag{3.38}$$

where $p(\lambda_1) = g_0(\lambda_1)$ is the density of the baseline distribution. From the above expression, it is clear that the main issue when calculating (3.37) is the computation of the sequence of predictive densities $p(\lambda_i|\lambda_{i-1},\ldots,\lambda_1)$, for $i = 2,\ldots,n$.

When conditioning on a one single observation, the probability measure of F given $Y_1 = y_1$ becomes,

$$\Pi(\mathrm{d}F|y_1) = \int \Pi(\mathrm{d}F|\lambda_1)P(\mathrm{d}\lambda_1|y_1), \tag{3.39}$$

with $\Pi(\mathrm{d}F|\lambda_1)$ being the conditional probability measure for $F(\lambda_1)$ under Π, and $p(\lambda_1|y_1) \propto K(y_1|\lambda_1)g_0(\lambda_1)$. And, the predictive density for Y_2 given $Y_1 = y_1$ turns out to take the

form

$$\begin{aligned}
\mathbb{E}_{\widetilde{\Pi}}[p_F(y_2)|y_1] &= \int p_F(y_2)\widetilde{\Pi}(\mathrm{d}p_F|y_1) \\
&= \int \left\{ \int K(y_2|\lambda_2)F(\mathrm{d}\lambda_2) \right\} \Pi(\mathrm{d}F|y_1) \\
&= \int \left\{ \int K(y_2|\lambda_2)F(\mathrm{d}\lambda_2) \right\} \int \Pi(\mathrm{d}F|\lambda_1)P(\mathrm{d}\lambda_1|y_1) \\
&= \int K(y_2|\lambda_2) \int \left\{ \int F(\mathrm{d}\lambda_2)\Pi(\mathrm{d}F|\lambda_1) \right\} P(\mathrm{d}\lambda_1|y_1) \\
&= \int \int K(y_2|\lambda_2)\, \mathbb{E}_{\Pi}[F(\lambda_2)|\lambda_1]P(\mathrm{d}\lambda_1|y_1), \qquad (3.40)
\end{aligned}$$

where $P(\lambda_1|y_1) \propto K(y_1|\lambda_1)P(\lambda_1)$, with $P = \mathbb{E}_{\Pi}(F)$.

Let us point out that the distribution (3.40) is generally intractable. Therefore, there is no analytic expression for it available, and its computation heavily relies on computational sampling algorithms; i.e., Markov chain Monte Carlo methods (see, Robert and Casella, 1999, among other references).

3.7.2 Discrete nonparametric mixtures models

When F is a discrete random probability measure of the form (3.4) the density function p_F can be written as an infinite component mixture model,

$$p_F(y) = \sum_{j=1}^{\infty} W_j K(y|\lambda_j), \qquad (3.41)$$

where, in general, the sequences $\{W_j\}_{j=1}^{\infty}$ and $\{\lambda_j\}_{j=1}^{\infty}$ are assumed to be random and defined as in Section 3.2.1. Notice that the above representation accommodates a vast variety of specifications for (3.41) by means of defining the random weights, (W_j), and the random locations, (λ_j), via the Dirichlet process, species sampling models or stick-breaking processes; with the Dirichlet process being the most used variant of the above mixture representation.

Letting aside the stochastic components ruling (W_j), in all of the cases mentioned above it is possible to satisfy $\mathbb{E}[F] = G_0$, with G_0 being a diffuse probability measure on

$(\mathcal{L}, \mathscr{B}_{\mathcal{L}})$. So, the follow relation follows naturally,

$$\mathbb{E}_{\widetilde{\Pi}}[p_F(y)] = \int K(y|\lambda)\, \mathbb{E}_{\Pi}(\mathrm{d}\lambda) = \int K(y|\lambda) G_0(\mathrm{d}\lambda), \tag{3.42}$$

with Π being the probability law governing (W_j) and (λ_j).

3.7.2.1 Computational considerations

A fundamental issue regarding nonparametric mixture models is the computation of the probability measure for p_F, conditionally on $(Y_1 = y_1, \ldots, Y_n = y_n)$, for a given kernel K. Equivalently, it is required to compute the conditional distribution for F, the mixing random distribution, given $(y_1, \ldots, y_n)$. As mentioned before, the computation of the conditional distribution for p_F or F, rely on computational methods, which can be catalogued in two types: i) Marginal sampling algorithm, or ii) Explicit sampling algorithms. The marginal sampling algorithms were designed not to compute explicitly the conditional distribution of F, but to focus on the observable and latent variables. These methods marginalize F in the sampling algorithm. See Escobar (1994) and Escobar and West (1998) for a description of this computational method.

The explicit sampling algorithm takes the random probability measure F explicitly into account. For this method it is fundamental to write F as a discrete random measure, with random weights (W_j) formed as a stick-breaking sequence of random variables. Notice that the conditional distribution of F will involve to take into account countable sums. So, the key idea here is to truncate the range of the sum (3.41) to a finite number of components. Such a truncation can be arbitrary (see, Ishwaran and James, 2001), or stochastic (see, Walker, 2007; Papaspiliopoulos and Roberts, 2008). Among the stochastic procedures to sample this mixture model we shall focus on Walker (2007), which truncate the range of (3.41) using an auxiliary slicing variable.

The idea developed by Walker (2007) consists in extending (3.41) by including an

additional latent variable u, such that the joint distribution for y and u becomes

$$\begin{aligned} p(y,u) &= \sum_{j=1}^{\infty} \mathbb{1}(u < W_j) K(y|\lambda_j) \\ &= \sum_{j=1}^{\infty} W_j \, \mathrm{U}(u|0, W_j) K(y|\lambda_j). \end{aligned} \tag{3.43}$$

The above extended expression induces a truncation of (3.41) to a finite number of components, as the number of (W_j) being greater than any given u is finite. That is due to the Poisson process structure the probability weights (W_j) have associated. So, as a consequence, the number of jumps of the Poisson process outside any neighbour of 0 will be finite.

It is direct to see that the original model (3.41) can be reproduced after integrating out u in (3.43). The truncation that u induces takes place on the probability weights (W_j), but it has a direct effect on the index set of $\{(W_j, \lambda_j)\}$, giving rise to a discrete finite sum over the sub-index set

$$A(u) = \{j : u < W_j\}, \tag{3.44}$$

i.e., the re-normalized density after conditioning on u becomes,

$$p_F(y|u) = \frac{1}{\#A(u)} \sum_{j \in A(u)} K(y|\lambda_j). \tag{3.45}$$

Then, drawing samples of the posterior distribution of (3.41) can be done via the Gibbs sampler on $\{(W_j, \lambda_j), u, k\}$, where k is the index in (3.45) from which the observation y comes, conditionally on u. Computing the full conditional distributions involving (W_j, λ_j) can be obtained using ideas developed by Ishwaran and James (2001). Updating procedures for the pair (u, k) were developed by Walker (2007).

Besides the simplification of the slice sampler to truncate (3.41) into a finite sum, the set $A(u)$ will typically be formed by a non-sequential set of indexes j. That is because, although the random weights (W_j) are decreasing in mean –as a feature of their stick-breaking construction–, there will be realizations of them with relative high values

in higher indexes. In Chapter 6 we shall use a similar truncation idea for which the truncation set of indexes is formed sequentially up to a given truncation variable.

3.7.3 Neutral-to-the-right mixture models

Neutral-to-the-right mixture models are those whose density functions p_F have the form

$$p_F(y) = \int K(y|\lambda)F(\mathrm{d}\lambda), \tag{3.46}$$

with F being neutral-to-the-right (NTR). Although NTR random probability measures are discrete almost surely, in general, they can not be written as a discrete random measure (see Subsection 3.2.1), as their paths are typically piecewise continuous (a. s.).

Finding analytic expressions for the conditional distributions of F given one single observation is doable, as we shall explore further in the upcoming chapters. However, handling the case of conditioning on $(y_1, \ldots, y_n)$ can present analytic difficulties, as the computational procedures developed for discrete mixture models are not applicable in this context. However, due to James (2006), it is possible to enrich or redefine the mixture model (3.46) by means of incorporating the mixing variable into the Lévy measure for a particular marked point process, which is defined on the extended sample space considering $\mathcal{L}$, where λ takes values. Such an extension gives rise to what is known as spatial neutral-to-the-right process. See, James (2006) and Doksum and James (2004) for an abbreviated account.

3.8 Discussion

In this chapter we have reviewed some basic notions regarding Bayesian nonparametric procedures, specifically the notion of random probability distribution and some probability measures with support on the space of cumulative distribution functions. Notice that most of the probability measures defined on the functional measurable space $(\mathcal{F}, \mathscr{B}_{\mathcal{F}})$ are connected and characterized in terms of Lévy processes, in some way or another. The

differences rely on the type of transformation used to define random probability measures from Lévy processes. Actually, such connections can be elegantly stated in terms of the notion of what is called "completely random measures" (see, Kingman, 1967, 1975). Lijoi and Prünster (2010) developed such connections between contemporary random probability measures existing in the literature. There is can also found a detailed review of other random probability measures and Bayesian nonparametric mixture models.

Let us notice that the essence of Bayesian nonparametric procedures, as described above, regards on the characterization of a joint probability law of a countable sequence of random variables, $\{Y_j\}_{j=1}^{\infty}$, and a functional random probability measure, F. Moreover, under the assumption of exchangeability and by means of de Finetti's representation theorem, the probability law of the sequence of exchangeable variables can be characterised by an unknown random component, F, taking values in the measurable space $(\mathcal{F}, \mathscr{B}_{\mathcal{F}})$. So, F in conjunction with its own probability measure, Π, define a joint probability measure on the product space $(\mathcal{Y} \times F, \mathscr{B}_{\mathcal{Y}} \otimes \mathscr{B}_{\mathcal{F}})$, which can be decomposed as

$$\mathbb{P}(\mathrm{d}y, \mathrm{d}F) = \mathbb{P}(\mathrm{d}y|F) \cdot \mathbb{P}(\mathrm{d}F) = F(\mathrm{d}y) \cdot \Pi(\mathrm{d}F). \tag{3.47}$$

Notice that one can actually manipulate the above joint product measure using standard tools developed in probability theory.

In the next chapter, we shall treat the product measure (3.47) as an instrument to construct models with a desired dependence structure, more than as a devise for statistical inference, as it is originally conceived in the Bayesian nonparametric framework.

So, let us mention here that in the sequel of this thesis we shall be exploiting the notions reviewed in this chapter with a double purpose: i) in order to construct flexible stationary time series models, in the spirit of Mena and Walker's approach; and ii) to make inference on dynamic versions of nonparametric models. When possible, and in order to avoid confusions to the reader, we shall point out such a difference in perspective at different parts during the course of this work. Additionally, let us point out that an alternative Bayesian nonparametric approach to inference on random probability

measures will, which was not covered in this chapter, will be exploited in Chapter 5.