Collegio Carlo Alberto

Models beyond the Dirichlet process

Antonio Lijoi

Igor Prünster

No. 129

December 2009

Carlo Alberto Notebooks

www.carloalberto.org/working_papers

Models beyond the Dirichlet process

Antonio Lijoi[1] and Igor Prünster[2]

[1] Dipartimento Economia Politica e Metodi Quantitatavi, Università degli Studi di Pavia, via San Felice 5, 27100 Pavia, Italy and CNR–IMATI, via Bassini 15, 20133 Milano.
E-mail: lijoi@unipv.it

[2] Dipartimento di Statistica e Matematica Applicata, Collegio Carlo Alberto and ICER, Università degli Studi di Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy.
E-mail: igor@econ.unito.it

September 2009

Abstract. Bayesian nonparametric inference is a relatively young area of research and it has recently undergone a strong development. Most of its success can be explained by the considerable degree of flexibility it ensures in statistical modelling, if compared to parametric alternatives, and by the emergence of new and efficient simulation techniques that make nonparametric models amenable to concrete use in a number of applied statistical problems. Since its introduction in 1973 by T.S. Ferguson, the Dirichlet process has emerged as a cornerstone in Bayesian nonparametrics. Nonetheless, in some cases of interest for statistical applications the Dirichlet process is not an adequate prior choice and alternative nonparametric models need to be devised. In this paper we provide a review of Bayesian nonparametric models that go beyond the Dirichlet process.

1 Introduction

Bayesian nonparametric inference is a relatively young area of research and it has recently undergone a strong development. Most of its success can be explained by the considerable degree of flexibility it ensures in statistical modelling, if compared to parametric alternatives, and by the emergence of new and efficient simulation techniques that make nonparametric models amenable to concrete use in a number of applied statistical problems. This fast growth is witnessed by some review articles and monographs providing interesting and accurate accounts on the state of the art in Bayesian nonparametrics. Among them we mention the discussion paper by Walker, Damien, Laud and Smith (1999), the book by Ghosh and Ramamoorthi (2003), the lecture notes by Regazzini (2001) and the review articles by Hjort (2003) and Müller and Quintana (2004). Here we wish to provide an update to all these excellent works. In particular, we focus on classes of nonparametric priors that go beyond the Dirichlet process.

The Dirichlet process has been a cornerstone in Bayesian nonparametrics since the seminal paper by T.S. Ferguson has appeared on the Annals of Statistics in 1973. Its success can be partly explained by its mathematical tractability and it has tremendously grown with the development of Markov chain Monte Carlo (MCMC) techniques whose implementation allows a full Bayesian analysis of complex statistical models based on the Dirichlet process prior. To date the most effective applications of the

Dirichlet process concern its use as a nonparametric distribution for latent variables within hierarchical mixture models employed for density estimation and for making inference on the clustering structure of the observations.

Nonetheless, in some cases of interest for statistical applications the Dirichlet process is not an adequate prior choice and alternative nonparametric models need to be devised. An example is represented by survival analysis: if a Dirichlet prior is used for the survival time distribution, then the posterior, conditional on a sample containing censored observations, is not Dirichlet. It is, then, of interest to find an appropriate class of random distributions which contain, as a special case, the posterior distribution of the Dirichlet process given censored observations. Moreover, in survival problems one might be interested in modelling hazard rate functions or cumulative hazards and the Dirichlet process cannot be used in these situations. On the other hand, in problems of clustering or species sampling, the predictive structure induced by the Dirichlet process is sometimes not flexible enough to capture important aspects featured by the data. Finally, in regression analysis one would like to elicit a prior which depends on a set of covariates, or on time, and the Dirichlet process is not able to accommodate for this modelling issue. Anyhow, besides these applied motivations, it is useful to view the Dirichlet process as a special case of a larger class of prior processes: this allows to gain a deeper understanding of the behaviour of the Dirichlet process itself.

Most of the priors we are going to present are based on suitable transformations of completely random measures: these have been introduced and studied by J.F.C. Kingman and are random measures giving rise to mutually independent random variables when evaluated on pairwise disjoint measurable sets. The Dirichlet process itself can be seen as the normalization of a so–called gamma completely random measure. Here it is important to emphasize that this approach sets up a unifying framework that we think is useful both for the understanding of the behaviour of commonly exploited priors and for the development of new models. Indeed, even if completely random measures are quite sophisticated probabilistic tools, their use in Bayesian nonparametric inference leads to intuitive *a posteriori* structures. We shall note this when dealing with: neutral to the right priors, priors for cumulative hazards, priors for hazard rate functions, normalized random measures with independent increments, hierarchical mixture models with discrete random mixing distribution. Recent advances in this area have strongly benefited from the contributions of J. Pitman who has developed some probabilistic concepts and models which fit very well within the Bayesian nonparametric framework.

The final part of this section is devoted to a concise summary of some basic notions that will be used throughout this paper.

1.1. Exchangeability assumption. Let us start by considering an (ideally) infinite sequence of observations $X^{(\infty)} = (X_n)_{n\geq 1}$, defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with each X_i taking values in a complete and separable metric space $\mathbb{X}$ endowed with the Borel σ–algebra $\mathscr{X}$. Throughout the present paper, as well as in the most commonly employed Bayesian models, $X^{(\infty)}$ is assumed to be *exchangeable*. In other terms, for any $n \geq 1$ and any permutation π of the indices $1, \ldots, n$, the probability distribution (p.d.) of the random vector $(X_1, \ldots, X_n)$ coincides with the p.d. of $(X_{\pi(1)}, \ldots, X_{\pi(n)})$. A celebrated result of de Finetti, known as de Finetti's representation theorem, states that the sequence $X^{(\infty)}$ is exchangeable if and only if it is a mixture of sequences of independent and identically distributed (i.i.d.) random variables.

Theorem 1. (de Finetti, 1937). *The sequence $X^{(\infty)}$ is exchangeable if and only if there exists a probability measure Q on the space $\mathcal{P}_{\mathbb{X}}$ of all probability measures on $\mathbb{X}$ such that, for any $n \geq 1$ and $A = A_1 \times \cdots \times A_n \times \mathbb{X}^{\infty}$, one has*

$$\mathbb{P}\left[X^{(\infty)} \in A\right] = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} p(A_i)\, Q(\mathrm{d}p)$$

where $A_i \in \mathscr{X}$ for any $i = 1, \ldots, n$ and $\mathbb{X}^{\infty} = \mathbb{X} \times \mathbb{X} \times \cdots$.

In the statement of the theorem, the space $\mathcal{P}_{\mathbb{X}}$ is equipped with the topology of weak convergence which makes it a complete and separable metric space. The probability Q is also termed the *de Finetti measure* of the sequence $X^{(\infty)}$. We will not linger on technical details on exchangeability and its connections with other dependence properties for sequences of observations. The interested reader can refer to the exhaustive and stimulating treatments of Aldous (1985) and Kallenberg (2005).

The exchangeability assumption is usually formulated in terms of conditional independence and identity in distribution, *i.e.*

$$\begin{aligned} X_i \,|\, \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \qquad\qquad i \geq 1 \\ \tilde{p} &\sim Q \end{aligned} \tag{1}$$

Hence, $\tilde{p}^n = \prod_{i=1}^{n} \tilde{p}$ represents the conditional p.d. of $(X_1, \ldots, X_n)$, given $\tilde{p}$. Here $\tilde{p}$ is some random probability measure defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $\mathcal{P}_{\mathbb{X}}$: its distribution Q takes on the interpretation of prior distribution for Bayesian inference. Whenever Q degenerates on a finite dimensional subspace of $\mathcal{P}_{\mathbb{X}}$, the inferential problem is usually called *parametric*. On the other hand, when the support of Q is infinite–dimensional then one typically speaks of a *nonparametric* inferential problem.

In the following sections we focus our attention on various families of priors Q: some of them are well–known and occur in many applications of Bayesian nonparametric statistics whereas some others have recently appeared in the literature and witness the great vitality of this area of research. We will describe specific classes of priors which are tailored for different applied statistical problems: each of them generalizes the Dirichlet process in a different direction, thus obtaining more modelling flexibility with respect to some specific feature of the prior process. This last point can be appreciated when considering the predictive structure implied by the Dirichlet process, which actually overlooks some important features of the data. Indeed, it is well–known that, in a model of the type (1), the family of predictive distributions induced by a Dirichlet process, with baseline measure α, are

$$\mathbb{P}\left[X_{n+1} \in A \,|\, X_1, \ldots, X_n\right] = \frac{\alpha(\mathbb{X})}{\alpha(\mathbb{X}) + n} P_0(A) + \frac{n}{\alpha(\mathbb{X}) + n} \frac{1}{n} \sum_{j=1}^{k} n_j \delta_{X_j^*}(A) \qquad \forall A \in \mathscr{X}$$

where δ_x denotes a point mass at $x \in \mathbb{X}$, $P_0 = \alpha/\alpha(\mathbb{X})$ and the X_j^*'s with frequency n_j denote the $k \leq n$ distinct observations within the sample. The previous expression implies that X_{n+1} will be a new observation X_{k+1}^* with probability $\alpha(\mathbb{X})/[\alpha(\mathbb{X}) + n]$, whereas it will coincide with any of the previous observations with probability $n/[\alpha(\mathbb{X}) + n]$. Since these probability masses depend neither on the number of clusters into which the data are grouped nor on their frequencies, an important piece of information for prediction is neglected. It is quite complicated to obtain a tractable generalization of the Dirichlet process incorporating dependence on both the number of clusters and the frequencies:

however, dependence on the number of clusters is achievable and the two parameter Poisson–Dirichlet process, with $\sigma \in (0,1)$ and $\theta > -\sigma$, represents a remarkable example. Details will be provided later, but here we anticipate that the predictive distribution implies that X_{n+1} will be a new value X^*_{k+1} with probability $[\theta + \sigma k]/[\theta + n]$, whereas X_{n+1} will coincide with a previously recorded observation with probability $[n - \sigma k]/[\theta + n]$. Hence, the probability of obtaining new values is monotonically increasing in k and the value of σ can be used to tune the strength of the dependence on k.

The analysis of general classes of priors implies that, in most of the cases and in contrast to what happens for the Dirichlet process, one has to work with non–conjugate models. This should not be a big concern, since conjugacy corresponds to mere mathematical convenience: from a conceptual point of view, there is no justification for requiring conjugacy. On the contrary, one may argue that conjugacy constrains the posterior to having the same structure as the prior which, in a nonparametric setup, may represent a limitation to the desired flexibility. So it is definitely worth exploring the potential of random probability measures which do not have this feature and it will be seen that, even if conjugacy fails, one can find many alternatives to the Dirichlet process which preserve mathematical tractability. Since most of these general classes of priors are obtained as suitable transformations of completely random measures, in the next subsection we provide a concise digression on this topic.

1.2. A concise account on completely random measures. We start with the definition of completely random measure, a concept introduced in Kingman (1967). Denote, first, by $\mathcal{M}_{\mathbb{X}}$ the space of boundedly finite measures on $(\mathbb{X}, \mathscr{X})$, this meaning that for any μ in $\mathcal{M}_{\mathbb{X}}$ and any bounded set A in $\mathscr{X}$ one has $\mu(A) < \infty$. Moreover, we let $\mathscr{M}_{\mathbb{X}}$ stand for the corresponding Borel σ–algebra on $\mathcal{M}_{\mathbb{X}}$. For technical details on $\mathcal{M}_{\mathbb{X}}$ and the construction of $\mathscr{M}_{\mathbb{X}}$, one can refer to Daley and Vere–Jones (1988).

Definition 1. Let $\tilde{\mu}$ be a measurable mapping from $(\Omega, \mathscr{F}, \mathbb{P})$ into $(\mathcal{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$ and such that for any $A_1, \ldots, A_n$ in $\mathscr{X}$, with $A_i \cap A_j = \varnothing$ for any $i \neq j$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent. Then $\tilde{\mu}$ is termed *completely random measure* (CRM).

An important property of CRMs is their almost sure discreteness (Kingman, 1993), which means that their realizations are discrete measures with probability 1. This fact essentially entails discreteness of random probability measures obtained as transformations of CRMs such as those presented in Sections 2 and 3. See, *e.g.*, James (2003). Discreteness of the Dirichlet process was first shown in Blackwell (1973).

A CRM on $\mathbb{X}$ can always be represented as the sum of two components: a completely random measure $\tilde{\mu}_c = \sum_{i=1}^{\infty} J_i \delta_{X_i}$, where both the positive jumps J_i's and the $\mathbb{X}$–valued locations X_i's are random, and a measure with random masses at fixed locations. Accordingly

$$\tilde{\mu} = \tilde{\mu}_c + \sum_{i=1}^{M} V_i \, \delta_{x_i} \tag{2}$$

where the fixed jump points $x_1, \ldots, x_M$, with $M \in \{1, 2, \ldots\} \cup \{\infty\}$, are in $\mathbb{X}$, the (non–negative) random jumps $V_1, \ldots, V_M$ are mutually independent and they are independent from $\tilde{\mu}_c$. Finally, $\tilde{\mu}_c$ is characterized by the *Lévy–Khintchine* representation which states that

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} f(x)\, \tilde{\mu}_c(\mathrm{d}x)}\right] = \exp\left\{-\int_{\mathbb{R}^+ \times \mathbb{X}} \left[1 - \mathrm{e}^{-sf(x)}\right] \, \nu(\mathrm{d}s, \mathrm{d}x)\right\} \tag{3}$$

where $f : \mathbb{X} \to \mathbb{R}$ is a measurable function such that $\int |f| \, \mathrm{d}\tilde{\mu}_c < \infty$ (almost surely) and ν is a measure on $\mathbb{R}^+ \times \mathbb{X}$ such that

$$\int_B \int_{\mathbb{R}^+} \min\{s, 1\} \, \nu(\mathrm{d}s, \mathrm{d}x) < \infty$$

for any B in $\mathscr{X}$. The measure ν characterizing $\tilde{\mu}_c$ is referred to as the *Lévy intensity* of $\tilde{\mu}_c$: it contains all the information about the distributions of the jumps and locations of $\tilde{\mu}_c$. Such a measure will play an important role throughout and many of the results to be presented are stated in terms of ν. For our purposes it will often be useful to separate the jump and location part of ν by writing it as

$$\nu(\mathrm{d}s, \mathrm{d}x) = \rho_x(\mathrm{d}s) \, \alpha(\mathrm{d}x) \tag{4}$$

where α is a measure on $(\mathbb{X}, \mathscr{X})$ and ρ a transition kernel on $\mathbb{X} \times \mathscr{B}(\mathbb{R}^+)$, *i.e.* $x \mapsto \rho_x(A)$ is $\mathscr{X}$–measurable for any A in $\mathscr{B}(\mathbb{R}^+)$ and ρ_x is a measure on $(\mathbb{R}^+, \mathscr{B}(\mathbb{R}^+))$ for any x in $\mathbb{X}$. If $\rho_x = \rho$ for any x, then the distribution of the jumps of $\tilde{\mu}_c$ is independent of their location and both ν and $\tilde{\mu}_c$ are termed *homogeneous*. Otherwise, ν and $\tilde{\mu}_c$ are termed *non–homogeneous*.

Example 1. (The gamma process). A homogeneous CRM $\tilde{\gamma}$ whose Lévy intensity is given by

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\mathrm{e}^{-s}}{s} \, \mathrm{d}s \, \alpha(\mathrm{d}x) \tag{5}$$

is a *gamma process* with parameter measure α on $\mathbb{X}$. It is characterized by its Laplace functional which is given by $\mathbb{E}\left[\mathrm{e}^{-\int f \, \mathrm{d}\tilde{\gamma}}\right] = \mathrm{e}^{-\int \log(1+f) \, \mathrm{d}\alpha}$ for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\int \log(1 + |f|) \, \mathrm{d}\alpha < \infty$. Now set $f = \lambda \mathbb{1}_B$ with $\lambda > 0$, $B \in \mathscr{X}$ such that $\alpha(B) < \infty$ and $\mathbb{1}_B$ denoting the indicator function of set B. In this case one obtains

$$\mathbb{E}\left[\mathrm{e}^{-\lambda \, \tilde{\gamma}(B)}\right] = [1 + \lambda]^{-\alpha(B)},$$

from which it is apparent that $\tilde{\gamma}(B)$ has a gamma distribution with scale and shape parameter equal to 1 and $\alpha(B)$, respectively. □

Example 2. (The σ–stable process). Let $\sigma \in (0, 1)$ and consider a CRM $\tilde{\mu}_\sigma$ with Lévy intensity defined by

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\sigma}{\Gamma(1 - \sigma) \, s^{1+\sigma}} \, \mathrm{d}s \, \alpha(\mathrm{d}x) \tag{6}$$

Then $\tilde{\mu}_\sigma$ is a *σ–stable process* with parameter measure α on $\mathbb{X}$. Moreover, for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\int |f|^\sigma \, \mathrm{d}\alpha < \infty$, the Laplace functional is of the form $\mathbb{E}\left[\mathrm{e}^{-\int f \, \mathrm{d}\tilde{\mu}_\sigma}\right] = \mathrm{e}^{-\int f^\sigma \, \mathrm{d}\alpha}$. Hence, the Laplace transform of $\tilde{\mu}_\sigma(B)$ is that of a positive stable random variable, namely $\mathbb{E}\left[\mathrm{e}^{-\lambda \tilde{\mu}_\sigma(B)}\right] = \mathrm{e}^{-\lambda^\sigma \, \alpha(B)}$. □

As one may note from (3), CRMs are also closely connected to Poisson processes. Indeed, $\tilde{\mu}_c$ can be represented as a linear functional of a Poisson process $\tilde{\Pi}$ on $\mathbb{R}^+ \times \mathbb{X}$ with mean measure ν. To state this precisely, $\tilde{\Pi}$ is a random subset of $\mathbb{R}^+ \times \mathbb{X}$ and if $N(A) = \mathrm{card}(\tilde{\Pi} \cap A)$ for any $A \subset \mathscr{B}(\mathbb{R}^+) \otimes \mathscr{X}$ such that $\nu(A) < \infty$, then

$$\mathbb{P}[N(A) = k] = \frac{(\nu(A))^k \, \mathrm{e}^{-\nu(A)}}{k!} \qquad k = 0, 1, 2, \ldots.$$

It can then be shown that

$$\tilde{\mu}_c(A) = \int_A \int_{\mathbb{R}^+} s \, N(\mathrm{d}s, \mathrm{d}x) \qquad \forall A \in \mathscr{X}. \tag{7}$$

A detailed treatment of this subject can be found in the superb book by Kingman (1993).

If $\tilde{\mu}$ is defined on $\mathbb{X} = \mathbb{R}$, one can also consider the càdlàg random distribution function induced by $\tilde{\mu}$, namely $\{\tilde{\mu}((-\infty, x]) : x \in \mathbb{R}\}$. Such a random function defines an *increasing additive process*, that is a process whose increments are non–negative, independent and possibly not stationary. See Sato (1999) for an exhaustive account. To indicate such processes we will also use the term *independent increments processes*, whereas in the Bayesian literature they are more frequently referred to as Lévy processes: this terminology is not completely appropriate since, in probability theory, the notion of Lévy process is associated to processes with independent and stationary increments. We rely on CRMs in most of our exposition since they represent an elegant, yet simple, tool for defining nonparametric priors. Moreover, one can easily realize that posterior inferences are achieved by virtue of the simple structure featured by CRMs conditional on the data. Indeed, in most of the examples we will illustrate, *a posteriori* a CRM turns out to be the sum of two independent components: (i) a CRM with no fixed points of discontinuity and whose Lévy intensity is obtained by applying an updating rule to the prior Lévy intensity; (ii) a sum of random jumps. These jumps occur at: a) the *a priori* fixed points of discontinuities with updated jump distribution; b) the new fixed points of discontinuity corresponding to the observations with jump distribution determined by the Lévy intensity of the CRM. Given this common structure, the specific updating of the involved quantities depends on the specific transformation of the CRM that has been adopted for defining the prior.

Finally, note that, without loss of generality, one can *a priori* consider CRMs with no fixed points of discontinuity which implies $\tilde{\mu} = \tilde{\mu}_c$. In the sequel we adopt this assumption when specifying some of the nonparametric priors we deal with and it will be pointed out how fixed points of discontinuity arise when evaluating the posterior distribution, given a sample $X_1, \ldots, X_n$.

2 Models for survival analysis

Survival analysis has been the focus of many contributions to Bayesian nonparametric theory and practice. Indeed, many statistical problems arising in the framework of survival analysis require function estimation and, hence, they are ideally suited for a nonparametric treatment. Moreover, this represents an area where the interest in generalizations of the Dirichlet process has emerged with particular emphasis. The main reason for this is due to the particular nature of survival data which are governed by some censoring mechanism. The breakthrough in the treatment of these issues in a Bayesian nonparametric setup can be traced back to Doksum (1974) where the notion of neutral to the right (NTR) random probability is introduced. The law of a NTR process can be used as a prior for the distribution function of survival times and the main advantage of Doksum's definition is that NTR priors are conjugate (in a sense to be made precise later), even when right–censored observations are present. While this enables one to model a random distribution function for the survival times, a different approach yields priors for cumulative hazards and hazard rates. This has been pursued in a number of papers such as Dykstra and Laud (1981), Lo and Weng (1989), Hjort (1990), Kim (1999), Nieto–Barajas and Walker (2004) and James (2005). All the proposals we are going to examine arise as suitable transformations of CRMs.

2.1. Neutral to the right priors. A simple and useful approach for defining a prior on the space of distribution functions on $\mathbb{R}^+$ has been devised by Doksum (1974) who introduces the notion

of neutral to the right prior.

Definition 2. A random distribution function $\tilde{F}$ on $\mathbb{R}^+$ is *neutral to the right* (NTR) if, for any $0 \leq t_1 < t_2 < \cdots < t_k < \infty$ and any $k \geq 1$, the random variables

$$\tilde{F}(t_1), \quad \frac{\tilde{F}(t_2) - \tilde{F}(t_1)}{1 - \tilde{F}(t_1)}, \quad \cdots \quad , \frac{\tilde{F}(t_k) - \tilde{F}(t_{k-1})}{1 - \tilde{F}(t_{k-1})}$$

are independent.

The concept of *neutrality* has been introduced in Connor and Mosimann (1969) and it designates a random vector $(\tilde{p}_1, \ldots, \tilde{p}_{k+1})$ of proportions with $\sum_{i=1}^{k+1} \tilde{p}_i = 1$ such that $\tilde{p}_1$ is independent from $\tilde{p}_2/(1-\tilde{p}_1)$, $(\tilde{p}_1, \tilde{p}_2)$ is independent from $\tilde{p}_3/(1-\tilde{p}_1-\tilde{p}_2)$ and so on. This can be seen as a method for randomly splitting the unit interval and, as will be shown in Section 3.4, it is also exploited in order to define the so–called stick–breaking priors. In the definition above, one has $\tilde{p}_i = \tilde{F}(t_i) - \tilde{F}(t_{i-1})$ for any $i = 1, \ldots, k$, where $\tilde{F}(t_0) = 0$.

We recall the connection between NTR priors and CRMs on $\mathbb{R}^+$ which has been pointed out by Doksum (1974).

Theorem 2. (Doksum, 1974). *A random distribution function $\tilde{F} = \{\tilde{F}(t) : t \geq 0\}$ is NTR if and only if it has the same p.d. of the process $\left\{1 - \mathrm{e}^{-\tilde{\mu}((0,t])} : t \geq 0\right\}$, for some CRM $\tilde{\mu}$ on $\mathbb{X} = \mathbb{R}^+$ such that $\mathbb{P}[\lim_{t\to\infty} \tilde{\mu}((0,t]) = \infty] = 1$.*

By virtue of this result one can characterize both the prior and, as we shall see, the posterior distribution of $\tilde{F}$ in terms of the Lévy intensity ν associated to $\tilde{\mu}$. For instance, one can evaluate the prior guess at the shape of $\tilde{F}$ since

$$\mathbb{E}[\tilde{F}(t)] = 1 - \mathbb{E}\left[\mathrm{e}^{-\tilde{\mu}((0,t])}\right] = 1 - \mathrm{e}^{-\int_{(0,t]} \int_{\mathbb{R}^+} \left[1-\mathrm{e}^{-s}\right] \rho_x(\mathrm{d}s)\, \alpha(\mathrm{d}x)}.$$

Another feature which makes NTR priors attractive for applications is their conjugacy property.

Theorem 3. (Doksum, 1974). *If $\tilde{F}$ is NTR($\tilde{\mu}$), then the posterior distribution of $\tilde{F}$, given the data $X_1, \ldots, X_n$, is NTR($\tilde{\mu}^*$) where $\tilde{\mu}^*$ is a CRM with fixed points of discontinuity.*

In light of the previous result it is worth remarking that, in a Bayesian nonparametric setup, the term "conjugacy" is used with slightly different meanings. For this reason, we introduce here a distinction between *parametric conjugacy* and *structural conjugacy*. The former occurs when the p.d. of the posterior process is the same as the p.d. of the prior process with updated parameters: for instance, the posterior distribution of the Dirichlet process with parameter–measure α, given uncensored data, is still a Dirichlet process with updated parameter–measure $\alpha + \sum_{i=1}^{n} \delta_{X_i}$. The latter, namely structural conjugacy, identifies a model where the posterior process has the same structure of the prior process in the sense that they both belong to the same general class of random probability measures. Hence, Theorem 3 establishes that NTR priors are structurally conjugate: the posterior of a NTR($\tilde{\mu}$) process is still NTR. Note that structural conjugacy does not necessarily imply parametric conjugacy: the posterior CRM $\tilde{\mu}^*$ characterizing the NTR process is not necessarily of the same type as the prior. On the other hand, parametric conjugacy of a specific prior implies structural conjugacy.

An explicit description of the posterior CRM $\tilde{\mu}^*$ has been provided in Ferguson (1974). Denote by $\bar{\Lambda}(x) := \sum_{i=1}^{n} \delta_{X_i}([x, \infty))$ the number of individuals still alive right before x,, *i.e.* the so–called

at risk process. Moreover, $X_1^*, \ldots, X_k^*$ represent the k distinct observations among $X_1, \ldots, X_n$, with $1 \leq k \leq n$. As mentioned before, we suppose, for notational simplicity, that $\tilde{\mu}$ does not have a priori fixed points of discontinuity.

Theorem 4. (Ferguson, 1974) *If $\tilde{F}$ is NTR($\tilde{\mu}$) and $\tilde{\mu}$ has Lévy intensity* (4)*, then*

$$\tilde{\mu}^* \stackrel{d}{=} \tilde{\mu}_c^* + \sum_{i=1}^{k} J_i \, \delta_{X_i^*} \tag{8}$$

where $\tilde{\mu}_c^*$ is independent from $J_1, \ldots, J_k$ and the J_i's are mutually independent. Moreover, the Lévy intensity of the CRM $\tilde{\mu}_c^*$ is updated as

$$\nu^*(\mathrm{d}s, \, \mathrm{d}x) = \mathrm{e}^{-\bar{\Lambda}(x)\, s} \, \rho_x(\mathrm{d}s) \, \alpha(\mathrm{d}x)$$

One can also determine an expression for the p.d. of the jumps J_i at the distinct observations. To this end, consider the distinct observations in an increasing order $X_{(1)}^* < \cdots < X_{(k)}^*$. Moreover, let $n_i = \sum_{j=1}^{n} \delta_{X_j}(\{X_{(i)}^*\})$ be the frequency of the i–th ordered observation in the sample: in terms of the at risk process one has $n_i = \bar{\Lambda}(X_{(i)}^*) - \bar{\Lambda}(X_{(i+1)}^*)$ for any $i = 1, \ldots, k$ with the proviso that $X_{(k+1)}^* = \infty$. The p.d. of J_i is given by

$$G_i(\mathrm{d}s) = \frac{(1 - \mathrm{e}^{-s})^{n_i} \, \mathrm{e}^{-s\, \bar{n}_{i+1}} \, \rho_{X_{(i)}^*}(\mathrm{d}s)}{\int_{\mathbb{R}^+} (1 - \mathrm{e}^{-v})^{n_i} \, \mathrm{e}^{-v\, \bar{n}_{i+1}} \, \rho_{X_{(i)}^*}(\mathrm{d}v)}$$

where, for the sake of simplicity, we have set $\bar{n}_i := \bar{\Lambda}(X_{(i)}^*) = \sum_{j=i}^{k} n_j$. If ν is homogeneous, then $\rho_{X_{(i)}^*} = \rho$ and the distribution of J_i does not depend on the location where the jump occurs.

The above posterior characterization does not take into account the possibility that the data are subject to a censoring mechanism according to which not all observations are exact. In particular, in survival analysis, in reliability and in other models for the time elapsing up to a terminal event, a typical situation is represented by right–censoring. For example, when studying the survival of a patient subject to a treatment in a hospital, the observation is right–censored if her/his survival time cannot be recorded after she/he leaves the hospital. Formally, right–censoring can be described as follows. Suppose $c_1, \ldots, c_n$ are n censoring times which can be either random or nonrandom. For ease of exposition we assume that the c_i are deterministic. To each survival time X_i associate $\Delta_i = \mathbb{1}_{(0, c_i]}(X_i)$ and set $T_i = \min\{X_i, c_i\}$. Clearly $\Delta_i = 1$ if X_i is observed exactly, and $\Delta_i = 0$ if X_i is right–censored and the observed data are then given by $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$. Supposing there are $k \leq n$ distinct observations among $\{T_1, \ldots, T_n\}$, we record them in an increasing order as $T_{(1)}^* < \cdots < T_{(k)}^*$. Correspondingly, define

$$n_i^c := \sum_{\{j:\, \Delta_j = 0\}} \delta_{T_j}(\{T_{(i)}^*\}) \quad \text{and} \quad n_i := \sum_{\{j:\, \Delta_j = 1\}} \delta_{T_j}(\{T_{(i)}^*\}) \tag{9}$$

as the number of right–censored and exact observations, respectively, occurring at $T_{(i)}^*$ for any $i = 1, \ldots, k$. Finally, set $\tilde{n}_i^c = \sum_{j=i}^{k} n_j^c$ and $\bar{n}_i = \sum_{j=i}^{k} n_j$.

Theorem 5. (Ferguson and Phadia, 1979). *Suppose $\tilde{F}$ is NTR($\tilde{\mu}$) where $\tilde{\mu}$ has no fixed jump points. Then the posterior distribution of $\tilde{F}$, given $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$, is NTR($\tilde{\mu}^*$) with*

$$\tilde{\mu}^* \stackrel{d}{=} \tilde{\mu}_c^* + \sum_{\{i:\, n_i \geq 1\}} J_i \, \delta_{T_{(i)}^*} \tag{10}$$

Hence, the posterior distribution of $\tilde{F}$ preserves the same structure of the uncensored case and the jumps occur only at the exact observations, *i.e.* those distinct observations for which n_i is positive. In (10) $\tilde{\mu}_c^*$ is a CRM without fixed points of discontinuity and it is independent from the jumps J_i. Its Lévy measure coincides with

$$\nu^*(\mathrm{d}s, \mathrm{d}x) = \mathrm{e}^{-\bar{\Lambda}(x)s}\, \rho_x(\mathrm{d}s)\, \alpha(\mathrm{d}x)$$

where $\bar{\Lambda}(x) = \sum_{i=1}^n \delta_{T_i}([x, \infty))$ is the at risk process based on both exact and censored observations.

Moreover, the p.d. of the jump J_i occurring at each exact distinct observation, *i.e.* $T^*_{(i)}$ with $n_i \geq 1$, is given by

$$G_i(\mathrm{d}s) = \frac{(1-\mathrm{e}^{-s})^{n_i}\, \mathrm{e}^{-(\bar{n}_{i+1}+\tilde{n}_i^c)s}\, \rho_{T^*_{(i)}}(\mathrm{d}s)}{\int_{\mathbb{R}^+}(1-\mathrm{e}^{-v})^{n_i}\, \mathrm{e}^{-(\bar{n}_{i+1}+\tilde{n}_i^c)v}\, \rho_{T^*_{(i)}}(\mathrm{d}v)}.$$

Also in this case, if $\rho_{T^*_{(i)}} = \rho$ the distribution of J_i does not depend on the location at which the jump occurs. We close this subsection with a detailed description of two important examples of NTR priors.

Example 3. (The Dirichlet process). One might wonder whether the Dirichlet process defined by Ferguson (1973) is also a NTR prior. This amounts to asking oneself whether there exists a CRM $\tilde{\mu}$ such that the random distribution function $\tilde{F}$ defined by $\tilde{F}(t) \stackrel{d}{=} 1 - \mathrm{e}^{-\tilde{\mu}((0,t])}$ for any $t > 0$ is generated by a Dirichlet process prior with parameter measure α on $\mathbb{R}^+$. The answer to such a question is affirmative. Indeed, if $\tilde{\mu}$ is a CRM whose Lévy intensity is defined by

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\mathrm{e}^{-s\,\alpha((x,\infty))}}{1-\mathrm{e}^{-s}}\, \alpha(\mathrm{d}x)\, \mathrm{d}s$$

then $\tilde{F} \stackrel{d}{=} 1 - \mathrm{e}^{-\tilde{\mu}}$ is a Dirichlet process with parameter measure α. See Ferguson (1974). One can, then, apply results from Ferguson and Phadia (1979) in order to characterize the posterior distribution of a Dirichlet random distribution function given right–censored data. It is to be mentioned that such an analysis has been originally developed by Susarla and Van Ryzin (1976) without resorting to the notion of NTR prior. They show that the Dirichlet process features the property of parametric conjugacy if the observations are all exact, whereas it does not in the presence of right–censored data. Indeed, Blum and Susarla (1977) characterize the posterior distribution of a Dirichlet process given right–censored data as a mixture of Dirichlet processes in the sense of Antoniak (1974). In the present setting, a simple application of Theorem 5 allows to recover the results in Susarla and Van Ryzin (1976). Moreover, Theorem 5 implies that the Dirichlet process, in the presence of right–censored observations, is structurally conjugate when seen as a member of the class of NTR priors. The posterior distribution of the Dirichlet random distribution function $\tilde{F}$ is NTR($\tilde{\mu}^*$) with $\tilde{\mu}^*$ as in (10). The Lévy intensity of $\tilde{\mu}_c^*$ coincides with

$$\nu^*(\mathrm{d}s,\, \mathrm{d}x) = \frac{\mathrm{e}^{-\{\alpha((x,\infty))+\bar{\Lambda}(x)\}\, s}}{1-\mathrm{e}^{-s}}\, \alpha(\mathrm{d}x)\, \mathrm{d}s$$

and the distribution of the jump J_i at each exact distinct observation (*i.e.* $T^*_{(i)}$ with $n_i \geq 1$) coincides with the distribution of the random variable $-\log(B_i)$ where $B_i \sim \mathrm{Beta}(\alpha((T^*_{(i)}, \infty)) + \bar{n}_{i+1} + \tilde{n}_i^c;\ n_i)$. Note that if the observations are all exact,, then $\tilde{F}$ given the data is a Dirichlet process with parameter measure $\alpha + \sum_{i=1}^n \delta_{X_i}$ which coincides with the well–known result proved by Ferguson (1973). □

Example 4. (The beta–Stacy process). Having pointed out the lack of parametric conjugacy of the Dirichlet process in a typical inferential problem for survival analysis, one might wonder whether, conditionally on a sample featuring right–censored data, there exists a NTR process prior which shares both structural and parametric conjugacy. The problem has been successfully faced in Walker and Muliere (1997), where the authors define the *beta–Stacy* NTR prior. Its description can be provided in terms of the Lévy intensity of $\tilde{\mu}$ where, as usual, we are supposing that *a priori* $\tilde{\mu}$ does not have fixed jump points. To this end, suppose that α is some probability measure on $\mathbb{R}^+$ which is absolutely continuous with respect to the Lebesgue measure and $c : \mathbb{R}^+ \to \mathbb{R}^+$ some piecewise continuous function. Use the notation F_α to denote the distribution function corresponding to α, *i.e.* $F_\alpha(x) = \alpha((0, x])$ for any x. A beta–Stacy process $\tilde{F}$ with parameters α and c is NTR($\tilde{\mu}$) if $\tilde{\mu}$ is a CRM whose Lévy intensity is defined by

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\mathrm{e}^{-s\, c(x)\, \alpha((x,\infty))}}{1 - \mathrm{e}^{-s}}\, c(x)\, \mathrm{d}s\, \alpha(\mathrm{d}x). \tag{11}$$

Note that one obtains $\mathbb{E}[\tilde{F}] = F_\alpha$ and that the Dirichlet process arises when $c(x) \equiv c$. It is to be said, however, that the definition originally provided in Walker and Muliere (1997) is more general and it allows possible choices of parameter measures α having point masses. Here, for ease of exposition we confine ourselves to this simplified case.

Theorem 6. (Walker and Muliere, 1997). *Let $\tilde{F}$ be a beta–Stacy process with parameters α and c satisfying the conditions given above. Then $\tilde{F}$, given $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$, is still a beta–Stacy process with updated parameters*

$$\alpha^*((0, t]) = 1 - \prod_{x \in [0,t]} \left\{ 1 - \frac{c(x)\, \mathrm{d}F_\alpha(x) + \mathrm{d}\Phi(x)}{c(x)\alpha([x, \infty)) + \bar{\Lambda}(x)} \right\}$$

$$c^*(x) = \frac{c(x)\alpha([x, \infty)) + \bar{\Lambda}(x) - \sum_{i=1}^n \delta_{T_i}(\{x\})\delta_{\Delta_i}(\{1\})}{\alpha^*([x, \infty))}$$

where $\Phi(x) = \sum_{i=1}^n \delta_{T_i}((0, x])\, \delta_{\Delta_i}(\{1\})$ is the counting process for the uncensored observations.

In the previous statement $\prod_{x \in [0,t]}$ denotes the *product integral*, a quite standard operator in the survival analysis literature. If $l_m = \max_{i=1,\ldots,m} |x_i - x_{i-1}|$, the following definition holds true

$$\prod_{x \in [a,b]} \{1 + \mathrm{d}Y(x)\} := \lim_{l_m \to 0} \prod_j \{1 + Y(x_j) - Y(x_{j-1})\}$$

where the limit is taken over all partitions of $[a, b]$ into intervals determined by the points $a = x_0 < x_1 < \cdots < x_m = b$ and these partitions get finer and finer as $m \to \infty$. See Gill and Johanssen (1990) for a survey of applications of product integrals to survival analysis. Finally, the Bayes estimator of $\tilde{F}$, under squared loss, coincides with the distribution function F_{α^*} associated to α^*. Interestingly, if the function c goes to 0 (pointwise) then F_{α^*} converges to the Kaplan–Meier estimator. □

Remark 1. An appealing feature of NTR processes is that they allow for quite a rich prior specification in terms of the parameters of the Lévy intensity: in addition to the prior guess at the shape $\mathbb{E}[\tilde{F}]$, it is often also possible to assign a functional form to $\mathrm{Var}[\tilde{F}]$, whereas in the Dirichlet case, after selecting $\mathbb{E}[\tilde{F}]$, one is left with a single constant parameter to fix. A few details on this can be found in Walker and Muliere (1997) and Walker and Damien (1998). □

Remark 2. The posterior characterizations in Theorems 4 and 5 may not seem particularly appealing at first glance: however, they reveal explicitly the posterior structure and constitute the fundamental element for devising a sampling strategy for achieving posterior inferences. Indeed, relying on some algorithm for simulating the trajectories of independent increment processes $\{\tilde{\mu}((0,x]) : x \geq 0\}$, thanks to Theorems 4 and 5 a full Bayesian analysis can be carried out: this allows to derive Bayes estimates such as $\mathbb{E}\left[\tilde{F}(t) \,|\, \text{data}\right]$ or any other posterior quantity of statistical interest. See *e.g.* , Ferguson and Klass (1972), Damien, Laud and Smith (1995), Walker and Damien (1998, 2000) and Wolpert and Ickstadt (1998). □

2.2. Priors for cumulative hazards: the beta process. An alternative approach to inference for survival analysis, due to Hjort (1990), consists in assessing a prior for the cumulative hazard defined as

$$\tilde{H}_x = \tilde{H}_x(\tilde{F}) = \int_0^x \frac{\mathrm{d}\tilde{F}(v)}{1-\tilde{F}(v^-)} \tag{12}$$

where $F(v^-) = \lim_{z \downarrow 0} F(v-z)$ and the integrand is the *hazard rate*, *i.e.* the conditional probability of observing a death/failure/event at time v given that the individual is still alive (or the system is still functioning or the event has not yet occurred) at v. From (12) one has the following product integral representation of $\tilde{F}$ in terms of the cumulative hazard $\tilde{H}_x$

$$\tilde{F}(t) = 1 - \prod_{x \in [0,t]} \left\{1 - \mathrm{d}\tilde{H}_x\right\}. \tag{13}$$

Hence assessing a prior for the distribution function $\tilde{F}$ is the same as specifying a prior for $\tilde{H} = \{\tilde{H}_x : x \geq 0\}$ or for the hazard rate. The relation (13) between $\tilde{F}$ and $\tilde{H}$ suggests that the prior for $\tilde{H}$ should be such that

$$0 \leq \tilde{H}_x - \tilde{H}_{x-} \leq 1 \qquad \forall x \tag{14}$$

almost surely.
The main idea is, then, to model $\tilde{H}$ as a CRM $\tilde{\mu}$ by setting $x \mapsto \tilde{H}_x := \tilde{\mu}((0,x])$. However, due to (14), such a CRM must have all jumps of size less than 1. As shown in Hjort (1990), this happens if and only if the jump part of the Lévy intensity ν is concentrated on $[0,1]$, *i.e.*

$$\rho_x((1,\infty)) = 0 \qquad \forall x > 0. \tag{15}$$

Within this context, Hjort's beta process prior stands, in terms of relevance, as the analogue of the Dirichlet process for modelling probability distributions. Let, again, $c : \mathbb{R}^+ \to \mathbb{R}^+$ be a piecewise continuous function and H_0 be the baseline cumulative hazard which, for simplicity, we assume to be absolutely continuous. Consider now the beta CRM $\tilde{\mu}$ on $\mathbb{R}^+$ which is characterized by the Lévy intensity

$$\nu(\mathrm{d}s, \mathrm{d}x) = c(x)\, s^{-1}\, (1-s)^{c(x)-1}\, \mathrm{d}s\, \mathrm{d}H_{0,x}$$

for any $x \geq 0$ and $0 < s < 1$. Then, the beta process is defined as $\tilde{H} = \{\tilde{\mu}((0,x]) : x \geq 0\}$. In symbols, we write $\tilde{H} \sim \text{Beta}(c, H_0)$. Note that $\mathbb{E}[\tilde{H}_x] = H_{0,x}$. The relation between modelling the cumulative hazard with a CRM and specifying a NTR prior for the distribution function is clarified by the following

Theorem 7. (Hjort, 1990). *A random distribution function $\tilde{F}$ is NTR($\tilde{\mu}$) for some CRM $\tilde{\mu}$ if and only if the corresponding cumulative hazard $\tilde{H}(\tilde{F}) = \{\tilde{H}_x(\tilde{F}) : x \geq 0\}$ is an independent increments process with Lévy intensity satisfying condition* (15).

For an interesting illustration of further connections between priors for cumulative hazards and NTR processes see Dey, Erickson and Ramamoorthi (2003).
In analogy with NTR processes, a posterior characterization in terms of an updated CRM with fixed points of discontinuity corresponding to the exact observations can be given for general CRM cumulative hazards. See Hjort (1990). For brevity, here we focus on the beta process. Indeed, an important aspect of the beta process, which makes it appealing for applications to survival analysis, is its parametric conjugacy with respect to right–censoring. Recall that $\Phi(x) = \sum_{i=1}^{n} \delta_{T_i}((0,x])\delta_{\Delta_i}(\{1\})$ is the number of uncensored observations occurring up to time x and $\bar{\Lambda}(x) = \sum_{i=1}^{n} \delta_{T_i}([x,\infty))$ is the at risk process. One, then, has

Theorem 8. (Hjort, 1990). *Let $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$ be a sample of survival times. If $\tilde{H} \sim \text{Beta}(c, H_0)$ then*

$$\tilde{H} \,|\, \text{data} \;\sim\; \text{Beta}\left(c + \bar{\Lambda}, \int \frac{\mathrm{d}\Phi + c\,\mathrm{d}H_0}{c + \bar{\Lambda}}\right). \tag{16}$$

It follows immediately that the Bayes estimators of $\tilde{H}$ and $\tilde{F}$, with respect to a squared loss function, are

$$\hat{H}_x = \int_0^x \frac{c\,\mathrm{d}H_0 + \mathrm{d}\Phi}{c + \bar{\Lambda}}$$

and

$$\hat{F}(t) = 1 - \prod_{[0,t]} \left\{1 - \frac{c\,\mathrm{d}H_0 + \mathrm{d}\Phi}{c + \bar{\Lambda}}\right\}$$

respectively. Again, if we let the function c tend to zero, one obtains in the limit the Nelson–Aalen and the Kaplan–Meier estimators for $\tilde{H}$ and $\tilde{F}$, respectively.

In order to highlight the underlying posterior structure, Theorem 8 can be reformulated as follows. Suppose there are $k \leq n$ distinct values among $\{T_1, \ldots, T_n\}$ so that the data can be equivalently represented as $(T^*_{(1)}, n^c_1, n_1), \ldots, (T^*_{(k)}, n^c_k, n_k)$ with n^c_i and n_i defined as in (9). If $\tilde{H} \sim \text{Beta}(c, H_0)$, then one has

$$\tilde{H} \,|\, \text{data} \;\stackrel{d}{=}\; \tilde{H}^* + \sum_{\{i:\, n_i \geq 1\}} J_i\, \delta_{T^*_{(i)}}, \tag{17}$$

where $\tilde{H}^* \stackrel{d}{=} \{\tilde{\mu}^*((0,x]) : x \geq 0\}$ and $\tilde{\mu}^*$ is a beta CRM with updated Lévy intensity

$$\nu(\mathrm{d}s, \mathrm{d}x) = s^{-1}\,(1-s)^{c(x)+\bar{\Lambda}(x)-1} c(x)\,\mathrm{d}H_{0,x}. \tag{18}$$

The random jump at each distinct exact observation (*i.e.* $T^*_{(i)}$ with $n_i \geq 1$) has the following distribution

$$J_i \;\sim\; \text{Beta}\left([c(T^*_{(i)}) + \bar{\Lambda}(T^*_{(i)})]\mathrm{d}H^*_{0,T^*_{(i)}} \;;\; [c(T^*_{(i)}) + \bar{\Lambda}(T^*_{(i)})]\left\{1 - \mathrm{d}H^*_{0,T^*_{(i)}}\right\}\right).$$

where $\mathrm{d}H^*_{0,x} = [\mathrm{d}H_{0,x} + \mathrm{d}\Phi(x)]/[c(x) + \bar{\Lambda}(x)]$. These jumps can be merged with the updated beta CRM in (18) yielding the posterior representation in (16): note that the posterior baseline hazard in (16) is not continuous anymore. This sets up an analogy with what happens in the updating of the Dirichlet process, to be clarified in Section 3.1.

Remark 3. Recently, an interesting Bayesian nonparametric approach for dealing with factorial models with unbounded number of factors has been introduced in Griffiths and Ghahramani (2006). The marginal process, termed *Indian buffet process*, represents the analogue of the Blackwell–MacQueen, or Chinese restaurant, process for the Dirichlet model. As shown in Thibaux and Jordan (2007), the de Finetti measure of the Indian buffet process is a beta process defined on a bounded space $\mathbb{X}$. Specifically, the Indian Buffet process is an i.i.d. mixture of suitably defined Bernoulli processes with mixing measure the beta process. Such developments show how classes of random measures can become important also for completely different applications than the ones they were designed for. This witnesses the importance of studying general classes of random measures independently of possible immediate applications. □

Two interesting extensions of Hjort's pioneering work can be found in Kim (1999) and James (2006a). The model adopted in Kim (1999) allows for more general censoring schemes. Let $\mathcal{N}_i = \{\mathcal{N}_{i,x} : x \geq 0\}$, for $i = 1, \ldots, n$, be counting processes where $\mathcal{N}_{i,x}$ denotes the number of events (these being, for instance, deaths or failures) observed up to time x for the i–th counting process. Moreover, let the process $Y_i = \{Y_{i,x} : x \geq 0\}$ be the cumulative intensity associated to $\mathcal{N}_i$, thus entailing that $\mathcal{N}_i - Y_i$ is a martingale with respect to some filtration $(\mathscr{F}_x)_{x\geq 0}$. If the cumulative intensity can be represented as

$$Y_{i,x} = \int_0^x Z_{i,s} \, \mathrm{d}\tilde{H}_s \tag{19}$$

where $Z_i = \{Z_{i,x} : x \geq 0\}$ is an $(\mathscr{F}_x)_{x\geq 0}$ adapted process, then we have a *multiplicative intensity model*, a general class of models introduced in Aalen (1978). Moreover, if survival times $X_1, \ldots, X_n$ are subject to right–censoring, with $c_1, \ldots, c_n$ denoting the n (possibly random) censoring times and $a \wedge b := \min\{a, b\}$, then $\mathcal{N}_{i,x} = \mathbb{1}_{(0,x\wedge c_i]}(X_i)$ is equal to 1 if the i–th observation is both uncensored and not greater than x. In this case the process Z_i is such that $Z_{i,x} = \mathbb{1}_{(0,T_i]}(x)$ where $T_i = X_i \wedge c_i$ is the possibly right–censored survival time (or time to failure or time to an event) for the i–th individual. On the other hand, when data are both left and right–censored with left and right–censoring times denoted by $\boldsymbol{e} = (e_1, \ldots, e_n)$ and on $\boldsymbol{c} = (c_1, \ldots, c_n)$, respectively, both independent from the X_i's, one is led to consider $\mathcal{N}_{i,x} = \mathbb{1}_{(e_i, c_i\wedge x]}(X_i)$. Hence, conditional on $\boldsymbol{e}$ and on $\boldsymbol{c}$, $\mathcal{N}_i$ is a counting process governed by a multiplicative intensity model (19) with $Z_{i,x} = \mathbb{1}_{(e_i,c_i]}(x)$, where e_i denotes an entrance time and c_i a censoring time. The main result proved in Kim (1999) is structural conjugacy of $\tilde{H} = \{\tilde{H}_x : x \geq 0\}$ in (19). Specifically, if $\tilde{H}$ is a process with independent increments, then $\tilde{H}|\text{data}$ is again a process with independent increments and fixed points of discontinuity in correspondence to the exact observation with random jumps expressed in terms of the Lévy intensity. For the case of right–censored observations with $\tilde{H}$ generated by a beta process, Hjort's result is recovered.

In James (2006a), the author proposes a new family of priors named *spatial neutral to the right* processes: this turns out to be useful when one is interested in modelling survival times X coupled with variables Y which take values in a general space. Typically, Y can be considered as a spatial component. A spatial neutral to the right process is a random probability measure associated to a cumulative hazard at y defined by

$$\tilde{H}_t(\mathrm{d}y) = \int_{(0,t]} \tilde{\mu}(\mathrm{d}x, \mathrm{d}y)$$

where $\tilde{\mu}$ is some CRM on $\mathbb{R}^+ \times \mathbb{Y}$ and $\mathbb{Y}$ is some complete and separable metric space. Hence, by (7), $\tilde{\mu}(\mathrm{d}x, \mathrm{d}y) = \int_{[0,1]} s\, N(\mathrm{d}s, \mathrm{d}x, \mathrm{d}y)$ where N is a Poisson random measure on $[0,1] \times \mathbb{R}^+ \times \mathbb{Y}$ whose intensity is

$$\nu(\mathrm{d}s, \mathrm{d}x, \mathrm{d}y) = \rho_x(\mathrm{d}s)\mathrm{d}A_0(x, y).$$

In accordance with the previous notation, ρ_x is, for any x in $\mathbb{R}^+$, a measure on $[0,1]$ and A_0 is some hazard measure on $\mathbb{R}^+ \times \mathbb{Y}$ which plays the role of baseline hazard. Correspondingly, one has

$$\tilde{S}(t^-) = 1 - \tilde{F}(t^-) = \exp\left\{ \int_{[0,1]\times(0,t)\times\mathbb{Y}} \log(1-s)\, N(\mathrm{d}s, \mathrm{d}x, \mathrm{d}y) \right\}$$

and $\tilde{p}(\mathrm{d}x, \mathrm{d}y) = \tilde{S}(x^-)\, \tilde{\mu}(\mathrm{d}x,\, \mathrm{d}y)$ is the random probability measure on $\mathbb{R}^+ \times \mathbb{Y}$ whose law acts as a prior for the distribution of (X, Y). James (2006a) shows also that the posterior distribution of $\tilde{p}$, given a sample of exchangeable observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, arises as the sum of two independent components: one has a similar form as the prior, the only difference being an updating of $\tilde{S}$ and $\tilde{\mu}$, and the other is given by fixed points of discontinuity corresponding to the distinct observations. The analysis provided by James (2006a) also offers an algorithm for sampling from the marginal distribution of the observations, which represents an analogue of the Blackwell-MacQueen urn scheme for these more general priors. Finally, as pointed out in James (2006a), there are some nice connections between this area of research in Bayesian nonparametrics and the theory of regenerative composition structures in combinatorics. See Gnedin and Pitman (2005b).

2.3. Priors for hazard rates. A number of papers have focused on the issue of specifying a prior for the hazard rate, instead of the cumulative hazard. For simplicity we assume that the data are generated by a p.d. on $\mathbb{R}^+$ which is absolutely continuous with respect to the Lebesgue measure. Then, the hazard rate is $h(x) = F'(x)/[1 - F(x^-)]$ and a prior for it can be defined in terms of a mixture with respect to a CRM. Let $k(\,\cdot\,|\,\cdot\,)$ be some kernel on $\mathbb{R}^+ \times \mathbb{Y}$, i.e. k is bimeasurable and for any bounded $B \in \mathscr{B}(\mathbb{R}^+)$ one has $\int_B k(x|y)\,\mathrm{d}x < \infty$ for any $y \in \mathbb{Y}$. Then, a prior for the hazard rate coincides with the p.d. of the random hazard rate defined by

$$\tilde{h}(x) = \int_{\mathbb{Y}} k(x|y)\, \tilde{\mu}(\mathrm{d}y) \tag{20}$$

where $\tilde{\mu}$ is a CRM on $(\mathbb{Y}, \mathscr{Y})$. The corresponding cumulative hazard is clearly given by $\tilde{H}_x = \int_0^x \tilde{h}(s)\mathrm{d}s$. From (20), provided $\tilde{H}_x \to \infty$ for $x \to \infty$ almost surely, one can define a random density function $\tilde{f}$ as

$$\tilde{f}(x) = \tilde{h}(x)\ \mathrm{e}^{-\tilde{H}_x}$$

where $\tilde{S}(x) = \exp(-\tilde{H}_x)$ is the survival function at x. Such models are often referred to as life–testing models. The random hazard $\tilde{h}$ in (20) can also be used to define the intensity rate of a counting process $\mathcal{N}_i = \{\mathcal{N}_{i,x} :\ x \geq 0\}$ as $Z_{i,x}\, \tilde{h}(x)$ where $Z_i = \{Z_{i,x} :\ x \geq 0\}$ is a process satisfying the same conditions pointed out in Kim (1999). Various specific models proposed in the literature fit within this framework according to the choices of k, $\tilde{\mu}$ and Z_i. For example, Dykstra and Laud (1981) consider the case where $k(x|y) \equiv \mathbb{1}_{(0,x]}(y)\beta(x)$ for some measurable and non–negative function β, $Z_i = \mathbb{1}_{(0,T_i]}(x)$ and $\tilde{\mu}$ is a gamma process characterized by the Lévy intensity (5).

The random hazard $\tilde{h} = \{\tilde{h}(x) :\ x \geq 0\}$ corresponding to the mixing kernel described above is termed *extended gamma process* with parameters α and β in Dykstra and Laud (1981) and is again

an independent increment process with non–homogeneous Lévy intensity

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\mathrm{e}^{-\beta(x)^{-1}\, s}}{s}\, \mathrm{d}s\, \alpha(\mathrm{d}x). \tag{21}$$

Lo and Weng (1989) consider $\tilde{h}$ in (20) with a generic kernel k and process Z_i, and with $\tilde{\mu}$ an extended gamma process, or weighted gamma process in their terminology. Due to linearity of the relation in (20), a characterization of the posterior distribution of $\tilde{\mu}$ given the data would easily yield a posterior representation of the hazard rate $\tilde{h}$. In order to determine a posterior characterization of $\tilde{\mu}$, it is convenient to interpret the variable y in the kernel $k(\,\cdot\,|y)$ as a latent variable: hence the posterior distribution of $\tilde{\mu}$ arises by mixing the conditional distribution of $\tilde{\mu}$, given the data and the latent, with respect to the posterior distribution of the latent variables, given the data. Such a strategy is pursued in James (2005) where the author achieves an explicit posterior characterization for general multiplicative intensity models with mixture random hazards (20) driven by a generic CRM $\tilde{\mu}$. For brevity, here we focus on the simple life–testing model case with exact observations denoted by $\boldsymbol{X} = (X_1, \ldots, X_n)$. The likelihood function is then given by

$$\mathscr{L}(\mu; \boldsymbol{x}) = \mathrm{e}^{-\int_{\mathbb{Y}} K_n(y)\mu(\mathrm{d}y)} \prod_{i=1}^{n} \int_{\mathbb{Y}} k(x_i|y)\mu(\mathrm{d}y), \tag{22}$$

where $K_n(y) = \sum_{i=1}^{n} \int_0^{x_i} k(s|y)\mathrm{d}s$. Now, augmenting the likelihood with respect to the latent variables $\boldsymbol{y} = (y_1, \ldots, y_n)$, (22) reduces to

$$\mathscr{L}(\mu; \boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\int_{\mathbb{Y}} K_n(y)\mu(\mathrm{d}y)} \prod_{i=1}^{n} k(x_i|y_i)\mu(\mathrm{d}x_i) = \mathrm{e}^{-\int_{\mathbb{Y}} K_n(y)\mu(\mathrm{d}y)} \prod_{j=1}^{k} \mu(\mathrm{d}y_j^*)^{n_j} \prod_{i \in C_j} k(x_i|y_j^*),$$

where $\boldsymbol{y}^* = (y_1^*, \ldots, y_k^*)$ denotes the vector of the $k \leq n$ distinct latent variables, n_j is the frequency of y_j^* and $C_j = \{r : y_r = y_j^*\}$. We are now in a position to state the posterior characterization of the mixture hazard rate.

Theorem 9. (James, 2005). *Let $\tilde{h}$ be a random hazard rate as defined in* (20)*. Then, given $\boldsymbol{X}$ and $\boldsymbol{Y}$, the posterior distribution of $\tilde{\mu}$ coincides with*

$$\tilde{\mu}^* \stackrel{d}{=} \tilde{\mu}_c^* + \sum_{i=1}^{k} J_i\, \delta_{Y_j^*} \tag{23}$$

where $\tilde{\mu}_c^*$ is a CRM with intensity measure

$$\nu^*(\mathrm{d}s, \mathrm{d}y) = \mathrm{e}^{-s\, K_n(y)}\, \rho_y(\mathrm{d}\, s)\, \alpha(\mathrm{d}y), \tag{24}$$

the jumps J_i $(i = 1, \ldots, k)$ are mutually independent, independent from $\tilde{\mu}_c^*$ and their distribution can be described in terms of the Lévy intensity of $\tilde{\mu}$.

Hence, we have again the posterior structure of an updated CRM with fixed points of discontinuity, the only difference being that in such a mixture setup one has to deal with both latent variables and observables. Moreover, the p.d. of the jumps J_i's corresponding to the latents Y_i^*'s is

$$G_i(\mathrm{d}s) \propto s^{n_i} \mathrm{e}^{-s\, K_n(y_i^*)} \rho_{y_i^*}(\mathrm{d}s).$$

To complete the description the distribution of the latent variables $\boldsymbol{Y}$ conditionally on the data is needed. Setting $\tau_{n_j}(u|y) = \int_{\mathbb{R}^+} s^{n_j} \mathrm{e}^{-us} \rho_y(\mathrm{d}s)$ for any $u > 0$, one has

$$f(\mathrm{d}y_1^*, .., \mathrm{d}y_k^* | \boldsymbol{X}) = \frac{\prod_{j=1}^k \tau_{n_j}(K_n(y_j^*)|y_j^*) \prod_{i \in C_j} k(x_i, y_j^*) \alpha(\mathrm{d}y_j^*)}{\sum_{k=1}^n \sum_{\boldsymbol{n} \in \Delta_{k,n}} \prod_{j=1}^k \int_{\mathbb{Y}} \tau_{n_j}(K_n(y)|y) \prod_{i \in C_j} k(x_i, y) \alpha(\mathrm{d}x)} \tag{25}$$

for any $k \in \{1, \ldots, n\}$ and $\boldsymbol{n} := (n_1, \ldots, n_k) \in \Delta_{n,k} := \{(n_1, \ldots, n_k) : n_j \geq 1, \sum_{j=1}^k n_j = n\}$. We also recall that an alternative posterior characterization, valid when modelling decreasing hazard rate functions, has been provided in Ho (2006) and it is formulated in terms of $\boldsymbol{S}$–paths. In the light of Theorem 9, the distribution of $\tilde{\mu}$, given $\boldsymbol{X}$, can in principle be evaluated exactly by integrating (23) with respect to (25). Performing such an integration is a very difficult task since one needs to average with respect to all possible partitions of the integers $\{1, \ldots, n\}$. Nonetheless, the posterior representation is crucial for devising suitable simulation algorithms such as those provided in Nieto–Barajas and Walker (2004) and Ishwaran and James (2004). The latter paper contains also a wealth of applications, which highlight the power of the mixture model approach to multiplicative intensity models.

A variation of the use of weighted gamma or of beta processes for modelling hazards is suggested in Nieto–Barajas and Walker (2002). Consider a sequence $(t_n)_{n \geq 1}$ of ordered points, $0 < t_1 < t_2 < \cdots$, and set λ_k to be the hazard in the interval $(t_{k-1}, t_k]$. A first attempt to model the different hazard rates might be based on independence of the λ_k's: this is done in Walker and Mallick (1997) where the λ_k's are taken to be independent gamma random variables. Alternatively, a discrete version of Hjort's model implies that, given a set of failure or death times $\{t_1, t_2, \ldots\}$, the hazard rates $\pi_k = \mathbb{P}[T = t_k \,|\, T \geq t_k]$ are independent beta–distributed random variables. However, in both cases it seems sensible to assume dependence among the λ_k's or among the π_k's. The simplest form of dependence one might introduce is Markovian and this is pursued in Nieto–Barajas and Walker (2002). Hence, if θ_k is the parameter of interest, one may set $\mathbb{E}[\theta_{k+1}|\theta_1, \ldots, \theta_k] = f(\theta_k)$ for some function f. This assumption gives rise to what the authors name *Markov gamma and beta* processes. The most interesting feature is that, conditionally on a latent variable, the hazard rates have a very simple structure which naturally yields an MCMC simulation scheme for posterior inferences. An early contribution to this approach is due to Arjas and Gasbarra (1994).

3 General classes of discrete nonparametric priors

In this Section we will describe in some detail a few recent probabilistic models that are natural candidates for defining nonparametric priors Q which select discrete distributions with probability 1. There are essentially two ways for exploiting such priors: a) they can be used to model directly the data when these are generated by a discrete distribution; b) they are introduced as basic building blocks in hierarchical mixtures if the data arise from a continuous distribution. The latter use will be detailed in Section 4.1.

3.1. Normalized random measures with independent increments. Among the various generalizations of the Dirichlet process, the one we will illustrate in the present section is inspired by a construction of the Dirichlet process provided in Ferguson (1973). Indeed, a Dirichlet process on a complete and separable metric space, $\mathbb{X}$, can also be obtained by normalizing the increments

of a gamma CRM $\tilde{\gamma}$ with parameter α as described in Example 1: the random probability measure $\tilde{p} = \tilde{\gamma}/\tilde{\gamma}(\mathbb{X})$ has the same distribution as the Dirichlet process on $\mathbb{X}$ with parameter measure α. Given this, one might naturally wonder whether a full Bayesian analysis can be performed if in the above normalization the gamma process is replaced by any CRM with a generic Lévy intensity $\nu(\mathrm{d}s, \mathrm{d}x) = \rho_x(\mathrm{d}s)\,\alpha(\mathrm{d}x)$. Though Bayesians have seldom considered "normalization" as a tool for defining random probability measures, this idea has been exploited and applied in a variety of contexts not closely related to Bayesian inference such as storage problems, computer science, population genetics, ecology, statistical physics, combinatorics, number theory and excursions of stochastic processes. See Pitman (2006) and references therein. Some important theoretical insight on the properties of normalized random measures was first given in Kingman (1975), where a random discrete distribution generated by the σ–stable subordinator is considered. Further developments can be found in Perman, Pitman and Yor (1992), where a description of the atoms of random probability measures, obtained by normalizing increasing processes with independent and stationary increments, in terms of a stick-breaking procedure, is provided. From a Bayesian perspective, the idea of normalization has been taken up again in Regazzini, Lijoi and Prünster (2003), where a normalized random measure with independent increments is introduced as a random probability measure on $\mathbb{R}$ obtained by normalizing a suitably time-changed increasing process with independent but not necessarily stationary increments. A definition stated in terms of CRMs is as follows.

Definition 3. Let $\tilde{\mu}$ be a CRM on $\mathbb{X}$ such that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ almost surely. Then, the random probability measure $\tilde{p} = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$ is termed *normalized random measure with independent increments* (NRMI).

Both finiteness and positiveness of $\tilde{\mu}(\mathbb{X})$ are clearly required for the normalization to be well–defined and it is natural to express such conditions in terms of the Lévy intensity of the CRM. Indeed, it is enough to have $\rho_x(\mathbb{R}^+) = \infty$ for every x and $0 < \alpha(\mathbb{X}) < \infty$. The former is equivalent to requiring that the CRM $\tilde{\mu}$ has infinitely many jumps on any bounded set: in this case $\tilde{\mu}$ is also called an *infinite activity* process. The previous conditions can also be strengthened to necessary and sufficient conditions but we do not pursue this here.
In the following we will speak of *homogeneous (non–homogeneous)* NRMIs, if the underlying CRM (or, equivalently, the Lévy intensity (4)) is homogeneous (non–homogeneous).

Example 5. (The σ–stable NRMI). Suppose $\sigma \in (0, 1)$ and let $\tilde{\mu}_\sigma$ be the σ–stable CRM examined in Example 2 with Lévy intensity (6). If α in (6) is finite, the required positivity and finiteness conditions are satisfied. One can, then, define a random probability measure $\tilde{p} = \tilde{\mu}_\sigma/\tilde{\mu}_\sigma(\mathbb{X})$ which takes on the name of *normalized σ–stable process* with parameter σ. This random probability measure was introduced in Kingman (1975) in relation to optimal storage problems. The possibility of application in Bayesian nonparametric inference was originally pointed out by A.F.M. Smith in the Discussion to Kingman (1975).

Example 6. (The generalized gamma NRMI). Consider now a generalized gamma CRM (Brix, 1999) which is characterized by a Lévy intensity of the form

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\sigma}{\Gamma(1-\sigma)}\, s^{-1-\sigma}\, \mathrm{e}^{-\tau s}\mathrm{d}s\, \alpha(\mathrm{d}x), \tag{26}$$

where $\sigma \in (0,1)$ and $\tau > 0$. Let us denote it by $\tilde{\mu}_{\sigma,\tau}$. Note that if $\tau = 0$ then $\tilde{\mu}_{\sigma,0}$ coincides with the σ–stable CRM $\tilde{\mu}_\sigma$, whereas if $\sigma \to 0$ the gamma CRM (5) is obtained. If α in (26) is finite, we have $0 < \tilde{\mu}_{\sigma,\tau}(\mathbb{X}) < \infty$ almost surely and a NRMI $\tilde{p} = \tilde{\mu}_{\sigma,\tau}/\tilde{\mu}_{\sigma,\tau}(\mathbb{X})$, which is termed *normalized generalized gamma process*. See Pitman (2003) for a discussion on its representation as Poisson–Kingman model, a class of random distributions described in Section 3.3. The special case of $\sigma = 1/2$, corresponding to the *normalized inverse Gaussian process*, has been examined in Lijoi, Mena and Prünster (2005) who also provide an expression for the family of finite dimensional distributions of $\tilde{p}$. □

Example 7. (THE EXTENDED GAMMA NRMI) A non–homogeneous NRMI arises by considering the extended gamma process of Dykstra and Laud (1981) characterized by the Lévy intensity (21). If the function $\beta : \mathbb{X} \to \mathbb{R}^+$ is such that $\int_{\mathbb{X}} \log[1+\beta(x)]\,\alpha(\mathrm{d}x) < \infty$, then the corresponding NRMI is well–defined and will be termed *extended gamma NRMI.* □

These examples, together with others one could think of by simply providing a Lévy intensity, suggest that NRMIs identify a very large class of priors and one might then wonder whether they are amenable of practical use for inferential purposes. A first thing to remark is that, apart from the Dirichlet process, NMRIs are not structurally conjugate. See James, Lijoi and Prünster (2006). Nonetheless one can still provide a posterior characterization of NRMIs in the form of a mixture representation. In the sequel, we will always work with NRMI, whose underlying Lévy intensity has a non–atomic α in (4). Suppose that the data are exchangeable according to model (1) where Q is the probability distribution of a NRMI. Since NRMIs are almost surely discrete, data can display ties and we denote by $X_1^*, \ldots, X_k^*$ the k distinct observations, with frequencies $n_1, \ldots, n_k$, present within the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. Before stating the posterior characterization, we introduce the key latent variable. For any $n \geq 1$, let U_n be a positive random variable whose density function, conditional on the sample $\boldsymbol{X}$, is

$$q_{\boldsymbol{X}}(u) \propto u^{n-1}\,\mathrm{e}^{-\psi(u)} \prod_{j=1}^{k} \tau_{n_j}(u|X_j^*), \tag{27}$$

where ψ is the *Laplace exponent* of $\tilde{\mu}$, *i.e.*

$$\psi(u) = \int_{\mathbb{X}} \int_{\mathbb{R}^+} (1 - \mathrm{e}^{-u\,v}) \rho_x(\mathrm{d}v)\,\alpha(\mathrm{d}x)$$

and, for any $m \geq 1$, $\tau_m(u|x) := \int_{\mathbb{R}^+} s^m\,\mathrm{e}^{-us}\,\rho_x(\mathrm{d}s)$. The following result states that the posterior distribution of $\tilde{\mu}$ and of $\tilde{p}$, given a sample $\boldsymbol{X}$, is a mixture of NRMIs with fixed points of discontinuity in correspondence to the observations and the mixing density is $q_{\boldsymbol{X}}$ in (27).

Theorem 10. (James, Lijoi and Prünster, 2005 & 2009). *If $\tilde{p}$ is a NRMI obtained by normalizing $\tilde{\mu}$, then*

$$\tilde{\mu}\,|\,(\boldsymbol{X}, U_n) \stackrel{d}{=} \tilde{\mu}_{U_n} + \sum_{i=1}^{k} J_i^{(U_n)}\,\delta_{X_i^*}, \tag{28}$$

where $\tilde{\mu}_{U_n}$ is a CRM with Lévy intensity $\nu^{(U_n)}(\mathrm{d}s, \mathrm{d}x) = \mathrm{e}^{-U_n s}\,\rho_x(\mathrm{d}s)\,\alpha(\mathrm{d}x)$, the non–negative jumps $J_i^{(U_n)}$'s are mutually independent and independent from $\tilde{\mu}_{U_n}$ with density function $f_i(s) \propto s^{n_i}\mathrm{e}^{-U_n s}\rho_{X_i^*}(\mathrm{d}s)$. Moreover,

$$\tilde{p}\,|\,(\boldsymbol{X}, U_n) \stackrel{d}{=} w\frac{\tilde{\mu}_{U_n}}{\tilde{\mu}_{U_n}(\mathbb{X})} + (1-w)\frac{\sum_{i=1}^{k} J_i^{(U_n)}\,\delta_{X_i^*}}{\sum_{r=1}^{k} J_r^{(U_n)}} \tag{29}$$

where $w = \tilde{\mu}_{U_n}(\mathbb{X})/[\tilde{\mu}_{U_n}(\mathbb{X}) + \sum_{i=1}^{k} J_i^{(U_n)}]$.

The above result displays the same posterior structure, namely CRM with fixed points of discontinuity, that has already occurred on several occasions in Section 2: here the only difference is that such a representation holds conditionally on a suitable latent variable, which makes it slightly more elaborate. This is due to the fact that the structural conjugacy property is not satisfied. Nonetheless, NRMIs give rise to more manageable predictive structures than, for instance, NTR processes. See also James, Lijoi and Prünster (2009).

Since the Dirichlet process is a special case of NRMI, it is interesting to see how the posterior representation of Ferguson (1973) is recovered. Indeed, if $\tilde{\mu}$ is a gamma CRM with parameter measure α on $\mathbb{X}$ such that $\alpha(\mathbb{X}) = \theta \in (0, \infty)$, then $\tilde{\mu}_{U_n} + \sum_{i=1}^{k} J_i^{(U_n)} \delta_{X_i^*}$ is a gamma CRM with Lévy intensity

$$\nu^{(U_n)}(\mathrm{d}s, \mathrm{d}x) = \frac{\mathrm{e}^{-(1+U_n)s}}{s} \, \mathrm{d}s \, \alpha_n^*(\mathrm{d}x) \tag{30}$$

where $\alpha_n^* = \alpha + \sum_{i=1}^{k} n_i \, \delta_{X_i^*}$. However, since the CRM characterized by (30) is to be normalized, we can, without loss of generality, set the scale parameter $1 + U_n$ in (30) equal to 1. The random probability in (29) turns out to be a Dirichlet process with parameter α_n^* and its distribution does not depend on U_n. Note also the analogy with the posterior updating of the beta process sketched after Theorem 8.

In analogy with NTR processes, the availability of a posterior representation is essential for the implementation of sampling algorithms in order to simulate the trajectories of the posterior CRM. A possible algorithm suggested by the representation (28) consists in

(i) Sample U_n from $q_{\boldsymbol{X}}$

(ii) Sample the jump $J_i^{(U_n)}$ at X_i^* from the density $f_i(s) \propto s^{n_i} \mathrm{e}^{-U_n s} \rho_{X_i^*}(\mathrm{d}s)$

(iii) Simulate a realization of $\tilde{\mu}_{U_n}$ with Lévy measure $\nu^{(U_n)}(\mathrm{d}x, \mathrm{d}s) = \mathrm{e}^{-U_n s} \, \rho_x(\mathrm{d}s) \, \alpha(\mathrm{d}x)$ via the Ferguson and Klass algorithm. See Ferguson and Klass (1972) and Walker and Damien (2000).

For an application of this computational technique see Nieto–Barajas and Prünster (2009).

Example 8. (The generalized gamma NRMI). Consider the normalized generalized gamma process defined in Example 6. The (posterior) distribution of $\tilde{\mu}$, given U_n and $\boldsymbol{X}$, coincides in distribution with the CRM $\tilde{\mu}_{U_n} + \sum_{i=1}^{k} J_i^{(U_n)} \, \delta_{X_i^*}$ where $\tilde{\mu}_{U_n}$ is a generalized gamma CRM with Lévy intensity $\nu^{(U_n)}(\mathrm{d}s, \mathrm{d}x) = \frac{\sigma}{\Gamma(1-\sigma)} \, s^{-1-\sigma} \, \mathrm{e}^{-(U_n+1)s} \mathrm{d}s \, \alpha(\mathrm{d}x)$, the fixed points of discontinuity coincide with the distinct observations X_i^* and the i–th jump $J_i^{(U_n)}$ is $\textsc{Gamma}(U_n+1, n_i-\sigma)$ distributed, for $i = 1, \ldots, k$. Finally, the density function of U_n, conditional on $\boldsymbol{X}$, is $q_{\boldsymbol{X}}(u) \propto u^{n-1} \, (1+u)^{k\sigma-n} \, \mathrm{e}^{-\alpha(\mathbb{X})(1+u)^{\sigma}}$. □

3.2. Exchangeable partition probability function. The nature of the realizations of NRMIs and, in general, of discrete random probability measures, quite naturally leads to analyze the partition structures among the observations that they generate. Indeed, given n observations $X_1, \ldots, X_n$ generated from model (1), discreteness of $\tilde{p}$ implies that there might be ties within the data, *i.e.* $\mathbb{P}[X_i = X_j] > 0$ for $i \neq j$. Correspondingly, define Ψ_n to be a random partition of the integers $\{1, \ldots, n\}$ such that any two integers i and j belong to the same set in Ψ_n if and only if $X_i = X_j$. Let $k \in \{1, \ldots, n\}$ and suppose $\{C_1, \ldots, C_k\}$ is a partition of $\{1, \ldots, n\}$ into k sets C_i. Hence, $\{C_1, \ldots, C_k\}$

is a possible realization of Ψ_n. A common and sensible specification for the probability distribution of Ψ_n consists in assuming that it depends on the frequencies of each set in the partition. To illustrate this point, recall that

$$\Delta_{n,k} := \left\{ (n_1, \ldots, n_k) : \ n_i \geq 1, \ \sum_{i=1}^{k} n_i = n \right\}$$

For $n_i = \mathrm{card}(C_i)$, then $(n_1, \ldots, n_k) \in \Delta_{n,k}$ and

$$\mathbb{P}[\Psi_n = \{C_1, \ldots, C_k\}] = \Pi_k^{(n)}(n_1, \ldots, n_k) \tag{31}$$

A useful and intuitive metaphor is that of species sampling: one is not much interested into the realizations of the X_i's, which stand as species labels thus being arbitrary, but rather in the probability of observing k distinct species with frequencies $(n_1, \ldots, n_k)$ in $n \geq k$ draws from a population.

Definition 4. Let $(X_n)_{n\geq 1}$ be an exchangeable sequence. Then, $\{\Pi_k^{(n)} : 1 \leq k \leq n, \ n \geq 1\}$ with $\Pi_k^{(n)}$ defined in (31) is termed *exchangeable partition probability function* (EPPF).

Indeed, the EPPF defines an important tool which has been introduced in Pitman (1995) and it determines the distribution of a random partition of $\mathbb{N}$. It is worth noting that the fundamental contributions J. Pitman has given to this area of research have been deeply influenced by, and appear as natural developments of, some earlier relevant work on random partitions by J.F.C. Kingman. See, *e.g.*, Kingman (1978, 1982).

From the above definition it follows that, for any $n \geq k \geq 1$ and any $(n_1, \ldots, n_k) \in \Delta_{n,k}$, $\Pi_k^{(n)}$ is a symmetric function of its arguments and it satisfies the addition rule $\Pi_k^{(n)}(n_1, \ldots, n_k) = \Pi_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} \Pi_k^{(n+1)}(n_1, \ldots, n_j + 1, \ldots, n_k)$. On the other hand, as shown in Pitman (1995), every non–negative symmetric function satisfying the addition rule is the EPPF of some exchangeable sequence. See Pitman (1995, 2006) for a thorough and useful analysis of EPPFs.

The availability of the EPPF yields, as a by–product, the system of predictive distributions induced by Q. Indeed, suppose Q in model (1) coincides with a discrete nonparametric prior and $\{\Pi_k^{(n)} : 1 \leq k \leq n, \ n \geq 1\}$ is the associated EPPF. If the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ contains k distinct values $X_1^*, \ldots, X_k^*$ and n_j of them are equal to X_j^* one has

$$\mathbb{P}[X_{n+1} = \text{ new} \,|\, \boldsymbol{X}] = \frac{\Pi_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1)}{\Pi_k^{(n)}(n_1, \ldots, n_k)}, \qquad \mathbb{P}[X_{n+1} = X_j^* \,|\, \boldsymbol{X}] = \frac{\Pi_k^{(n+1)}(n_1, \ldots, n_j + 1, \ldots, n_k)}{\Pi_k^{(n)}(n_1, \ldots, n_k)}$$

If $\tilde{p}$ is a NRMI (with non–atomic parameter measure α), the associated EPPF is

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{1}{\Gamma(n)} \int_0^{\infty} u^{n-1} \, \mathrm{e}^{-\psi(u)} \left\{ \prod_{j=1}^{k} \int_{\mathbb{X}} \tau_{n_j}(u|x) \, \alpha(\mathrm{d}x) \right\} \mathrm{d}u \tag{32}$$

and from it one can deduce the system of predictive distributions of X_{n+1}, given $\boldsymbol{X}$,

$$\mathbb{P}\left[X_{n+1} \in \mathrm{d}x_{n+1} \,|\, \boldsymbol{X}\right] = w_k^{(n)} \, P_0(\mathrm{d}x_{n+1}) + \frac{1}{n} \sum_{j=1}^{k} w_{j,k}^{(n)} \, \delta_{X_j^*}(\mathrm{d}x_{n+1}) \tag{33}$$

where $P_0 = \alpha/\alpha(\mathbb{X})$ and

$$w_k^{(n)} = \frac{1}{n} \int_0^{+\infty} u \, \tau_1(u|x_{n+1}) \, q_{\boldsymbol{X}}(u) \, \mathrm{d}u \qquad w_{j,k}^{(n)} = \int_0^{+\infty} u \, \frac{\tau_{n_j+1}(u|X_j^*)}{\tau_{n_j}(u|X_j^*)} \, q_{\boldsymbol{X}}(u) \, \mathrm{d}u \tag{34}$$

In the homogeneous case, *i.e* $\rho_x = \rho$, the previous formulae reduce to those given in Pitman (2003). Closed form expressions are derivable for some specific NRMI. For example, if $\tilde{p}$ is the σ–stable NRMI, then $w_k^{(n)} = k\,\sigma/n$ and $w_{j,k}^{(n)} = (n_j - \sigma)$. See Pitman (1996). On the other hand, if $\tilde{p}$ is the normalized generalized gamma process, one has (33) with

$$w_k^{(n)} = \frac{\sigma}{n}\,\frac{\sum_{i=0}^{n}\binom{n}{i}(-1)^i\,\beta^{i/\sigma}\Gamma\left(k+1-\frac{i}{\sigma};\,\beta\right)}{\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\,\beta^{i/\sigma}\,\Gamma\left(k-\frac{i}{\sigma};\,\beta\right)} \qquad w_{j,k}^{(n)} = \frac{\sum_{i=0}^{n}\binom{n}{i}(-1)^i\,\beta^{i/\sigma}\,\Gamma\left(k-\frac{i}{\sigma};\,\beta\right)}{\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\beta^{i/\sigma}\,\Gamma\left(k-\frac{i}{\sigma};\,\beta\right)}\,(n_j-\sigma).$$

See Lijoi, Mena and Prünster (2007a). The availability of closed form expressions of the predictive distributions is essential for the implementation of Blackwell–MacQueen–type sampling schemes, which are a key tool for drawing inference in complex mixture models. Nonetheless, even when no closed form expressions are available, drawing samples from the predictive is still possible by conditioning on the latent variable U_n. Indeed, one has

$$\mathbb{P}[X_{n+1}\in \mathrm{d}x_{n+1}\,|\,\boldsymbol{X}, U_n = u] \propto \kappa_1(u)\,\tau_1(u|x_{n+1})\,P_0(\mathrm{d}x_{n+1}) + \sum_{j=1}^{k}\frac{\tau_{n_j+1}(u|X_j^*)}{\tau_{n_j}(u|X_j^*)}\,\delta_{X_j^*}(\mathrm{d}x_{n+1})$$

where $\kappa_1(u) = \int_{\mathbb{X}}\tau_1(u|x)\,\alpha(\mathrm{d}x)$. From this one can implement an analog of the Blackwell–MacQueen urn scheme in order to draw a sample $X_1,\ldots,X_n$ from $\tilde{p}$. Let $m(\mathrm{d}x|u) \propto \tau_1(u|x)\,\alpha(\mathrm{d}x)$ and, for any $i \geq 2$, set $m(\mathrm{d}x_i|x_1,\ldots,x_{i-1},u) = \mathbb{P}[X_i \in \mathrm{d}x_i|X_1,\ldots,X_{i-1},U_{i-1}=u]$. Moreover, set U_0 to be a positive random variable whose density function is given by $q_0(u) \propto \mathrm{e}^{-\psi(u)}\,\int_{\mathbb{X}}\tau_1(u|x)\,\alpha(\mathrm{d}x)$. The sampling scheme can be described as follows

1) Sample U_0 from q_0
2) Sample X_1 from $m(\mathrm{d}x|U_0)$
3) At step i

   3a) Sample U_{i-1} from $q_{\boldsymbol{X}_{i-1}}(u)$, where $\boldsymbol{X}_{i-1} = (X_1,\ldots,X_{i-1})$

   3b) Generate ξ_i from $f_i(\xi) \propto \tau_1(U_{i-1}|\xi)\,P_0(\mathrm{d}\xi)$

   3c) Sample X_i from $m(\mathrm{d}x|\boldsymbol{X}_{i-1},U_{i-1})$ which implies

$$X_i = \begin{cases} \xi_i & \text{prob } \propto \kappa_1(U_{i-1}) \\ X_{j,i-1}^* & \text{prob } \propto \tau_{n_{j,i-1}+1}(U_{i-1}|X_{j,i-1}^*)/\tau_{n_{j,i-1}+1}(U_{i-1}|X_{j,i-1}^*) \end{cases}$$

   where $X_{j,i-1}^*$ is the j–th distinct value among $X_1,\ldots,X_{i-1}$ and $n_{j,i-1}$ is the cardinality of the set $\{X_s:\ X_s = X_{j,i-1}^*,\ \ s=1,\ldots,i-1\}$.

3.3. Poisson-Kingman models and Gibbs–type priors. Consider a discrete random probability measure $\tilde{p} = \sum_{i\geq 1}\tilde{p}_i\delta_{X_i}$ where the locations X_i's are i.i.d. from a non–atomic probability measure P_0 on $\mathbb{X}$. Furthermore, suppose the locations are independent from the weights $\tilde{p}_i$'s. The specification of $\tilde{p}$ is completed by assigning a distribution for the weights. Pitman (2003) identifies a method for achieving this goal: he derives laws, which are termed Poisson–Kingman distributions, for sequences of ranked random probability masses $\tilde{p}_i$'s. To be more specific, consider a homogeneous CRM $\tilde{\mu}$ whose intensity $\nu(\mathrm{d}s,\mathrm{d}x) = \rho(\mathrm{d}s)\alpha(\mathrm{d}x)$ is such that $\rho(\mathbb{R}^+) = \infty$ and $\alpha = P_0$ is a non–atomic probability measure. Denote by $J_{(1)} \geq J_{(2)} \geq \cdots$ the ranked jumps of the CRM, set $T = \sum_{i\geq 1}J_{(i)}$

and assume that the p.d. of the total mass T is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. Next, define

$$\tilde{p}_{(i)} = \frac{J_{(i)}}{T} \tag{35}$$

for any $i = 1, 2, \ldots$ and denote by $S^* = \{(p_1, p_2, \ldots) : p_1 \geq p_2 \geq \cdots \geq 0, \ \sum_{i\geq 1} p_i = 1\}$ the set of all sequences of ordered non–negative real numbers that sum up to 1.

Definition 5. Let $P_{\rho,t}$ be the conditional distribution of the sequence $(\tilde{p}_{(i)})_{i\geq 1}$ of ranked random probabilities generated by a CRM through (35), given $T = t$. Let η be a probability distribution on $\mathbb{R}^+$. The distribution

$$\int_{\mathbb{R}^+} P_{\rho,t}\, \eta(\mathrm{d}t)$$

on S^* is termed *Poisson–Kingman distribution* with Lévy intensity ρ and mixing distribution η. It is denoted by $\mathrm{PK}(\rho, \eta)$.

If η coincides with the p.d. of T, we use the notation $\mathrm{PK}(\rho)$ to indicate the corresponding random probability with masses in S^*. The discrete random probability measure $\tilde{p} = \sum_{i\geq 1} \tilde{p}_{(i)}\delta_{X_i}$, where the $\tilde{p}_{(i)}$'s follow a $\mathrm{PK}(\rho, \eta)$ distribution, is termed *PK(ρ, η) random probability measure*. It is important to remark that $\mathrm{PK}(\rho)$ random probability measures are equivalent to homogeneous NRMIs defined in Section 3.1. Pitman (2003) derives an expression for the EPPF of a general $\mathrm{PK}(\rho, \eta)$ model but it is difficult to evaluate. However, in the special case of a $\mathrm{PK}(\rho)$ model it reduces to the simple expression implied by (32) when the dependence on the locations of the jumps is removed. Although the potential use of general $\mathrm{PK}(\rho, \eta)$ random probability measures for statistical inference is quite limited, their theoretical importance can be traced back to two main reasons: (i) the two parameter Poisson–Dirichlet process is a $\mathrm{PK}(\rho, \eta)$ model, whereas it is not a NRMI; (ii) $\mathrm{PK}(\rho, \eta)$ models generate the class of Gibbs–type random probability measure which possess a conceptually very appealing predictive structure. Both examples involve $\mathrm{PK}(\rho, \eta)$ models based on the σ–stable CRM.

Example 9. (The two parameter Poisson–Dirichlet process). One of the main reasons of interest for the class of $\mathrm{PK}(\rho, \eta)$ priors is due to the fact that it contains, as a special case, the *two parameter Poisson–Dirichlet process*, introduced in Perman, Pitman and Yor (1992). This process and the distribution of the ranked probabilities, termed *two parameter Poisson–Dirichlet distribution*, were further studied in the remarkable papers by Pitman (1995) and Pitman and Yor (1997a). Its name is also explained by the fact that it can be seen as a natural extension of the one parameter Poisson–Dirichlet distribution of Kingman (1975), which corresponds to the distribution of the ranked probabilities of the Dirichlet process.

Let ρ_σ be the jump part of the Lévy intensity corresponding to a σ–stable CRM, *i.e.* $\rho_\sigma(s) = \sigma s^{-1-\sigma}/\Gamma(1-\sigma)$, and consider a parameter $\theta > -\sigma$. Further denote by f_σ the density of a σ–stable random variable and define $\eta_{\sigma,\theta}(\mathrm{d}t) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\, t^{-\theta}\, f_\sigma(t)\, \mathrm{d}t$. Then, as shown in Pitman (2003), the $\mathrm{PK}(\rho_\sigma, \eta_{\sigma,\theta})$ random probability measure is a two parameter Poisson–Dirichlet process, to be abbreviated as $\mathrm{PD}(\sigma, \theta)$ process. In many recent papers, especially within the machine learning community, such a process is often referred to as *Pitman–Yor process*. In Section 3.4 we will present an alternative stick–breaking construction of the $\mathrm{PD}(\sigma, \theta)$ process.

Among all generalizations of the Dirichlet process, the PD(σ,θ) process stands out for its tractability. The EPPF, which characterizes the induced random partition, of a PD(σ,θ) process is

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1} \tag{36}$$

where $(a)_n = \Gamma(a+n)/\Gamma(a) = a(a+1)\cdots(a+n-1)$ for any $n \geq 1$ and $(a)_0 \equiv 1$. Note that if one lets $\sigma \to 0$ then the EPPF above reduces to $\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{\theta^k}{(\theta)_n} \prod_{j=1}^{k}\Gamma(n_j)$ which coincides with the EPPF for the Dirichlet process as provided by Antoniak (1974). On the other hand, if $\theta = 0$, one obtains the σ–stable NRMI presented in Example 5. Now, denote by $m_j \geq 0$, $j = 1,\ldots,n$, the number of sets in the partition which contain j objects or, using again the species metaphor, the number of species appearing j–times in a sample of size n. Then, an alternative equivalent formulation of (36), known as *Pitman's sampling formula*, is given by

$$\Pi^*(m_1,\ldots,m_n) = n!\,\frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}\prod_{i=1}^{n} m_i!} \prod_{i=1}^{n}\left[\frac{(1-\sigma)_{i-1}}{i!}\right]^{m_i}$$

for any $n \geq 1$ and $m_1,\ldots,m_n$ such that $m_i \geq 0$ and $\sum_{i=1}^{n} i\,m_i = n$. The above expression represents a two parameter generalization of the celebrated Ewens' sampling formula in population genetics, which can be recovered by letting $\sigma \to 0$. See Ewens (1972). As highlighted in Section 3.2, the availability of the EPPF in (36) allows one to determine the system of predictive distributions associated with the PD(σ,θ) process. Indeed, if $\boldsymbol{X} = (X_1,\ldots,X_n)$ is a sample consisting of k distinct values $X_1^*,\ldots,X_k^*$ and n_j of them are equal to X_j^*, then

$$\mathbb{P}[X_{n+1} \in \mathrm{d}x \,|\, \boldsymbol{X}] = \frac{\theta+k\sigma}{\theta+n}\,P_0(\mathrm{d}x) + \frac{1}{\theta+n}\sum_{j=1}^{k}(n_j-\sigma)\,\delta_{X_j^*}(\mathrm{d}x)$$

As observed in Section 1.1, for the PD(σ,θ) process the probability of observing a new value depends, in contrast to the Dirichlet process, also on the number of distinct observations. Another distinctive feature, if compared with the Dirichlet process, is represented by the asymptotic behaviour of the number of groups K_n generated by the first n observations, as $n \to \infty$. For the Dirichlet process, as shown in Korwar and Hollander (1973), $K_n \sim \theta\log(n)$ almost surely as $n \to \infty$. Hence, the number of distinct observations increases at a logarithmic rate. On the other hand, when the observations are governed by a PD(σ,θ) process, then $K_n \sim S_{\sigma,\theta}n^\sigma$ as $n \to \infty$ where $S_{\sigma,\theta}$ is a positive random variable whose p.d. has a density on $\mathbb{R}^+$ depending on σ and θ. See Pitman (2003). In other terms, the number of distinct observations under a PD(σ,θ) increases at a higher rate, n^σ, than in the Dirichlet case. □

An interesting and closely related class of random probability measures is given by Gibbs–type priors, introduced in Gnedin and Pitman (2005a). We first aim at defining such priors and highlighting some of their features. Afterwards we will explain their connection to Poisson–Kingman models.

By looking at the EPPF of the PD(σ,θ) process (36) one immediately recognizes that it arises as a product of two factors: the first one depends only on (n,k), whereas the second one depends on the frequencies $(n_1,\ldots,n_k)$ via the product $\prod_{j=1}^{k}(1-\sigma)_{n_j-1}$. This structure is the main ingredient for defining a general family of exchangeable random partitions, namely the *Gibbs–type random partitions* and the associated *Gibbs–type priors*.

Definition 6. Let $\tilde{p} = \sum_{i \geq 1} \tilde{p}_i \, \delta_{X_i}$ be a discrete random probability measure for which the locations X_i's are independent from the weights $\tilde{p}_i$'s and are i.i.d. from a non–atomic probability measure P_0 on $\mathbb{X}$. Then $\tilde{p}$ is termed *Gibbs–type random probability measure* if, for all $1 \leq k \leq n$ and for any $(n_1, \ldots, n_k)$ in $\Delta_{n,k}$, the EPPF can be represented as

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}, \tag{37}$$

for some $\sigma \in [0, 1)$. The random partition of $\mathbb{N}$ determined by (37) is termed *Gibbs–type random partition*.

It is worth noting that Gibbs–type random partitions identify particular exchangeable product partition models of the type introduced by Hartigan (1990). Indeed, if the cohesion function $c(\,\cdot\,)$ in Hartigan's definition depends on the cardinalities of the groups, a result of Gnedin and Pitman (2005a) states that it must be of the form $c(n_j) = (1-\sigma)_{n_j - 1}$ for $j = 1, \ldots, k$. See Lijoi, Mena and Prünster (2007a) for more explanations on this connection.

From (37), it follows that the predictive distributions induced by a Gibbs–type prior are of the form

$$\mathbb{P}\left[X_{n+1} \in \mathrm{d}x \,\middle|\, \boldsymbol{X}\right] = \frac{V_{n+1,k+1}}{V_{n,k}} \, P_0(\mathrm{d}x) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma) \, \delta_{X_j^*}(\mathrm{d}x). \tag{38}$$

The structure of (38) provides some insight into the inferential implications of the use of Gibbs–type priors. Indeed, the prediction rule can be seen as resulting from a two step procedure: the $(n+1)$–th observation X_{n+1} is either "new" (*i.e.* not coinciding with any of the previously observed X_i^*'s) or "old" with probability depending on n and k but not on the frequencies n_i's. Given X_{n+1} is "new", it is sampled from P_0. Given X_{n+1} is "old" (namely X_{n+1} is equal to one of the already sampled X_i^*'s), it will coincide with a particular X_j^* with probability $(n_j - \sigma)/(n - k\sigma)$. By comparing the predictive distributions (38) with those arising from the models dealt with so far, one immediately sees that the PD(σ, θ) process (hence, *a fortiori* the Dirichlet process) and the normalized generalized gamma process belong to the class of Gibbs priors. Considered as a member of this general class, the Dirichlet process is the only prior for which the probability of sampling a "new" or "old" observation does not depend on the number of distinct ones present in the sample. On the other hand, one may argue that it is desirable to have prediction rules for which the assignment to "new" or "old" depends also on the frequencies n_i's: however, this would remarkably increase the mathematical complexity and so Gibbs priors appear to represent a good compromise between tractability and richness of the predictive structure. An investigation of the predictive structure arising from Gibbs–type priors can be found in Lijoi, Prünster and Walker (2008a).

An important issue regarding the class of Gibbs–type priors is the characterization of its members. In other terms, one might wonder which random probability measures induce an EPPF of the form (37). An answer has been successfully provided by Gnedin and Pitman (2005a). Let ρ_σ be the jump part of the intensity of a σ–stable CRM and consider PK(ρ_σ, η) random probability measures with arbitrary mixing distribution η: for brevity we refer to them as the σ–stable PK models. Then, $\tilde{p}$ is a Gibbs–type prior with $\sigma \in (0, 1)$ if and only if it is a σ–stable PK model. Hence, the corresponding

$V_{n,k}$, which specify the prior completely, are of the form

$$V_{n,k} = \int_{\mathbb{R}^+} \frac{\sigma^k t^{-n}}{\Gamma(n-k\sigma)\, f_\sigma(t)} \int_0^t s^{n-k\sigma-1} f_\sigma(t-s)\, \mathrm{d}s\eta(\mathrm{d}t),$$

where f_σ denotes, as before, the σ–stable density. Moreover, $\tilde{p}$ is a Gibbs–type prior with $\sigma = 0$ if and only if it is a mixture, with respect to the parameter $\theta = \alpha(\mathbb{X})$, of a Dirichlet process. See Pitman (2003, 2006) and Gnedin and Pitman (2005a) for more details and interesting connections to combinatorics. Finally, the only NRMI, which is also of Gibbs–type with $\sigma \in (0,1)$, is the normalized generalized gamma process (Lijoi, Prünster and Walker, 2008b).

3.4. Species sampling models. Species sampling models, introduced and studied in Pitman (1996), are a very general class of discrete random probability measures $\tilde{p} = \sum_{j\geq 1} \tilde{p}_j\, \delta_{X_j}$ in which the weights $\tilde{p}_j$ are independent of the locations X_j. Such a generality provides some insight on the structural properties of these random probability measures; however, for possible uses in concrete applications, a distribution for the weights $\tilde{p}_j$'s has to be specified. Indeed, homogeneous NRMI and Poisson–Kingman models belong to this class and can be seen as completely specified species sampling models. On the other hand, NTR and non–homogeneous NRMI do not fall within this framework.

Definition 7. Let $(\tilde{p}_j)_{j\geq 1}$ be a sequence of non–negative random weights such that $\sum_{j\geq 1} \tilde{p}_j \leq 1$ and suppose that $(\xi_n)_{n\geq 1}$ is a sequence of i.i.d. random variables with non–atomic p.d. P_0. Moreover, let the ξ_i's be independent from the $\tilde{p}_j$'s. Then, the random probability measure

$$\tilde{p} = \sum_{j\geq 1} \tilde{p}_j\, \delta_{\xi_j} + \left(1 - \sum_{j\geq 1} \tilde{p}_j\right) P_0$$

is a *species sampling model*.

Accordingly, a sequence of random variables $(X_n)_{n\geq 1}$, which is conditionally i.i.d. given a species sampling model, is said to be a *species sampling sequence*. Moreover, if in the previous definition one has $\sum_{j\geq 1} \tilde{p}_j = 1$, almost surely, then the model is termed *proper*. We will focus on this specific case and provide a description of a few well–known species sampling models.

The use of the terminology species sampling is not arbitrary. Indeed, discrete nonparametric priors are not well suited for modelling directly data generated by a continuous distribution (in such cases they are used at a latent level within a hierarchical mixture). However, as already noted in Pitman (1996), when the data come from a discrete distribution as it happens for species sampling problems in ecology, biology and population genetics, it is natural to assign a discrete nonparametric prior to the unknown proportions. More precisely, suppose that a population consists of an ideally infinite number of species: one can think of $\tilde{p}_i$ as the proportion of the i–th species in the population and ξ_i is the label assigned to species i. Since the labels ξ_i are generated by a non–atomic distribution they are almost surely distinct: hence, distinct species will have distinct labels attached. The following characterization provides a formal description of the family of predictive distributions induced by a species sampling model.

Theorem 11. (Pitman, 1996) *Let $(\xi_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with p.d. P_0. Then $(X_n)_{n\geq 1}$ is a species sampling sequence if and only if there exists a collection of weights*

$\{p_{j,n}(n_1,\ldots,n_k):\ 1\le j\le k,\ 1\le k\le n,\ n\ge 1\}$ *such that* $X_1=\xi_1$ *and, for any* $n\ge 1$,

$$X_{n+1}\,|\,(X_1,\ldots,X_n)=\begin{cases}\xi_{n+1} & \text{with prob } p_{k_n+1,n}(n_1,\ldots,n_{k_n},1)\\ X^*_{n,j} & \text{with prob } p_{k_n,n}(n_1,\ldots,n_j+1,\ldots,n_{k_n})\end{cases}$$

where k_n *is the number of distinct values* $X^*_{n,1},\ldots,X^*_{n,k_n}$ *among the conditioning observations.*

The main issue with the statement above lies in the fact that it guarantees the existence of the predictive weights $p_{j,n}(n_1,\ldots,n_k)$, but it does not provide any hint on their form. As mentioned earlier, in order to evaluate the predictive distribution it is necessary to assign a p.d. to the weights $\tilde p_j$. An alternative to the normalization procedure used for NRMI and PK models, is represented by the *stick–breaking* mechanism which generates species sampling models with stick–breaking weights. Let $(V_i)_{i\ge1}$ be a sequence of independent random variables taking values in $[0,1]$ and set

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

These random weights define a proper species sampling model if and only if $\sum_{i\ge1}\mathbb{E}\left[\log(1-V_i)\right]=-\infty$. See Ishwaran and James (2001). The rationale of the construction is apparent. Suppose one has a unit length stick and breaks it into two bits of length V_1 and $1-V_1$. The first bit represents $\tilde p_1$ and in order to obtain $\tilde p_2$ it is enough to split the remaining part, of length $1-V_1$, into two parts having respective lengths $V_2(1-V_1)$ and $(1-V_2)(1-V_1)$. The former will coincide with $\tilde p_2$ and the latter will be split to generate $\tilde p_3$, and so on. The Dirichlet process with parameter measure α represents a special case, which corresponds to the Sethuraman (1994) series representation: if $\alpha(\mathbb{X})=\theta$, then the V_i's are i.i.d. with Beta$(1,\theta)$ distribution. Another nonparametric prior which admits a stick–breaking construction is the PD(σ,θ) process. If in the stick–breaking construction one takes independent V_i's such that

$$V_i\sim\text{Beta}(\theta+i\sigma,\,1-\sigma),$$

the resulting $\tilde p$ is a PD(σ,θ) process. See Pitman (1995). Moreover, Teh, Görür and Ghahramani (2007) derived a simple and interesting construction of the beta process, which is based on a variation of the stick–breaking scheme described above.

Remark 4. There has recently been a growing interest for stick–breaking priors as a tool for specifying priors within regression problems. Based on an initial idea set forth by MacEachern (1999, 2000, 2001) who introduced the so–called dependent Dirichlet process, many subsequent papers have provided variants of the stick–breaking construction so to allow either the random masses $\tilde p_j$ or the random locations X_i to depend on a set of covariates $\boldsymbol{z}\in\mathbb{R}^d$. In this respect, stick–breaking priors are particularly useful, since they allow to introduce dependence in a relatively simple way. This leads to a family of random probability measures $\{\tilde p_{\boldsymbol{z}}:\ \boldsymbol{z}\in\mathbb{R}^d\}$ where

$$\tilde p_{\boldsymbol{z}}=\sum_{j\ge1}\tilde p_{j,\boldsymbol{z}}\,\delta_{X_{j,\boldsymbol{z}}}.$$

A natural device for incorporating dependence on $\boldsymbol{z}$ into the $\tilde p_j$'s is to let the variables V_i depend on $\boldsymbol{z}\in\mathbb{R}^d$: for example one might have $V_{i,\boldsymbol{z}}\sim\text{Beta}(a_{\boldsymbol{z}},b_{\boldsymbol{z}})$. As for the dependence of the locations on $\boldsymbol{z}$, the most natural approach is to take the $X_{i,\boldsymbol{z}}$ i.i.d. with distribution $P_{0,\boldsymbol{z}}$. Anyhow, we will not enter

the technical details related to these priors: these, and other interesting proposals, are extensively described Dunson (2010). □

Turning attention back to the PD(σ,θ) process as a species sampling model, the weights $p_{j,n}$ defining the predictive distribution induced by $\tilde{p}$ are known. Indeed, if $\xi_1,\ldots,\xi_n$ are i.i.d. random variables with distribution P_0, then $X_1=\xi_1$ and, for any $i\geq 2$, one has

$$X_{n+1}\,|\,(X_1,\ldots,X_n)=\begin{cases}\xi_{n+1} & \text{with prob } (\theta+\sigma k_n)/(\theta+n)\\ X^*_{n,j} & \text{with prob } (n_{n,j}-\sigma)/(\theta+n)\end{cases}$$

with $X^*_{n,j}$ being the j–th of the k_n distinct species observed among $X_1,\ldots,X_n$ and $n_{n,j}$ is the number of times the j–th species $X^*_{n,j}$ has been observed. Besides the characterization in terms of predictive distributions, Pitman (1996) has also provided a representation of the posterior distribution of a PD(σ,θ) process $\tilde{p}$, given the data $\boldsymbol{X}$. Suppose $\mathbb{E}[\tilde{p}]=P_0$ and let $\boldsymbol{X}=(X_1,\ldots,X_n)$ be such that it contains $k\leq n$ distinct values $X^*_1,\ldots,X^*_k$, with respective frequencies $n_1,\ldots,n_k$. Then

$$\tilde{p}\,|\,\boldsymbol{X}\stackrel{d}{=}\sum_{j=1}^{k}p^*_j\,\delta_{X^*_j}+\left(1-\sum_{j=1}^{k}p^*_j\right)\tilde{p}^{(k)} \tag{39}$$

where $\tilde{p}^{(k)}$ is a PD$(\sigma,\,\theta+k\sigma)$ such that $\mathbb{E}[\tilde{p}^{(k)}]=P_0$ and $(p^*_1,\ldots,p^*_k)\sim\text{Dir}(n_1-\sigma,\ldots,n_k-\sigma,\theta+k\sigma)$. The posterior distribution of a PD(σ,θ) process can also be described in terms of a mixture with respect to a latent random variable, thus replicating the structure already encountered for NRMI. Let $\boldsymbol{X}$ be, as usual, the set of n data with $k\leq n$ distinct values $X^*_1,\ldots,X^*_k$ and let U_k be a positive random variable with density

$$q_{\sigma,\theta,k}(u)=\frac{\sigma}{\Gamma(k+\theta/\sigma)}\,u^{\theta+k\sigma-1}\,\mathrm{e}^{-u^\sigma}$$

It can be shown that the distribution of a PD(σ,θ) process, conditional on the data $\boldsymbol{X}$ and on U_k, coincides with the distribution of a normalized CRM

$$\tilde{\mu}_{U_k}+\sum_{i=1}^{k}J_i^{(U_k)}\,\delta_{X^*_i}$$

where $\tilde{\mu}_{U_k}$ is a generalized gamma process with $\rho_x^{(U_k)}(\mathrm{d}s)=\rho^{(U_k)}(\mathrm{d}s)=\frac{\sigma}{\Gamma(1-\sigma)}\,s^{-1-\sigma}\,\mathrm{e}^{-U_k\,s}\,\mathrm{d}s$. The jumps $J_i^{(U_k)}$ at the observations X^*_i are independent gamma random variables with $\mathbb{E}[J_i^{(U_k)}]=(n_i-\sigma)/U_k$. Moreover, the jumps $J_i^{(U_k)}$ and the random measure $\tilde{\mu}_{U_k}$ are, conditional on U_k, independent. This characterization shows quite nicely the relation between the posterior behaviour of the PD(σ,θ) process and of the generalized gamma NRMI, detailed in Example 8. Finally, note that the posterior representation in (39) is easily recovered by integrating out U_k.

Remark 5. Species prediction problems based on these models have been considered by Lijoi, Mena and Prünster (2007b). Specifically, they assume that data are directed by a Gibbs–type prior. Conditionally on $X_1,\ldots,X_n$, exact evaluations are derived for the following quantities: the p.d. of the number of new species that will be detected among the observations $X_{n+1},\ldots,X_{n+m}$; the probability that the observation X_{n+m+1} will show a new species. Various applications, such as, *e.g.*, gene discovery prediction in genomics, illustrate nicely how discrete nonparametric priors can be successfully used to model directly the data, if these present ties. In this context the need for predictive structures, which exhibit a more flexible clustering mechanism than the one induced by the Dirichlet process, becomes apparent.

4 Models for density estimation

Up to now we have mainly focused on nonparametric priors, which select almost surely discrete probability measures. Due to the nonparametric nature of the models, it is clear that the set of such discrete distributions is not dominated by a fixed σ–finite measure. In the present section we illustrate two different approaches for defining priors whose realizations yield, almost surely, p.d.'s admitting a density function with respect to (w.r.t.) some σ–finite measure λ on $\mathbb{X}$. The results we are going to describe are useful, for example, when one wants to model directly data generated by a continuous distribution on $\mathbb{X} = \mathbb{R}$.

4.1. Mixture models. An important and general device for defining a prior on densities has been first suggested by Lo (1984). The basic idea consists in introducing a sequence of exchangeable latent variables $(\theta_n)_{n\geq 1}$ governed by some discrete random probability measure $\tilde{p}$ on Θ, a Polish space endowed with the Borel σ–field, which is convoluted with a suitable kernel k. To be more precise, k is a jointly measurable application from $\mathbb{X} \times \Theta$ to $\mathbb{R}^+$ and, given the dominating measure λ, the application $C \mapsto \int_C k(x,\theta)\lambda(\mathrm{d}x)$ defines a probability measure on $\mathbb{X}$ for any $\theta \in \Theta$. Hence, for any θ, $k(\,\cdot\,,\theta)$ is a density function on $\mathbb{X}$ w.r.t. λ. A hierarchical mixture model can, then, be defined as follows

$$\begin{aligned} X_i \,|\, \theta_i, \tilde{p} &\overset{\text{ind}}{\sim} k(\,\cdot\,,\theta_i) \\ \theta_i \,|\, \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \\ \tilde{p} &\sim Q \end{aligned}$$

This is the same as saying that, given the random density

$$x \mapsto \tilde{f}(x) = \int_\Theta k(x,\theta)\, \tilde{p}(\mathrm{d}\theta) = \sum_{j\geq 1} k(x,\theta_j)\, \tilde{p}_j, \tag{40}$$

the observations X_i are independent and identically distributed and the common p.d. has density function $\tilde{f}$. In (40), the $\tilde{p}_j$'s are the probability masses associated to the discrete mixing distribution $\tilde{p}$. The original formulation of the model provided by Lo (1984) sets $\tilde{p}$ to coincide with a Dirichlet process: hence it takes on the name of *mixture of Dirichlet process* whose acronym MDP is commonly employed in the Bayesian literature. It is apparent that one can replace the Dirichlet process in (40) with any of the discrete random probability measures examined in Section 3. As for the choice of the kernels the most widely used is represented by the Gaussian kernel: in this case, if the nonparametric prior is assigned to both mean and variance, then $\tilde{p}$ is defined on $\Theta = \mathbb{R} \times \mathbb{R}^+$. Such an approach to density estimation yields, as a by–product, a natural framework for investigating the clustering structure within the observed data. Indeed, given the discreteness of $\tilde{p}$, there can be ties among the latent variables in the sense that $\mathbb{P}[\theta_i = \theta_j] > 0$ for any $i \neq j$. Possible coincidences among the θ_i's induce a partition structure within the observations. Suppose, for instance, that there are $k \leq n$ distinct values $\theta_1^*, \ldots, \theta_k^*$ among $\theta_1, \ldots, \theta_n$ and let $C_j := \{i : \theta_i = \theta_j^*\}$ for $j = 1, \ldots, k$. According to such a definition, any two different indices i and l belong to the same group C_j if and only if $\theta_i = \theta_l = \theta_j^*$. Hence, the C_j's describe a clustering scheme for the observations X_i: any two observations X_i and X_l belong to the same cluster if and only if $i, l \in I_j$ for some j. In particular, the number of distinct values θ_i^* among the latent θ_i's identifies the number of clusters into which

the n observations can be partitioned. Within the framework of nonparametric hierarchical mixture models, one might, then, be interested in determining an estimate of the density $\tilde{f}$ and in evaluating the posterior distribution of the number of clusters featured by the observed data. There are, however, some difficulties that do not allow for an exact numerical evaluation of the quantities of interest. Just to give an idea of the computational problems that arise, let $\mathscr{L}(\,\cdot\,|\boldsymbol{X})$ denote the posterior distribution of the k distinct latent variables θ_i^*, given the data $\boldsymbol{X} = (X_1, \ldots, X_n)$. If $\mathbb{E}[\tilde{p}] = P_0$ for some non–atomic p.d. P_0, then one has that

$$\mathscr{L}(\mathrm{d}\theta_1^* \cdots \mathrm{d}\theta_k^* \,|\, \boldsymbol{X}) \propto \Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} \prod_{i \in C_j} k(X_i, \theta_j^*)\, P_0(\mathrm{d}\theta_j^*)$$

where it is to be emphasized that the partition sets C_j depend on the specific vector $(n_1, \ldots, n_k)$ in $\Delta_{n,k}$ and $\Pi_k^{(n)}$ is the EPPF induced by $\tilde{p}$. In this case a Bayesian estimate of $\tilde{f}$ would be defined by

$$\mathbb{E}\left[\tilde{f}(x) \,\middle|\, \boldsymbol{X}\right] = \sum_{k=1}^{n} \int_{\Theta^k} \int_{\Theta} k(x, \theta) \sum_{\pi \in \mathscr{P}_{n,k}} \mathbb{E}\left[\tilde{p}(\mathrm{d}\theta) \,|\, \theta_1^*, \ldots, \theta_k^*\right] \mathscr{L}(\mathrm{d}\theta_1^* \cdots \mathrm{d}\theta_k^* \,|\, \boldsymbol{X})$$

where $\mathscr{P}_{n,k}$ is the space of all partitions $\boldsymbol{\pi}$ of $\{1, \ldots, n\}$ into $n(\boldsymbol{\pi}) = k$ sets. In the previous expression, the quantity $\mathbb{E}\left[\tilde{p}(\mathrm{d}\theta) \,|\, \theta_1^*, \ldots, \theta_k^*\right]$ is the predictive distribution which, as seen in the previous Section, can be determined in closed form for various priors. Hence, the source of problems in the above expression is the evaluation of the sum over $\mathscr{P}_{n,k}$. Analogous difficulties need to be faced when trying to determine the posterior distribution of the number of clusters K_n among the n observations $X_1, \ldots, X_n$. These technical issues can be overcome by resorting to well–established MCMC algorithms applicable to hierarchical mixture models. The main reference in this area is represented by the algorithm devised in Escobar (1988, 1994) and Escobar and West (1995) and originally developed for the MDP model. Here below we provide a description which applies to any discrete random probability measure $\tilde{p}$ for which the EPPF or, equivalently, the induced system of predictive distributions is known in explicit form, a fact first noted in Ishwaran and James (2001, 2003). In order to sample $\theta_1, \ldots, \theta_n$ from the posterior $\mathscr{L}(\,\cdot\,|\boldsymbol{X})$, one exploits the following predictive distributions

$$\mathbb{P}[\theta_i \in \mathrm{d}\theta_i | \, \boldsymbol{\theta}_{-i}, \boldsymbol{X}] = q_{i,0}^*\, P_{0,i}(\mathrm{d}\theta_i) + \sum_{j=1}^{k_{i,n-1}} q_{i,j}^*\, \delta_{\theta_{i,j}^*}(\mathrm{d}\theta_i) \tag{41}$$

where

$$P_{0,i}(\mathrm{d}\theta_i) = \frac{k(X_i, \theta_i) P_0(\mathrm{d}\theta_i)}{\int_\Theta k(X_i, \theta) P_0(\mathrm{d}\theta)},$$

$\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ and $k_{i,n-1}$ is the number of distinct values $\theta_{i,j}^*$ in the vector $\boldsymbol{\theta}_{-i}$, with n_j being the frequency of $\theta_{i,j}^*$ in $\boldsymbol{\theta}_{-i}$. As far as the weights in (41) are concerned, they are given by

$$q_{i,0}^* \propto \Pi_{k_{i,n-1}+1}^{(n)}(n_{i,1}, \ldots, n_{i,k_{i,n-1}}, 1) \int_\Theta k(X_i, \theta)\, P_0(\mathrm{d}\theta)$$
$$q_{i,j}^* \propto \Pi_{k_{i,n-1}}^{(n)}(n_{i,1}, \ldots, n_{i,j} + 1, \ldots, n_{i,k_{i,n-1}})\, k(X_i, \theta_{i,j}^*)$$

and are such that $\sum_{j=0}^{k_{i,n-1}} q_{i,j}^* = 1$. Note that these weights reduce to

$$q_{i,0}^* \propto \theta \int_\Theta k(X_i, \theta)\, P_0(\mathrm{d}\theta), \qquad q_{i,j}^* \propto n_{i,j}\, k(X_i, \theta_{i,j}^*),$$

with $n_{i,j}$ being the frequency with which $\theta^*_{i,j}$ appears in $\boldsymbol{\theta}_{-i}$, when $\tilde{p}$ is the Dirichlet process prior. The algorithm which allows to sample $\theta_1, \ldots, \theta_n$ from the posterior, given $\boldsymbol{X}$, works as follows

1) Sample i.i.d. initial values $\theta_1^{(0)}, \ldots, \theta_n^{(0)}$ from P_0

2) At each subsequent iteration $t \geq 1$ generate the vector $(\theta_1^{(t)}, \ldots, \theta_n^{(t)})$ from the corresponding distributions

$$\begin{aligned}
\theta_1^{(t)} &\sim \mathbb{P}\left[\theta_1^{(t)} \in \mathrm{d}\theta_1 \,\middle|\, \theta_2^{(t-1)}, \ldots, \theta_n^{(t-1)}, \boldsymbol{X}\right] \\
\theta_2^{(t)} &\sim \mathbb{P}\left[\theta_2^{(t)} \in \mathrm{d}\theta_2 \,\middle|\, \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_n^{(t-1)}, \boldsymbol{X}\right] \\
&\vdots \quad \vdots \\
\theta_n^{(t)} &\sim \mathbb{P}\left[\theta_n^{(t)} \in \mathrm{d}\theta_n \,\middle|\, \theta_1^{(t)}, \ldots, \theta_{n-1}^{(t)}, \boldsymbol{X}\right]
\end{aligned}$$

Each iteration from the algorithm will yield the number $k^{(t)}$ of clusters and the distinct values $\theta^*_{1,t}, \ldots, \theta^*_{k^{(t)},t}$. Using the output of N iterations, after a suitable number of burn–in period sweeps, one can evaluate a posterior estimate of $\tilde{f}$

$$\hat{f}(x) = \frac{1}{N} \sum_{t=1}^{N} \int_{\Theta} k(x, \theta)\, \mathbb{E}\left[\tilde{p}(\mathrm{d}\theta) \,\middle|\, \theta^*_{1,t}, \ldots, \theta^*_{k^{(t)},t}\right]$$

and the posterior distribution of the number of clusters

$$\mathbb{P}\left[K_n = k \,\middle|\, \boldsymbol{X}\right] \approx \frac{1}{N} \sum_{t=1}^{N} \mathbb{1}_{\{k\}}\left(k^{(t)}\right).$$

Remark 6. There are two possible problems related to the above Gibbs sampling scheme. The first one consists in a slow mixing of the chain. This drawback usually appears when the weights $q^*_{i,j}$ are much greater than $q^*_{i,0}$. A remedy is represented by the addition of a further *acceleration step*. Once the number $k^{(t)}$ of distinct latents has been sampled according to the scheme above, one proceeds to re–sampling the values of the $k^{(t)}$ distinct latent variables from their marginal distribution. In other terms, given $k^{(t)}$ and the vector $\boldsymbol{\theta}^{(t)} = (\theta^*_{1,t}, \ldots, \theta^*_{k^{(t)},t})$, one re–samples the labels of $\boldsymbol{\theta}^{(t)}$ from the distribution

$$\mathbb{P}\left[\theta^*_{1,t} \in \mathrm{d}\theta_1, \ldots, \theta^*_{k^{(t)},t} \in \mathrm{d}\theta_{k^{(t)}} \,\middle|\, \boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right] \propto \prod_{j=1}^{k^{(t)}} \prod_{i \in C_{j,t}} k(X_i, \theta_j)\, P_0(\mathrm{d}\theta_j)$$

where the $C_{j,t}$ are sets of indices denoting the membership to each of the $k^{(t)}$ clusters at iteration t. Such an additional sampling step has been suggested in MacEachern (1994) and Bush and MacEachern (1996). See also Ishwaran and James (2001). Another difficulty arises for non–conjugate models where it is not possible to sample from $P_{0,i}(\mathrm{d}\theta_i)$ and evaluate exactly the weights $q^*_{i,0}$. A variant to the sampler in this case has been proposed by MacEachern and Müller (1998), Neal (2000) and Jain and Neal (2007). Note that, even if these remedies where devised for the MDP, they work for any mixture of random probability measure. □

Remark 7. According to a terminology adopted in Papaspiliopoulos and Roberts (2008), the previous Gibbs sampling scheme can be seen as a *marginal method* in the sense that it exploits the integration

with respect to the underlying $\tilde{p}$. The alternative family of algorithms is termed *conditional methods*: those rely on the simulation of the whole model and, hence, of the latent random probability measure as well. The simulation of $\tilde{p}$ can be achieved either by resorting to the Ferguson and Klass (1972) algorithm or by applying MCMC methods tailored for stick–breaking priors. See Ishwaran and James (2001, 2003b), Papaspiliopoulos and Roberts (2008) and Walker (2007). Here we do not pursue this point and refer the interested reader to the above mentioned articles. In particular, Papaspiliopoulos and Roberts (2008) discuss a comparison between the two methods. It is important to stress that both approaches require an analytic knowledge of the posterior behaviour of the latent random probability measure: for marginal methods the key ingredient is represented by the predictive distributions, whereas for conditional methods a posterior representation for $\tilde{p}$ is essential. □

We now describe a few examples where the EPPF is known and a full Bayesian analysis for density estimation and clustering can be carried out using marginal methods.

Example 10. (Mixture of the PD(σ,θ) process). These mixtures have been examined by Ishwaran and James (2001) and, within the more general framework of species sampling models, by Ishwaran and James (2003a). For a PD(σ,θ) process $\tilde{p}$, equation (36) yields the following weights

$$q^*_{i,0} \propto (\theta + \sigma k_{i,n-1}) \int_\Theta k(X_i,\theta)\, P_0(\mathrm{d}\theta), \qquad q^*_{i,j} \propto (n_{i,j} - \sigma)\, k(X_i, \theta^*_{i,j})$$

for any $j = 1,\dots,k_{i,n-1}$. As expected, when $\sigma \to 0$ one obtains the weights corresponding to the Dirichlet process. □

Example 11. (Mixture of the generalized gamma NRMI). If the mixing $\tilde{p}$ is a normalized generalized gamma process described in Example 6, one obtains a mixture discussed in Lijoi, Mena and Prünster (2007a). The Gibbs sampler is again implemented in a straightforward way since the EPPF is known: the weights $q^*_{i,j}$, for $j = 0,\dots,k_{i,n-1}$, can be determined from the weights of the predictive, $w^{(n-1)}_{k_{i,n-1}}$ and $w^{(n-1)}_{j,k_{i,n-1}}$ as displayed in Subsection 3.2. In Lijoi, Mena and Prünster (2007a) it is observed that the parameter σ has a significant influence on the description of the clustering structure of the data. First of all, the prior distribution on the number of components of the mixture, induced by $\tilde{p}$, is quite flat if σ is not close to 0. This is in clear contrast to the highly peaked distribution corresponding to Dirichlet case. Moreover, values of σ close to 1 tend to favour the formation of a large number of clusters most of which of size (frequency) $n_j = 1$. This phenomenon gives rise to a reinforcement mechanism driven by σ: the mass allocation, in the predictive distribution, is such that clusters of small size are penalized whereas those few groups with large frequencies are reinforced in the sense that it is much more likely that they will be re–observed. The role of σ suggests a slight modification of the Gibbs sampler above and one needs to consider the full conditional of σ as well. Hence, if it is supposed that the prior for σ is some density q on $[0,1]$, one finds out that the conditional distribution of σ, given the data $\boldsymbol{X}$ and the latent variables $\boldsymbol{\theta}$, is

$$q\left(\sigma \,|\, \boldsymbol{X}, \boldsymbol{\theta}\right) = q(\sigma \,|\, \boldsymbol{\theta}) \propto q(\sigma)\, \sigma^{k-1} \left(\prod_{j=1}^{k} (1-\sigma)_{n_j - 1}\right) \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i\, \beta^{i/\sigma}\, \Gamma\left(k - \frac{i}{\sigma};\, \beta\right)$$

where, again, $n_1,\dots,n_k$ are the frequencies with which the $K_n = k$ distinct values among the θ_i's are recorded. This strategy turns out to be very useful when inferring on the number of clusters featured by the data. It is apparent that similar comments about the role of σ apply to the PD(σ,θ) process as well. □

We close this Subsection with another interesting model of mixture introduced in Petrone (1999a,b): random Bernstein polynomials.

Example 12. (Random Bernstein polynomials). A popular example of nonparametric mixture model for density estimation has been introduced by Petrone (1999a,b). The definition of the prior is inspired by the use of Bernstein polynomials for the approximation of real functions. Indeed, it is well–known that if F is a continuous function defined on $[0,1]$ then the polynomial of degree m defined by

$$B_m^F(x) = \sum_{j=0}^{m} F\left(\frac{j}{m}\right) \binom{m}{j} x^j (1-x)^{k-j} \tag{42}$$

converges, uniformly on $[0,1]$, to F as $m \to \infty$. The function B_m^F in (42) takes on the name of *Bernstein polynomial* on $[0,1]$. It is clear that, when F is a distribution function on $[0,1]$, then B_m^F is a distribution function as well. Moreover, if the p.d. corresponding to F does not have a positive mass on $\{0\}$ and $\beta(x;a,b)$ denotes the density function of a beta random variable with parameters a and b, then

$$b_m^F(x) = \sum_{j=1}^{m} \left[F(j/m) - F((j-1)/m)\right] \beta(x; j, m-j+1) \tag{43}$$

for any $x \in [0,1]$ is named a *Bernstein density*. If F has density f, it can be shown that $b_m^F \to f$ pointwise as $m \to \infty$. These preliminary remarks on approximation properties for Bernstein polynomials suggest that a prior on the space of densities on $[0,1]$ can be constructed by randomizing both the polynomial degree m and the weights of the mixture (43). In order to properly define a random Bernstein prior, let $\tilde{p}$ be, for instance, some NRMI generated by a CRM $\tilde{\mu}$ with Lévy intensity $\rho_x(\mathrm{d}s)\alpha(\mathrm{d}x)$ concentrated on $\mathbb{R}^+ \times [0,1]$ and $\alpha([0,1]) = a \in (0,\infty)$. Next, for any integer $m \geq 1$, introduce a discretization of α as follows

$$\alpha^{(m)} = \sum_{j=1}^{m} \alpha_{j,m}\, \delta_{j/m}$$

where the weights $\alpha_{j,m}$ are non–negative and such that $\sum_{j=1}^{m} \alpha_{j,m} = a$. One may note that the intensity $\nu^{(m)}(\mathrm{d}s, \mathrm{d}x) = \rho_x(\mathrm{d}s)\, \alpha^{(m)}(\mathrm{d}x)$ defines a NRMI $\tilde{p}_m$ which is still concentrated on $S_m := \{1/m, \ldots, (m-1)/m, 1\}$, *i.e.*

$$\tilde{p}_m = \sum_{j=1}^{m} \tilde{p}_{j,m}\, \delta_{j/m}$$

where $\tilde{p}_{j,m} = p(((j-1)/m,\, j/m])$, for any $j = 2, \ldots, m$, and $\tilde{p}_{1,m} = \tilde{p}([0, 1/m])$. Hence, if π is a prior on $\{1, 2, \ldots\}$, a *Bernstein random polynomial prior* is defined as the p.d. of the random density $\tilde{f}(x) = \sum_{m \geq 1} \pi(m)\, \tilde{f}_m(x)$, where

$$\tilde{f}_m(x) = \int_{[0,1]} \beta(x; my, m - my + 1)\, \tilde{p}_m(\mathrm{d}y). \tag{44}$$

is a mixture of the type (40). Conditional on m, $\tilde{f}_m$ defines a prior on the space of densities on $[0,1]$. The previous definition can be given by introducing a vector of latent variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and function $x \mapsto Z_m(x) = \sum_{j=1}^{m} j\, \mathbb{1}_{B_{j,m}}(x)$ where $B_{1,m} = [0, 1/m]$ and $B_{j,m} = ((j-1)/m,\, j/m]$ for

any $j = 2, \ldots, m$. Hence, a Bernstein random polynomial prior can be defined through the following hierarchical mixture model

$$\begin{aligned} X_j \,|\, m, \tilde{p}, Y_j &\overset{\text{ind}}{\sim} \text{Beta}(Z_m(Y_j),\, m - Z_m(Y_j) + 1) \qquad j = 1, \ldots, n \\ Y_j \,|\, m, \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \\ \tilde{p} \,|\, m &\sim Q \\ m &\sim \pi \end{aligned}$$

The original definition provided in Petrone (1999a) involves a Dirichlet process, $\tilde{p}$, with parameter measure α and the author refers to it as a *Bernstein–Dirichlet prior* with parameters (π, α). The use of the Dirichlet process is very useful, especially when implementing the MCMC strategy devised in Petrone (1999a,b) since, conditional on m, the vector of weights $(\tilde{p}_{1,m}, \ldots, \tilde{p}_{m-1,m})$ in (44) turns out to be distributed according to an $(m-1)$–variate Dirichlet distribution with parameters $(\alpha_{1,m}, \ldots, \alpha_{m,m})$. Nonetheless, the posterior distribution of $(m, \tilde{p}_m)$, given $\boldsymbol{X} = (X_1, \ldots, X_n)$, is proportional to $\pi(m)\pi(p_{1,m}, \ldots, p_{m-1,m}) \prod_{i=1}^n \tilde{f}_m(X_i)$ which is analytically intractable since it consists of a product of mixtures. For example, it is impossible to evaluate the posterior distribution $\pi(m|X_1, \ldots, X_n)$ which is of great interest since it allows to infer on the number of components in the mixture and, hence, on the number of clusters in the population. As for density estimation, the Bayesian estimate of $\tilde{f}$ with respect to a squared loss function is given by

$$\mathbb{E}[\tilde{f}(x) \,|\, X_1, \ldots, X_n] = \sum_{m \geq 1} \tilde{f}^*_m(x)\, \pi(m|X_1, \ldots, X_m)$$

with $\tilde{f}^*_m(x) = \sum_{j=1}^m \mathbb{E}[\tilde{p}_{j,m} \,|\, m, X_1, \ldots, X_n]\, \beta(x;\, j,\, m-j+1)$. This entails that the posterior estimate of $\tilde{f}$ is still a Bernstein random polynomial with updated weights. See Petrone (1999b).

Given the analytical difficulties we have just sketched, performing a full Bayesian analysis asks for the application of suitable computational schemes such as the MCMC algorithm devised in Petrone (1999b). The implementation of the algorithm is tailored to the Bernstein–Dirichlet process prior. It is assumed that the distribution function $x \mapsto F_0(x) = \alpha([0, x])/a$ is absolutely continuous with density f_0. Next, by making use of the latent variables $\boldsymbol{Y}$, a simple application of Bayes' theorem shows that

$$\pi(m|\boldsymbol{Y}, \boldsymbol{X}) \propto \pi(m) \prod_{i=1}^m \beta(X_i;\, Z_m(Y_j),\, m - Z_m(Y_j) + 1).$$

On the other hand, since $\tilde{p}$ is the Dirichlet process with parameter measure α, one has the following predictive structure for the latent variables

$$\pi(Y_j \,|\, m, \boldsymbol{Y}_{-j}, \boldsymbol{X}) \propto q(X_j, m)\, f_0(Y_j)\, \beta(X_j;\, Z_m(Y_j),\, m - Z_m(Y_j) + 1) + \sum_{i \neq j} q^*_i(X_j, m)\, \delta_{Y_i} \tag{45}$$

with $\boldsymbol{Y}_{-j}$ denoting the vector of latent variables obtained by deleting Y_j, the density $b^{F_0}_m$ defined as in (43) and

$$q(X_j, m) \propto a\, b^{F_0}_m(X_j), \qquad q^*_i(X_j, m) \propto \beta(X_j;\, Z_m(Y_i),\, m - Z_m(Y_i) + 1)$$

such that $q(X_j, m) + \sum_{i \neq j} q^*_i(X_j, m) = 1$. The predictive distribution in (45) implies that: (i) with probability $q(X_j, m)$ the value of Y_j is sampled from a density $f(y) \propto f_0(y)\, \beta(X_j;\, Z_m(y),\, m -$

$Z_m(y)+1$) and (ii) with probability $q_i^*(X_j, m)$ the value of Y_j coincides with Y_i. Hence, one can apply the following Gibbs sampling algorithm in order to sample from the posterior distribution of $(m, \boldsymbol{Y}, \tilde{p}_{1,m}, \ldots, \tilde{p}_{m,m})$. Starting from initial values $(m^{(0)}, \boldsymbol{Y}^{(0)}, p_{1,m}^{(0)}, \ldots, p_{m,m}^{(0)})$, at iteration $t \geq 1$ one samples

(1) $m^{(t)}$ from $\pi(m \,|\, \boldsymbol{Y}^{(t-1)}, \boldsymbol{X})$

(2) $Y_i^{(t)}$ from the predictive $\pi(Y_i \,|\, m^{(t)}, Y_1^{(t)}, \ldots, Y_{i-1}^{(t)}, Y_{i+1}^{(t-1)}, \ldots, Y_n^{(t-1)}, \boldsymbol{X})$ described in (45)

(3) $(p_{1,m}^{(t)}, \ldots, p_{m,m}^{(t)})$ from an $(m-1)$–variate Dirichlet distribution with parameters $(\alpha_{1,m^{(t)}} + n_1, \ldots, \alpha_{m^{(t)},m^{(t)}} + n_{m^{(t)}})$, where n_j is the number of latent variables in $(Y_1^{(t)}, \ldots, Y_n^{(t)})$ in $B_{j,m^{(t)}}$.

For further details, see Petrone (1999a). □

4.2. Pólya trees. Pólya trees are another example of priors which, under suitable conditions, are concentrated on absolutely continuous p.d.'s with respect to the Lebesgue measure on $\mathbb{R}$. A first definition of Pólya trees can be found in Ferguson (1974) and a systematic treatment is provided by Lavine (1992, 1994) and Mauldin, Sudderth and Williams (1992). A useful preliminary concept is that of *tailfree prior* introduced by Freedman (1963). Let $\Gamma = \{\Gamma_k : k \geq 1\}$ be a nested tree of measurable partitions of $\mathbb{X}$. This means that Γ_{k+1} is a refinement of Γ_k, *i.e.* each set in Γ_{k+1} is the union of sets in Γ_k, and that $\cup_{k\geq 1}\Gamma_k$ generates $\mathscr{X}$, with $\mathscr{X}$ denoting the Borel σ–algebra of $\mathbb{X}$. One can, then, give the following

Definition 8. A random probability measure $\tilde{p}$ on $\mathbb{X} \subset \mathbb{R}$ is *tailfree* with respect to Γ if there exist non–negative random variables $\{V_{k,B} : k \geq 1,\ B \in \Gamma_k\}$ such that

(i) the families $\{V_{1,B} : B \in \Gamma_1\}$, $\{V_{2,B} : B \in \Gamma_2\}$, ..., are independent

(ii) if $B_k \subset B_{k-1} \subset \cdots \subset B_1$, with $B_j \in \Gamma_j$, then $\tilde{p}(B_k) = \prod_{j=1}^k V_{j,B_j}$

For tailfree processes a structural conjugacy property holds true: if $\tilde{p}$ is tailfree with respect to Γ, then $\tilde{p}$ given the data is still tailfree with respect to Γ.

Pólya trees can be recovered as special case of tailfree processes with the $V_{k,B}$ variables having a beta distribution. To illustrate the connection, consider the family Γ of partitions described as follows

$$\Gamma_1 = \{B_0, B_1\}, \quad \Gamma_2 = \{B_{00}, B_{01}, B_{10}, B_{11}\}, \quad \Gamma_3 = \{B_{000}, B_{001}, B_{010}, \ldots, B_{111}\}$$

and so on. In the above definition of the Γ_i's we set $B_0 = B_{00} \cup B_{01}$, $B_1 = B_{10} \cup B_{11}$ and, given sets $B_{\boldsymbol{\varepsilon}0}$ and $B_{\boldsymbol{\varepsilon}1}$ in Γ_{k+1}, one has

$$B_{\boldsymbol{\varepsilon}0} \cup B_{\boldsymbol{\varepsilon}1} = B_{\boldsymbol{\varepsilon}}$$

for any $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k) \in E^k = \{0,1\}^k$. With this notation, the k–th partition can be described as $\Gamma_k = \{B_{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in E^k\}$. Finally, let $E^* = \cup_{k\geq 1} E^k$ be the set of all sequences of zeros and ones and $\mathscr{A} = \{\alpha_{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in E^*\}$ a set of non–negative real numbers.

Definition 9. A random probability measure $\tilde{p}$ is a *Pólya tree* process with respect to $\Gamma = \{\Gamma_k : k \geq 1\}$ and $\mathscr{A}$, in symbols $\tilde{p} \sim \mathrm{PT}(\mathscr{A}, \Gamma)$, if

(i) $\{\tilde{p}(B_{\boldsymbol{\varepsilon}0}|B_{\boldsymbol{\varepsilon}}) : \boldsymbol{\varepsilon} \in E^*\}$ is a collection of independent random variables

(ii) $\tilde{p}(B_{\boldsymbol{\varepsilon}0}|B_{\boldsymbol{\varepsilon}}) \sim \text{Beta}(\alpha_{\boldsymbol{\varepsilon}0}, \alpha_{\boldsymbol{\varepsilon}1})$

The existence of a Pólya tree with respect to the parameters $\mathscr{A}$ is guaranteed by the validity of the following conditions expressed in terms of infinite products

$$\frac{\alpha_{\boldsymbol{\varepsilon}0}}{\alpha_{\boldsymbol{\varepsilon}0}+\alpha_{\boldsymbol{\varepsilon}1}}\,\frac{\alpha_{\boldsymbol{\varepsilon}00}}{\alpha_{\boldsymbol{\varepsilon}00}+\alpha_{\boldsymbol{\varepsilon}01}}\,\cdots = 0, \qquad \frac{\alpha_{1}}{\alpha_{0}+\alpha_{1}}\,\frac{\alpha_{11}}{\alpha_{10}+\alpha_{11}}\,\cdots = 0$$

These ensure that the Pólya random probability measure is countably additive, almost surely. For a proof of this fact see, *e.g.* , Ghosh and Ramamoorthi (2003).

One of the most relevant properties of a Pólya tree prior $\text{PT}(\mathscr{A}, \Gamma)$ is that, under a suitable condition on the parameters in $\mathscr{A}$, the realizations of $\tilde{p}$ are, almost surely, p.d.'s that are absolutely continuous. In order to illustrate such a condition we confine ourselves to the case where $\mathbb{X} = [0, 1]$, the extension to the case $\mathbb{X} = \mathbb{R}$ being straightforward. Suppose that Γ is a sequence of dyadic partitions of $[0, 1]$, *i.e.* with $\boldsymbol{\varepsilon} \in E^k$ one has $B_{\boldsymbol{\varepsilon}} = (\sum_{j=1}^{k} \varepsilon_j\, 2^{-j},\ \sum_{j=1}^{k} \varepsilon_j\, 2^{-j} + 2^{-k}]$. As noted in Ferguson (1974), using a result in Kraft (1964), one can show that if $\tilde{p} \sim \text{PT}(\mathscr{A}, \Gamma)$ and the $\alpha_{\varepsilon_1 \cdots \varepsilon_k}$'s, seen as function of the level on the partition tree, increase at a rate of at k^2 or faster, then the p.d. of $\tilde{p}$ is concentrated on the set of probability measures that are absolutely continuous with respect to the Lebesgue measure.

The beta distribution in the definition above allows for a straightforward characterization of the marginal distribution of the observations. Indeed, if $(X_n)_{n\geq 1}$ is an exchangeable sequence of observations governed by a $\text{PT}(\mathscr{A}, \Gamma)$ according to model (1), then any $B_{\boldsymbol{\varepsilon}} \in \Gamma_k$ is such that $B_{\boldsymbol{\varepsilon}} = \cap_{i=1}^{k} B_{\varepsilon_1 \cdots \varepsilon_i}$ and

$$\begin{aligned} \mathbb{P}[X_1 \in B_{\boldsymbol{\varepsilon}}] = \mathbb{E}\left[\tilde{p}(B_{\boldsymbol{\varepsilon}})\right] &= \mathbb{E}\left[\prod_{i=1}^{k} \tilde{p}(B_{\varepsilon_1 \cdots \varepsilon_i} | B_{\varepsilon_1 \cdots \varepsilon_{i-1}})\right] \\ &= \prod_{i=1}^{k} \mathbb{E}\left[\tilde{p}(B_{\varepsilon_1 \cdots \varepsilon_i} | B_{\varepsilon_1 \cdots \varepsilon_{i-1}})\right] = \frac{\alpha_{\varepsilon_1}}{\alpha_0 + \alpha_1} \prod_{i=2}^{k} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_i}}{\alpha_{\varepsilon_1 \cdots \varepsilon_{i-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{i-1} 1}} \end{aligned} \tag{46}$$

where we have set, by convention, $\tilde{p}(B_{\varepsilon_1}|B_{\varepsilon_1 \varepsilon_0}) = \tilde{p}(B_{\varepsilon_1})$ and the last two equalities follow, respectively, from the independence among the $\tilde{p}(B_{\varepsilon_1 \cdots \varepsilon_i}|B_{\varepsilon_1 \cdots \varepsilon_{i-1}})$ and the fact that each of these random variables has beta distribution. Similar arguments lead one to determine the posterior distribution of a $\text{PT}(\mathscr{A}, \Gamma)$ prior. See Ferguson (1974), Lavine (1992) and Mauldin, Sudderth and Williams (1992).

Theorem 12. *Let $\tilde{p} \sim \text{PT}(\mathscr{A}, \Gamma)$ and $(X_n)_{n\geq 1}$ is an exchangeable sequence of random elements taking values in $\mathbb{X}$ and governed by the p.d. of $\tilde{p}$. Then*

$$\tilde{p}\,|\,\boldsymbol{X} \;\sim\; \text{PT}(\mathscr{A}_n^*, \Gamma)$$

where $\mathscr{A}_n^* = \{\alpha_{n,\boldsymbol{\varepsilon}}^* : \boldsymbol{\varepsilon} \in E^*\}$ is the updated set of parameters defined by $\alpha_{n,\boldsymbol{\varepsilon}}^* = \alpha_{\boldsymbol{\varepsilon}} + \sum_{i=1}^{n} \mathbb{1}_{B_{\boldsymbol{\varepsilon}}}(X_i)$ and $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Hence, Pólya trees feature parametric conjugacy. The posterior distribution can be employed in order to deduce the system of predictive distributions associated to $\tilde{p}$. Since $\mathbb{P}\left[X_{n+1} \in B_{\boldsymbol{\varepsilon}}\,|\,\boldsymbol{X}\right] = \mathbb{E}\left[\tilde{p}(B_{\boldsymbol{\varepsilon}})\,|\,\boldsymbol{X}\right]$ one can combine the previous theorem with the marginal distribution in (46) to obtain a characterization of the predictive distribution of X_{n+1} given the data. Indeed, since $\tilde{p}|\boldsymbol{X} \stackrel{d}{=} \tilde{p}_n \sim \text{PT}(\mathscr{A}_n^*, \Gamma)$, then for any $\boldsymbol{\varepsilon} \in E^k$

$$\mathbb{P}\left[X_{n+1} \in B_{\boldsymbol{\varepsilon}} \,|\, \boldsymbol{X}\right] = \prod_{i=1}^{k} \mathbb{E}\left[\tilde{p}_n(B_{\varepsilon_1 \cdots \varepsilon_i} | B_{\varepsilon_1 \cdots \varepsilon_{i-1}})\right]$$

$$= \frac{\alpha_{\varepsilon_1} + n_{\varepsilon_1}}{\alpha_0 + \alpha_1 + n} \prod_{j=2}^{k} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_j} + n_{\varepsilon_1 \cdots \varepsilon_j}}{\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 1} + n_{\varepsilon_1 \cdots \varepsilon_{j-1}}}$$

where $n_{\varepsilon_1 \cdots \varepsilon_j} = \sum_{i=1}^{n} \mathbb{1}_{B_{\varepsilon_1 \cdots \varepsilon_j}}(X_i)$ is the number of observations in $B_{\varepsilon_1 \cdots \varepsilon_j}$ for $j \in \{1, \ldots, k\}$. The displayed expression suggests that, even if the predictive density exists, it can be discontinuous and the discontinuities will depend on the specific sequence of partitions Γ.

The partition tree Γ and the parameters $\mathscr{A}$ can be used to incorporate prior opinions on the unknown distribution function. Lavine (1992) provides some hints in this direction. Suppose, *e.g.* , that the prior guess at the shape of $\tilde{p}$ is P_0. Hence, one would like to fix the Pólya tree such that $\mathbb{E}[\tilde{p}] = P_0$. If $F_0(x) = P_0((-\infty, x])$, for any x in $\mathbb{R}$, and $F_0^{-1}(y) = \inf\{x : F(x) \geq y\}$ is the quantile function of F_0, for any $y \in [0, 1]$, then the sequence Γ of partitions can be fixed in such a way that

$$B_{\boldsymbol{\varepsilon}} = \left(F_0^{-1}\left(\sum_{i=1}^{k} \varepsilon_i \, 2^{-i}\right),\ F_0^{-1}\left(\sum_{i=1}^{k} \varepsilon_i \, 2^{-i} + 2^{-k}\right)\right]$$

for any $k \geq 1$ and $\boldsymbol{\varepsilon} \in E^k$. Then, by setting $\alpha_{\boldsymbol{\varepsilon}0} = \alpha_{\boldsymbol{\varepsilon}1}$ for any $\boldsymbol{\varepsilon} \in E^*$, one has

$$\mathbb{E}\left[\tilde{p}(B_{\boldsymbol{\varepsilon}})\right] = \prod_{i=1}^{k} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_i}}{\alpha_{\varepsilon_1 \cdots \varepsilon_{i-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{i-1} 1}} = 2^{-k} = P_0(B_{\boldsymbol{\varepsilon}})$$

for any $k \geq 1$ and $\boldsymbol{\varepsilon} \in E^k$. Since $\cup_{k \geq 1} \Gamma_k$ generates $\mathscr{B}(\mathbb{R})$, this implies $\mathbb{E}[\tilde{p}] = P_0$. Having centered the prior on the desired P_0, one still has to face the issue of specifying the actual values of the α_ε's. These control the strength of the prior belief in P_0, in the sense that large α_ε's tend to concentrate the Pólya tree around the prior guess P_0. Moreover, and more importantly, the choice of $\mathscr{A}$ determines the almost sure realizations of $\tilde{p}$. As we have already noted, if $\mathbb{X} = [0, 1]$ and Γ is a sequence of nested partitions of $\mathbb{X}$ into dyadic intervals, then $\alpha_{\boldsymbol{\varepsilon}} = k^2$, for any $\boldsymbol{\varepsilon} \in E^k$ and $k \geq 1$, implies that $\tilde{p}$ is (almost surely) absolutely continuous. If, on the other hand $\alpha_{\boldsymbol{\varepsilon}} = 2^{-k}$, for any $\boldsymbol{\varepsilon} \in E^k$ and $k \geq 1$, then $\tilde{p}$ is a Dirichlet process, which selects discrete probabilities with probability 1. Finally, if $\alpha_{\boldsymbol{\varepsilon}} = 1$ for any $\boldsymbol{\varepsilon} \in E^*$, then $\tilde{p}$ is continuous singular with probability 1. See Ferguson (1974) and Mauldin, Sudderth and Williams (1992) for some comments on this issue and further results.

Also alternative strategies are available for selecting the tree of partitions Γ. For example, suppose the data consist of censored observations, with censoring times occurring at $c_1 < c_2 < \cdots < c_n$. Within the partitions $\Gamma_1, \ldots, \Gamma_n$, choose $B_1 = (c_1, \infty)$, $B_{11} = (c_2, \infty)$, and so on. If $\tilde{p} \sim \mathrm{PT}(\mathscr{A}, \Gamma)$, then the posterior of $\tilde{p}$, given the n censored data, is $\mathrm{PT}(\mathscr{A}^*, \Gamma)$. The parameters in $\mathscr{A}^*$ are identical to those in $\mathscr{A}$, with the exception of $\alpha_1^* = \alpha_1 + n$, $\alpha_{11}^* = \alpha_{11} + n - 1$, ..., $\alpha_{11\cdots1}^* = \alpha_{11\cdots1} + 1$. For an application of Pólya trees to survival analysis see Muliere and Walker (1997).

Pólya trees represent an important extension of the Dirichlet process since they stand as priors for absolutely continuous distributions on $\mathbb{R}$: nonetheless, they feature a serious drawback, since the inferences deduced from a Pólya tree prior heavily depend on the specific sequence of partitions Γ. In order to overcome the issue, Lavine (1992) suggests the use of mixtures of Pólya trees. This amounts to assuming the existence of random variables θ and ξ such that

$$\tilde{p} \,|\, (\theta, \xi) \quad \sim \quad \mathrm{PT}(\mathscr{A}^\theta, \Gamma^\xi)$$

$$(\theta, \xi) \quad \sim \quad \pi$$

If the prior π on the mixing parameters satisfies some suitable conditions, then the dependence on the partitions is smoothed out and the predictive densities can be continuous. A similar device is adopted in Paddock, Ruggeri, Lavine and West (2003) where the authors introduce a sequence of independent random variables which determine the end points partition elements in Γ_k, for any $k \geq 1$. Mixtures of Pólya trees are also used in Hanson and Johnson (2002) to model the regression error and the authors investigate applications to semiparametric accelerated failure time models.

5 Random means

The investigation of general classes of priors as developed in the previous sections is of great importance when it comes to study some quantities of statistical interest. Among these, here we devote some attention to random means, namely to linear functionals of random probability measures $\tilde{p}(f) = \int f \, \mathrm{d}\tilde{p}$, with f being some measurable function defined on $\mathbb{X}$. For instance, if the data are lifetimes, $\int x \, \tilde{p}(\mathrm{d}x)$ represents the random expected lifetime. The reason for focusing on this topic lies not only in the statistical issues that can be addressed in terms of means, but also because many of the results obtained for means of nonparametric priors do have important connections with seemingly unrelated research topics such as, *e.g.* , excursions of Bessel processes, the moment problem, special functions and combinatorics.

The first pioneering fundamental contributions to the study of means are due to D.M. Cifarelli and E. Regazzini. In their papers (Cifarelli and Regazzini, 1979a, 1979b and 1990) they provide useful insight into the problem and obtain closed form expressions for the p.d. of $\tilde{p}(f)$ when $\tilde{p}$ is a Dirichlet process. They first determine the remarkable identity for means of the Dirichlet process

$$\mathbb{E}\left[\frac{1}{\{1 + \mathrm{i}t\tilde{p}(f)\}^{\theta}}\right] = \exp\left\{-\int \log(1 + \mathrm{i}tf) \, \mathrm{d}\alpha\right\} \qquad \forall t \in \mathbb{R} \tag{47}$$

where f is any measurable function on $\mathbb{X}$ such that $\int \log(1 + |f|) \, \mathrm{d}\alpha < \infty$ and $\theta = \alpha(\mathbb{X}) \in (0, \infty)$. The left–hand side of (47) is the Stieltjes transform of order θ of the p.d., say $M_{\alpha,f}$, of the Dirichlet mean $\tilde{p}(f)$, while the right–hand side is the Laplace transform of $\int f \, \mathrm{d}\tilde{\gamma}$ where $\tilde{\gamma}$ is a gamma process with parameter measure α. Equation (47) has been termed *Markov–Krein identity* because of its connections to the Markov moment problem, whereas it is named the *Cifarelli–Regazzini identity* in James (2006b). By resorting to (47), Cifarelli and Regazzini (1990) apply an inversion formula for Stieltjes transforms and obtain an expression for $M_{\alpha,f}$. For example, if $\theta = 1$, the density function corresponding to $M_{\alpha,f}$ coincides with

$$m_{\alpha,f}(x) = \frac{1}{\pi} \, \sin\left(\pi F_{\alpha^*}(x)\right) \, \exp\left\{-\mathrm{PV} \int_{\mathbb{R}} \log|y - x| \, \alpha^*(\mathrm{d}y)\right\}$$

where $\alpha^*(B) = \alpha(\{x \in \mathbb{R} : f(x) \in B\})$ is the image measure of α through f, F_{α^*} is the corresponding distribution function and $\mathrm{PV}\int$ means that the integral is a principal–value integral. In Diaconis and Kemperman (1996) one can find an interesting discussion with some applications of the formulae of Cifarelli and Regazzini (1990). Alternative expressions for $M_{\alpha,f}$ can be found in Regazzini, Guglielmi and Di Nunno (2002) where the authors rely on an inversion formula for characteristic functions due to Gurland (1948).

Since, in general, the exact analytic form of $M_{\alpha,f}$ is involved and difficult to evaluate, it is desirable to devise some convenient method to sample from $M_{\alpha,f}$ or to approximate it numerically. For example, Muliere and Tardella (1998) make use of the stick–breaking representation of the Dirichlet process and suggest an approximation based on a random stopping rule. In Regazzini, Guglielmi and Di Nunno (2002) one can find numerical approximation of $M_{\alpha,f}$.

In Lijoi and Regazzini (2004) it is noted that when the baseline measure α is concentrated on a finite number of points, then the left–hand side of (47) coincides with the fourth Lauricella hypergeometric function. See Exton (1977). Such a connection has been exploited in order to provide an extension of (47) where the order of the Stieltjes transform does not need to coincide with the total mass of the baseline measure α. Other interesting characterizations of $M_{\alpha,f}$ can also be found in Hjort and Ongaro (2005). It is worth noting that Romik (2004, 2005) has recently pointed out how the p.d. $M_{\alpha,f}$ of a Dirichlet random mean coincides with the limiting distribution of a particular hook walk: it precisely represents the p.d. of the point where the hook walk intersects, on the plane, the graph of a continual Young diagram. Recall that a continual Young diagrams is a positive increasing function g on some interval $[a, b]$ and it can be seen as the continuous analog of the Young diagram which is a graphic representation of a partition of an integer n. Romik (2004, 2005) has considered the problem of determining a formula for the baseline measure α (with support a bounded interval $[\xi_1, \xi_2]$) corresponding to a specified distribution $M_{\alpha,f}$ for the Dirichlet random mean. The solution he obtains is described by

$$F_\alpha(x) = \frac{1}{\pi} \operatorname{arccot}\left(\frac{1}{\pi\, m_{\alpha,f}(x)} \operatorname{PV} \int_{[\xi_1,\xi_2]} \frac{m_{\alpha,f}(y)}{y-x}\, \mathrm{d}y \right).$$

See also Cifarelli and Regazzini (1993) for an alternative representation of F_α and Hill and Monticino (1998) for an allied contribution.

There have also been recent contributions to the analysis of linear functionals of more general classes of priors of the type we have been presenting in this paper. In Regazzini, Lijoi and Prünster (2003) the authors resort to Gurland's inversion formula for characteristic functions and provide an expression for the distribution function of linear functionals $\tilde{p}(f)$ of NRMIs. This approach can be naturally extended to cover means of the mixture of a Dirichlet process (Nieto–Barajas, Prünster and Walker, 2004). In Epifani, Lijoi and Prünster (2003) one can find an investigation of means of NTR priors which are connected to exponential functionals of Lévy processes: these are of great interest in the mathematical finance literature. The determination of the p.d. of a linear functional of a two–parameter Poisson–Dirichlet process has been the focus of James, Lijoi and Prünster (2008). They rely on a representation of the Stieltjes transform of $\tilde{p}(f)$ as provided in Kerov (1998) and invert it. The formulae they obtain are of relevance also for the study of excursions of Bessel processes, which nicely highlights the connection of Bayes Nonparametrics with other areas in strong development. Indeed, let $Y = \{Y_t, t \geq 0\}$ denote a real–valued process, such that: (i) the zero set $\mathcal{Z}$ of Y is the range of a σ–stable process and (ii) given $|Y|$, the signs of excursions of Y away from zero are chosen independently of each other to be positive with probability p and negative with probability $\bar{p} = 1 - p$. Examples of this kind of process are: Brownian motion ($\sigma = p = 1/2$); skew Brownian motion ($\sigma = 1/2$ and $0 < p < 1$); symmetrized Bessel process of dimension $2 - 2\sigma$; skew Bessel process of dimension $2 - 2\sigma$.

Then for any random time T which is a measurable function of $|Y|$,

$$A_T = \int_0^T \mathbb{1}_{(0,+\infty)}(Y_s)\,\mathrm{d}s \tag{48}$$

denotes the time spent positive by Y up to time T and A_T/T coincides in distribution with the distribution of $\tilde{p}(f)$ where $\tilde{p}$ is a $\mathrm{PD}(\sigma,\sigma)$ process and $f = \mathbb{1}_C$, the set C being such that $\alpha(C)/\theta = p$. See Pitman and Yor (1997b) for a detailed analysis. A recent review on means of random probability measures is provided in Lijoi and Prünster (2009).

6 Concluding remarks

In the present paper we have provided an overview of the various classes of priors which generalize the Dirichlet process. As we have tried to highlight, most of them are suitable transformations of CRMs and they all share a common *a posteriori* structure. As far as the tools for deriving posterior representations are concerned, there are essentially two general techniques and both take the Laplace functional in (3) as starting point. The first one, set forth in James (2002) and developed and refined in subsequent papers, is termed Poisson partition calculus: the key idea consists in facing the problem at the level of the Poisson process underlying the CRM, according to (7), and then to use Fubini–type arguments. The second approach, developed by the two authors of the present review and first outlined in Prünster (2002), tackles directly the problem at the CRM level, interprets observations as derivatives of the Laplace functional and then obtains the posterior representations as Radon–Nikodým derivatives.

A last remark concerns asymptotics, a research area under strong development which has been accounted for, e.g., in Ghosal (2010). Among the asymptotic properties, consistency plays a predominant role. Despite the general validity of proper Bayesian Doob–style consistency, the "what if" or frequentist approach to consistency set forth by Diaconis and Freedman (1986) has recently gained great attention. The evaluation of a Bayesian procedure according to such a frequentist criterion is appropriate when one believes that data are i.i.d. from some "true" distribution P_0 and, nonetheless, assumes exchangeability as a tool which leads to a sensible rule for making predictions and for inductive reasoning. One is, then, interested to ascertain whether the posterior distribution accumulates in suitable neighbourhoods of P_0 as the sample size increases. A few examples of inconsistency provide a warning and suggest a careful treatment of this issue. Many sufficient conditions ensuring frequentist consistency are now available and results on rates of convergence have been derived as well. If one adheres to such a frequentist point of view, then one should choose, among priors for which consistency has been proved, the one featuring the fastest rate of convergence. When dealing with the discrete nonparametric priors examined in Sections 2 and 3 these considerations are clearly of interest: in fact, most of them, with the exceptions of the Dirichlet and the beta processes, are inconsistent if used to model directly continuous data. However, even an orthodox Bayesian who does not believe in the existence of a "true" P_0 and, hence, specifies priors regardless of frequentist asymptotic properties, would hardly use a discrete nonparametric prior on continuous data: this would mean assuming a model, which generates ties among observations with probability tending to 1 as the sample size diverges, for data which do not contain ties with probability 1. On the other hand, all the discrete priors we have been describing are consistent when exploited in situations they are structurally designed for.

Specifically, they are consistent when used for modelling data arising from discrete distributions and, moreover, they are also consistent, under mild conditions, when exploited in a hierarchical mixture setup for continuous data. Thus, we have agreement of the two viewpoints on the models to use. Finally, note that rates of convergence seem not to discriminate between different discrete priors in a mixture, since they are derived assuming i.i.d. data. In such cases we have to reverse the starting question and ask "what if the data are not i.i.d. but, indeed, exchangeable"? Then, the assessment of a prior should naturally be guided by considerations on the flexibility of the posterior and on the richness of the predictive structure, which also allow for a parsimonious model specification.

Ackowledgements. Both authors wish to express their gratitude to Eugenio Regazzini who has introduced them to the world of Bayesian statistics and has transmitted enthusiasm and skills of great help for the development of their own research. This research was partially supported by MIUR–Italy.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701–726.

Aldous D.J. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour XIII*, Lecture Notes in Math., Vol. **1117**. Springer, Berlin, 1–198.

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.

Arjas, E. and Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. *Statist. Sinica* **4**, 505–524.

Blackwell, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1**, 356-358.

Blum, J. and Susarla, V. (1977). On the posterior distribution of a Dirichlet process given randomly right censored observations. *Stochastic Processes Appl.* **5**, 207–211.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.* **31**, 929–953.

Bush, C.A. and MacEachern, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

Cifarelli, D.M. and Regazzini, E. (1979a). A general approach to Bayesian analysis of nonparametric problems. The associative mean values within the framework of the Dirichlet process. I. (Italian) *Riv. Mat. Sci. Econom. Social.* **2**, 39–52.

Cifarelli, D.M. and Regazzini, E. (1979b). A general approach to Bayesian analysis of nonparametric problems. The associative mean values within the framework of the Dirichlet process. II. (Italian) *Riv. Mat. Sci. Econom. Social.* **2**, 95–111.

Cifarelli, D.M. and Regazzini, E. (1990). Distribution functions of means of a Dirichlet process. *Ann. Statist.*, **18**, 429-442 (Correction in *Ann. Statist.* (1994) **22**, 1633-1634).

Cifarelli, D.M. and Regazzini, E. (1993). Some remarks on the distribution functions of means of a Dirichlet process. *Technical Report* **93.4**, IMATI–CNR, Milano.

Connor, R.J and Mosimann, J.E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Amer. Statist. Assoc.* **64**, 194–206

Daley, D.J. and Vere–Jones, D. (1988). *An introduction to the theory of point processes.* Springer, New York.

Damien, P., Laud, P. and Smith, A.F.M. (1995). Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *J. Roy. Statist. Soc. B* **57**, 547–563.

Dey, J., Erickson, R.V. and Ramamoorthi, R.V. (2003). Some aspects of neutral to the right priors. *Internat. Statist. Rev.* **71**, 383–401.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–26.

Diaconis, P. and Kemperman, J. (1996). Some new tools for Dirichlet priors. In *Bayesian statistics 5*, 97–106. Oxford Univ. Press, New York.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183–201.

Dunson, D.B. (2010). Nonparametric Bayes applications to biostatistics. To appear in *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C., Müller, P., Walker S.G. Eds.), Cambridge University Press, Cambridge.

Dykstra, R. L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9**, 356–367.

Epifani, I., Lijoi, A. and Prünster, I. (2003). Exponential functionals and means of neutral–to–the–right priors. *Biometrika* **90**, 791–808.

Escobar, M.D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. *PhD dissertation*, unpublished. Yale University, Dept. of Statistics.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268–277.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Stat. Assoc.* **90**, 577–588.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3** 87-112.

Exton, H. (1976). *Multiple hypergeometric functions and applications.* Ellis Horwood, Chichester.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615–629.

Ferguson, T.S. and Klass, M.J. (1972). A representation of independent increments processes without Gaussian components. *Ann. Math. Statist.* **43**, 1634–1643.

Ferguson, T.S. and Phadia, E.G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7**, 163-186.

de Finetti, B. (1937). La prévision : ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68.

Freedman, D.A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386–1403.

Ghosal, D.B. (2010). Dirichlet process, related priors and posterior asymptotics. To appear in *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C., Müller, P., Walker S.G. Eds.), Cambridge University Press, Cambridge.

Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian nonparametrics.* Springer, New York.

Gill, R.D. and Johansen, S. (1990). A survey of product integration with a view towards survival analysis. *Ann. Statist.* **18**, 1501–1555.

Gnedin, A. and Pitman, J. (2005a). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **325**, 83–102.

Gnedin, A. and Pitman, J. (2005b). Regenerative composition structures. *Ann. Probab.* **33**, 445–479.

Griffiths, T.L. and Ghahramani, Z. (2006) Infinite Latent Feature Models and the Indian Buffet Process. In *Advances in Neural Information Processing Systems* **18** (Eds. Weiss, Y., Schölkopf, B. and Platt, J.), MIT Press, Cambridge, USA, 475–482.

Gurland, J. (1948). Inversion formulae for the distributions of ratios. *Ann. Math. Statist.*, **19**, 228-237.

Hanson, T. and Johnson, W.O. (2002). Modeling regression error with a mixture of Polya trees. *J. Amer. Statist. Assoc.* **97**, 1020–1033.

Hartigan, J.A. (1990). Partition models. *Comm. Statist. Theory Methods* **19**, 2745–2756.

Hill, T. and Monticino, M. (1998). Constructions of random distributions via sequential barycenters. *Ann. Statist.* **26**, 1242–1253.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259–1294.

Hjort, N.L. (2003). Topics in non-parametric Bayesian statistics. In: *Highly structured stochastic systems* (P. Green, N.L. Hjort and S.Richardson, Eds.), 455–487. Oxford Univ. Press, Oxford.

Hjort, N.L. and Ongaro, A. (2005). Exact inference for random Dirichlet means. *Stat. Inference Stoch. Process.* **8**, 227–254.

Ho, M.-W. (2006). A Bayes method for a monotone hazard rate via S-paths. *Ann. Statist.* **34**, 820–836.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Stat. Assoc.* **96**, 161–173.

Ishwaran, H. and James, L.F. (2003a). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13**, 1211–1235.

Ishwaran, H. and James, L.F. (2003b). Some further developments for stick-breaking priors: finite and infinite clustering and classification. *Sankhyā* **65**, 577–592.

Ishwaran, H. and James, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: Proportional hazards, marked point processes, and panel count data. *J. Amer. Stat. Assoc.* **99**, 175–190.

Jain, S. and Neal, R.M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis* **2**, 445–472.

James, L.F. (2002). Poisson Process Partition Calculus with applications to exchangeable models and Bayesian Nonparametrics. Unpublished Manuscript. *Mathematics ArXiv*, math.PR/0205093.

James, L.F. (2003). A simple proof of the almost sure discreteness of a class of random measures. *Statist. Probab. Lett.* **65**, 363–368.

James, L.F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Ann. Statist.* **33**, 1771–1799.

James, L.F. (2006a). Poisson calculus for spatial neutral to the right processes. *Ann. Statist.* **34**, 416–440.

James, L.F. (2006b). Functionals of Dirichlet processes, the Cifarelli-Regazzini identity and beta-gamma processes. *Ann. Statist.* **33**, 647–660.

James, L.F., Lijoi, A. and Prünster, I. (2005). Bayesian inference via classes of normalized random measures. Unpublished Manuscript. *Mathematics ArXiv*, `math/0503394`.

James, L.F., Lijoi, A. and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.* **33**, 105–120.

James, L.F., Lijoi, A. and Prünster, I. (2008). Distributions of linear functionals of two parameter Poisson–Dirichlet random measures. *Ann. Appl. Probab.* **18**, 521-551.

James, L.F., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.* **36**, 76-97.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles.* Springer, New York.

Kerov, S. (1998). Interlacing measures. In *Kirillov's seminar on representation theory.* Amer. Math. Soc. Transl. Ser. 2, Vol. **181**. Amer. Math. Soc., Providence, 35–83.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Ann. Statist.* **27**, 562–588.

Kingman, J.F.C. (1967). Completely random measures. *Pacific J. Math.* **21**, 59–78.

Kingman, J.F.C. (1975). Random discrete distributions (with discussion). *J. Roy. Statist. Soc. Ser. B* **37**, 1–22.

Kingman, J.F.C. (1978). The representation of partition structures. *J. London Math. Soc.* **18**, 374–380.

Kingman, J.F.C. (1982). The coalescent. *Stochast. Processes Appl.* **13**, 235–248.

Kingman, J.F.C. (1993). *Poisson processes.* Oxford University Press, Oxford.

Korwar, R.M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1**, 705–711.

Kraft, C.H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probability* **1**, 385–388

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222–1235.

Lavine, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **22**, 1161–1176.

Lijoi, A. and Regazzini, E. (2004). Means of a Dirichlet process and multiple hypergeometric functions. *Ann. Probab.* **32**, 1469–1495.

Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *J. Amer. Statist. Assoc.* **100**, 1278–1291.

Lijoi, A., Mena, R.H. and Prünster, I. (2007a). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 715–740.

Lijoi, A., Mena, R.H. and Prünster, I. (2007b). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786.

Lijoi, A. and Prünster, I. (2009). Distributional properties of means of random probability measures. *Statistics Surveys*, **3**, 47–95.

Lijoi, A., Prünster, I. and Walker, S.G. (2008a). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18**, 1519–1547.

Lijoi, A., Prünster, I. and Walker, S.G. (2008b). Investigating nonparametric priors with Gibbs structure. *Statist. Sinica* **18**, 1653–1668.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12**, 351–357.

Lo, A.Y. and Weng, C.-S. (1989). On a class of Bayesian nonparametric estimates. II. Hazard rate estimates. *Ann. Inst. Statist. Math.* **41**, 227–245.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Statist. Simulation Comp.* **23**, 727–741.

MacEachern, S.N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria: American Statistical Association, 50-55.

MacEachern, S.N. (2000). Dependent Dirichlet processes. *Technical Report*. Department of Statistics, Ohio State University.

MacEachern, S.N. (2001). Decision theoretic aspects of dependent nonparametric processes. In *Bayesian methods with applications to science, policy and official statistics* (E. George, Ed.), 551–560. International Society for Bayesian Analysis, Crete.

MacEachern, S.N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comp. Graph. Statist.* **7**, 223–238.

Mauldin, R.D., Sudderth, W.D. and Williams, S.C. (1992). Pólya trees and random distributions. *Ann. Statist.* **20**, 1203–1221.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canad. J. Statist.* **26** 283–297.

Müller, P. and Quintana, F.A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19**, 95–110.

Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Statist.* **9**, 249–265

Nieto–Barajas, L.E. and Prünster, I. (2009). A sensitivity analysis for Bayesian nonparametric density estimators. *Statist. Sinica* **19**, 685-705.

Nieto-Barajas, L.E., Prünster, I. and Walker, S.G. (2004). Normalized random measures driven by increasing additive processes. *Ann. Statist.* **32**, 2343–2360.

Nieto-Barajas, L.E. and Walker, S.G. (2002). Markov beta and gamma processes for modelling hazard rates. *Scand. J. Statist.* **29**, 413–424.

Nieto-Barajas, L.E. and Walker, S.G. (2004). Bayesian nonparametric survival analysis via Lévy driven Markov processes. *Statist. Sinica* **14**, 1127–1146.

Muliere, P. and Walker, S.G. (1997). A Bayesian non–parametric approach to survival analysis using Pólya trees. *Scand. J. Statist.* **24**, 331–340.

Paddock, S.M, Ruggeri, F., Lavine, M. and West, M. (2003). Randomized Pólya tree models for nonparametric Bayesian inference. *Statist. Sinica* **13**, 443–460.

Papaspiliopoulos, O. and Roberts, G.O. (2008). Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika* **95**, 169–186.

Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92**, 21–39.

Petrone, S. (1999a). Random Bernstein polynomials. *Scand. J. Statist.* **26**, 373–393.

Petrone, S. (1999b). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **27**, 105–126.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102**, 145–158.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (T.S. Ferguson, L.S. Shapley and J.B. MacQueen, Eds.). IMS Lecture Notes Monogr. Ser., Vol. **30**. Inst. Math. Statist., Hayward, 245–267.

Pitman, J. (2003). Poisson-Kingman partitions. In *Statistics and science: a Festschrift for Terry Speed* (D.R Goldstein, Ed.). IMS Lecture Notes Monogr. Ser., Vol. **40**. Inst. Math. Statist., Beachwood, 1–34.

Pitman, J. (2006). *Combinatorial stochastic processes.* Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002. Lecture Notes in Math., Vol. **1875**. Springer, Berlin.

Pitman, J. and Yor, M. (1997a). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.

Pitman, J., and Yor, M. (1997b). On the relative lengths of excursions derived from a stable subordinator. In *Séminaire de Probabilités XXXI* (Azema, J., Emery, M. and Yor, M., Eds.), Lecture Notes in Math., Vol. **1655**. Springer, Berlin, 287–305.

Prünster, I. (2002). *Random probability measures derived from increasing additive processes and their application to Bayesian statistics.* Ph.d thesis, University of Pavia.

Regazzini, E. (2001). *Foundations of Bayesian statistics and some theory of Bayesian nonparametric methods.* Lecture Notes, Stanford University.

Regazzini, E., Guglielmi, A. and Di Nunno, G. (2002). Theory and numerical analysis for exact distribution of functionals of a Dirichlet process. *Ann. Statist.* **30**, 1376–1411.

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560–585.

Romik, D. (2004). Explicit formulas for hook walks on continual Young diagrams. *Adv. in Appl. Math.* **32**, 625–654.

Romik, D. (2005). Roots of the derivative of a polynomial. *Amer. Math. Monthly* **112**, 66–69.

Sato, K. (1999). *Lévy processes and infinitely divisible distributions.* Cambridge University Press, Cambridge.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statist. Sinica* **2**, 639–650.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897–902.

Teh, Y.W., Görür, D. and Ghahramani, Z. (2007). Stick-breaking Construction for the Indian Buffet. *Proceedings of 11th International Conference on Artifical Intelligence and Statistics (AISTATS 2007).*

Thibaux, R. and Jordan, M.I. (2007). Hierarchical beta processes and the Indian buffet process. *Proceedings of 11th International Conference on Artifical Intelligence and Statistics (AISTATS 2007).*

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simul. Computat.* **36**, 45–54.

Walker, S.G. and Damien, P. (1998). A full Bayesian non–parametric analysis involving a neutral to the right process. *Scand. J. Statist.* **25**, 669–680.

Walker, S.G., Damien, P., Laud, P.W. and Smith, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions. *J. R. Stat. Soc. Ser. B* **61**, 485–527.

Walker, S.G. and Damien, P. (2000). Representation of Lévy processes without Gaussian components. *Biometrika* **87**, 477–483.

Walker, S.G. and Mallick, B.K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *J. Roy. Statist. Soc. Ser. B* **59**, 845–860.

Walker, S.G. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Statist.* **25**, 1762–1780.

Wolpert, R.L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, 251–267.