Brazilian Journal of Probability and Statistics
2012, Vol. 26, No. 4, 423–449
DOI: 10.1214/11-BJPS176

Predictive construction of priors in Bayesian nonparametrics

Sandra Fortini and Sonia Petrone
Bocconi University

Abstract. The characterization of models and priors through a predictive approach is a fundamental problem in Bayesian statistics. In the last decades, it has received renewed interest, as the basis of important developments in Bayesian nonparametrics and in machine learning. In this paper, we review classical and recent work based on the predictive approach in these areas. Our focus is on the predictive construction of priors for Bayesian nonparametric inference, for exchangeable and partially exchangeable sequences. Some results are revisited to shed light on theoretical connections among them.

1 Introduction

The characterization of models and priors through a predictive approach is a fundamental, long studied problem in Bayesian statistics. It is a point often underlined by de Finetti that one can only express a subjective probability on observable facts, and parametric models are just a link of the chain that leads from past experience to the probability of future observable facts. See the discussion and references in Cifarelli and Regazzini (1996), and Kallenberg (2005).

However, these fundamental problems have often been thought as mainly theoretical, and their potentiality in applications somehow undervalued. In fact, in the recent years they have received a new, exciting vigor in research areas with strong applied motivations and impact, in particular in the machine learning community and in Bayesian nonparametrics. In this note, we give a brief overview of some classical and recent work based on the predictive approach in these areas. Some results are revisited to shed light on theoretical connections among them. In Section 2, we discuss predictive constructions of nonparametric priors for exchangeable sequences. In Section 3, we consider Markov exchangeable sequences. Section 4 briefly discusses more recent developments and open problems, and concludes the paper.

2 Exchangeable sequences

The notion of exchangeability has a fundamental role in Bayesian statistics. Let $(X_n, n \geq 1)$ be a sequence of exchangeable random quantities, with probability

Key words and phrases. Exchangeability, Dirichlet process, random probability measures, mixtures of Markov chains, infinite hidden Markov models, urn schemes.
Received July 2011; accepted October 2011.

law P. For the sake of simplicity, we will only consider the case of real valued random variables (r.v.'s) X_i, but the results can be extended to more general spaces. Let $\hat{F}_n(\cdot) = \sum_{i=1}^n \delta_{X_i}(\cdot)/n$ be the empirical distribution of $(X_1, \ldots, X_n)$, where δ_x denotes a probability measure degenerate on x, and $\mathcal{F}$ be the class of all probability distributions on $\mathbb{R}$, equipped with the sigma-field induced by the metric of weak convergence. We will usually denote by the same symbol a probability measure and the corresponding distribution function (d.f.).

The strong law of large numbers for exchangeable sequences establishes that the sequence of empirical distributions $(\hat{F}_n)$ converges weakly to a random distribution $\tilde{F}$, a.s., as $n \to \infty$. On this basis, we have the celebrated representation theorem for exchangeable sequences (de Finetti (1937)).

Theorem 1 (de Finetti representation theorem). *Let $(X_n, n \geq 1)$ be a sequence of exchangeable random variables with probability law P. Then there exists a unique probability measure μ such that, for any $n \geq 1$,*

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \int_{\mathcal{F}} \prod_{i=1}^n F(x_i)\, d\mu(F).$$

Furthermore, P almost surely, the sequence of the empirical distributions $\hat{F}_n$ converges weakly to a random d.f. $\tilde{F}$ that is distributed according to μ.

Through de Finetti representation theorem, exchangeability assumes a basic role in Bayesian inference, as the probability assumption on observable quantities that gives the fundamental justification of the *hypothetical approach*, where $X_i \mid \tilde{F} \overset{\text{i.i.d.}}{\sim} \tilde{F}$ (the statistical model) and $\tilde{F} \sim \mu$ (the prior). On the other hand, in the *predictive approach*, the emphasis is on the probability law P on the observable quantities (X_n), and model and prior are then, at least in principle, characterized by P.

This is a fundamental result, but of course, in practice one has to specify P, or the *de Finetti measure* μ. Note that μ is a probability law on the space $\mathcal{F}$ of all distributions on the sample space, and in this sense, its choice indissolubly implies a choice of the "statistical model," given by its support. We will refer to μ as a parametric prior, if it has support on a class $\mathcal{F}_\theta = \{F_\theta, \theta \in \mathbb{R}^p\}$ indexed by a finite-dimensional parameter; while a *nonparametric* prior typically has full (weak) support $\mathcal{F}$.

For years, a basic problem on which many authors have been working is what further conditions on the observable quantities, therefore on P, restrict the support of μ to a parametric class, and what relationship with the observable quantities have the parameters of the model. Characterizations of parametric models based on invariance and sufficiency conditions include Freedman (1963), Kingman (1972), Dawid (1978), Smith (1981), Diaconis and Freedman (1984), Diaconis, Eaton and Lauritzen (1992), Eaton et al. (1993), Iglesias et al. (2009). Further conditions on observable quantities are needed to also characterize the class of prior

distributions on the parameters of the model; see Diaconis and Ylvisaker (1979), Zabell (1982), Arellano-Valle, Bolfarine and Iglesias (1994), Arellano-Valle and Bolfarine (1995), Loschi, Iglesias and Arellano-Valle (2003).

In fact, with the seminal papers on Bayesian nonparametrics in the 1960–70s, another basic problem that arose is how to characterize *nonparametric* priors with full support $\mathcal{F}$, through a predictive approach. This is the problem on which we focus in the next sections.

2.1 Predictive constructions

In a predictive approach, one can at least in principle characterize the prior through the sequence of predictive distributions. Here we review some fundamental properties. A first basic result establishes the asymptotic behavior of the sequence of predictive distributions for an exchangeable sequence. For brevity, we will use the notation $P(X_{n+1} \le x \mid X_1 = x_1, \ldots, X_n = x_n) = P_n(x \mid x_1, \ldots, x_n)$, or sometimes, shortly, $P_n(x)$. Let $\tilde{F}$ be the a.s. limit of the sequence of the empirical distributions. Then the sequence of predictive distributions converges almost surely to $\tilde{F}$ (see Fortini, Ladelli and Regazzini (2000), Proposition 5.7). This result is based on de Finetti's work about the approximation of the predictive distribution through the empirical distribution; in the case of 0–1 exchangeable random variables, he proved that $|P(X_{n+1} = 1 \mid X_1, \ldots, X_n) - \sum_{i=1}^{n} X_i/n| \to 0$, a.s. P; see Cifarelli and Regazzini (1996). A stronger result has been recently given by Berti and Rigo (1997), who prove a Glivenko–Cantelli theorem for exchangeable sequences, establishing that $\sup_x |P_n(x) - \hat{F}_n(x)| \to 0$, a.s. P. It follows that $P_n \to \tilde{F}$ weakly, a.s. P. Based on this result, de Finetti representation theorem can be stated in terms of the predictive distributions.

Theorem 2 (de Finetti representation theorem in terms of predictive distributions). *Let $(X_n, n \ge 1)$ be an exchangeable sequence of r.v.'s, with probability law P, and let $X_1 \sim P_1$ and P_n be the predictive distribution of $X_{n+1} \mid X_1, \ldots, X_n$, for $n \ge 1$. Then the sequence (P_n) converges weakly to a random d.f. $\tilde{F}$, a.s. P. Furthermore for any $n \ge 1$,*

$$P(X_1 \le x_1, \ldots, X_n \le x_n) = \int \prod_{i=1}^{n} F(x_i)\, d\mu(F),$$

where μ is the probability distribution of $\tilde{F}$.

Thus, the law P of an exchangeable sequence can be represented in terms of a *statistical model* $\tilde{F}$ (a random probability measure) which is the limit, under P, of the sequence of the predictive distributions (P_n). The de Finetti measure μ here arises as the limiting probability law of such sequence.

Theorem 2 suggests that one can *construct* an exchangeable probability law P by specifying the sequence of predictive distributions. Let $P_0(x)$ be a probability distribution and, for every $n \geq 1$, let $P_n(x \mid x_1, \ldots, x_n)$ be a *transition probability distribution* (i.e., P_n is a measurable function of $x_1, \ldots, x_n$). According to Ionescu–Tulcea theorem there exists a unique probability measure P on $\mathbb{R}^\infty$ such that P_0 is the distribution of the first coordinate and (P_n) is the sequence of predictive distributions.

Further conditions are needed in order to have an *exchangeable* P. Necessary and sufficient conditions such that P is exchangeable are given by Fortini, Ladelli and Regazzini (2000). Suppose first that P is exchangeable, with predictive distributions (P_n). Then it holds a.s.

(a) $P_n(A \mid x_1, \ldots, x_n) = P_n(A \mid x_{i_1}, \ldots, x_{i_n})$, for any permutation $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$ and $n \geq 2$;
(b) $\int_B P_{n+1}(A \mid x_1, \ldots, x_{n+1}) P_n(dx_{n+1} \mid x_1, \ldots, x_n) = \int_A P_{n+1}(B \mid x_1, \ldots, x_{n+1}) P_n(dx_{n+1} \mid x_1, \ldots, x_n)$

for every $A, B, x_1, \ldots, x_n$ and $n \geq 0$ (for $n = 0$ set $P_n(\cdot \mid x_1, \ldots, x_n) \equiv P_0(\cdot)$).

In terms of the exchangeable sequence (X_n), condition (b) states that $P(X_{n+1} \in B, X_{n+2} \in A \mid X_1, \ldots, X_n) = P(X_{n+1} \in A, X_{n+2} \in B \mid X_1, \ldots, X_n)$. Fortini, Ladelli and Regazzini (2000), Theorem 3.1, show that these conditions are also sufficient in order that the sequence of predictive distributions is consistent with an exchangeable P. Thus, one can construct an exchangeable probability law P by assigning a sequence of transition probabilities that satisfy conditions (a) and (b).

In this approach, parametric models can be characterized through further assumptions on the predictive structure. By exchangeability, the empirical d.f. $\hat{F}_n$ is sufficient to predict $X_{n+1}, X_{n+2}, \ldots$. Assume that a further summary $T(\hat{F}_n)$, with values in an Euclidean space, is predictive sufficient, that is, the conditional probability law of $X_{n+1}, X_{n+2}, \ldots \mid \hat{F}_n$ depends on $\hat{F}_n$ only through $T(\hat{F}_n)$. Then, in particular,

$$P(X_{n+1} \leq x \mid X_1, \ldots, X_n) = P\big(X_{n+1} \leq x \mid T(\hat{F}_n)\big).$$

By the previous results, $P_n(x \mid X_1, \ldots, X_n) = P_n(x \mid T(\hat{F}_n))$ converges weakly to a random distribution $\tilde{F}$, a.s. P. Under regularity assumptions, such limit can be expressed in parametric form, depending on $T(\tilde{F})$ (Fortini, Ladelli and Regazzini (2000), Theorem 7.1). Informally, $T(\hat{F}_n) \to T(\tilde{F}) \equiv \tilde{\theta}$, a.s. P, and

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \int_\Theta \prod_{i=1}^n P(X_i \leq x \mid \theta)\, d\mu^*(\theta),$$

where μ^* is the limit distribution of the sequence of predictive sufficient statistics $T(\hat{F}_n)$. Thus, the parametric model arises as the limit of the predictive distributions, and the prior is interpretable as the limit law of the predictive sufficient statistics $T(\hat{F}_n)$. An example of this construction is the predictive characterization

of the exponential family (see Fortini, Ladelli and Regazzini (2000) and references therein). The basic result of Zabell (1982), that characterizes the Dirichlet conjugate prior for multinomial observations through Johnson's sufficiency postulate, can be also regarded in this framework.

In the next section, we focus on predictive constructions of *nonparametric* priors, whose support cannot be indexed by a finite-dimensional parameter θ. From the above results, this implies that the complete empirical distribution is needed for prediction of future observations, that is, a finite-dimensional summary cannot be predictive sufficient. A main reference on Bayesian nonparametric inference is the monograph by Ghosh and Ramamoorthi (2003). Recent developments are discussed in Müller and Quintana (2004) and Hjort et al. (2010).

2.2 Pólya sequences and Dirichlet process

The Dirichlet process is the most popular prior for Bayesian nonparametric inference. Blackwell and MacQueen (1973) described the construction of a Dirichlet process prior by a generalization of the k-colors Pólya's urn scheme. Given a positive measure $\alpha(\cdot)$ on the Borel subset of $\mathbb{R}$, with $0 < \alpha(\mathbb{R}) < \infty$, let $\alpha = \alpha(\mathbb{R})$ and $F_0(\cdot) = \alpha(\cdot)/\alpha(\mathbb{R})$. Blackwell and McQueen define a *Pólya sequence* as a sequence of random variables (X_n) such that $X_1 \sim F_0$ and for $n \geq 1$,

$$X_{n+1} \mid X_1, \ldots, X_n \sim P_n = \frac{\alpha F_0 + \sum_{i=1}^{n} \delta_{X_i}}{\alpha + n}. \tag{2.1}$$

Theorem 3 (Blackwell and McQueen (1973)). *Let (X_n) be a Pólya sequence of r.v.'s such that $X_1 \sim F_0$ and (2.1) holds for $n \geq 1$. Then*

(i) *P_n converges a.s. as $n \to \infty$ to a random discrete distribution $\tilde{F}$;*
(ii) *$\tilde{F}$ has a Dirichlet process distribution, with parameter αF_0, $\tilde{F} \sim \mathrm{DP}(\alpha F_0)$;*
(iii) *$X_1, X_2, \ldots$ are a random sample from $\tilde{F}$, in the sense that $X_i \mid \tilde{F} \overset{i.i.d.}{\sim} \tilde{F}$.*

The discrete nature of the Dirichlet process, established by (ii) above, implies that ties are observed in a random sample $(X_1, \ldots, X_n)$ from $\tilde{F}$ with positive probability, as it is also clearly shown by the predictive rule (2.1). Thus, the Dirichlet process induces a random partition of $\{1, \ldots, n\}$, defined as $i \sim j$ (i and j are in the same group) if $X_i = X_j$.

The probability law of the random partition is usefully described by Hoppe's urn scheme (Hoppe (1984)). Consider sampling from an urn that initially contains $\alpha > 0$ black balls. At time n a ball is picked at random from the urn. If it is black, it is returned together with an additional ball of a previously unobserved color; if it is colored, it is returned together with an additional ball of the same color. Natural numbers are used to label the colors and they are chosen sequentially as the need arises. The sampling generates a process $(S_n, n \geq 1)$, where the random variable S_n is the label of the additional ball returned after the nth drawing. Initially there are only black balls, thus $S_1 = 1$; then $S_2 = 1$ or 2, $S_3 = 1, 2$ or 3, etc.

In this scheme, the actual colors have no relevance, the interest being in their labels (S_n) and, consequently, in the random partition that they describe. Indeed, for any $n \geq 1$, the random vector $(S_1, \ldots, S_n)$ defines a random partition ρ_n of $\{1, 2, \ldots, n\}$ in k nonempty sets $(A_1, \ldots, A_k)$ defined by $A_j = \{i \in \{1, 2, \ldots, n\} : S_i = j\}$ for $j = 1, \ldots, k$, where $k = \max(S_1, \ldots, S_n)$ is the number of distinct labels in $(S_1, \ldots, S_n)$. Note that the partition is ordered, specifically the A_j's are in order of appearance. It is easy to check that, if $(A_1, \ldots, A_k)$ is the partition corresponding to a realization $(s_1, \ldots, s_n)$ of $(S_1, \ldots, S_n)$, then

$$P\big(\rho_n = (A_1, \ldots, A_k)\big) = P(S_1 = s_1, \ldots, S_n = s_n) = \frac{\alpha^k}{\alpha^{[n]}} \prod_{j=1}^{k} (n_j - 1)!, \tag{2.2}$$

where $\alpha^{[n]} = \alpha(\alpha + 1) \cdots (\alpha + n - 1)$ and n_j is the number of elements of A_j, $j = 1, \ldots, k$. The probability law of the random partition ρ_n, for $n \geq 1$, is called *partition probability function*.

In genetics, it is of interest to consider the *allelic partition* corresponding to the occupancy numbers $(n_1, \ldots, n_k)$, described by $a_n = (m_1, \ldots, m_n)$ where m_j is the number of $(n_1, \ldots, n_k)$ that are equal to j; hence $\sum_{j=1}^{n} j m_j = n$. It corresponds to the partition $(A_1, \ldots, A_k)$ but now the order of the elements is not relevant. Thus, its probability is computed by multiplying (2.2) by the number of possible partitions with the same occupancy numbers $(n_1, \ldots, n_k)$, which gives (Antoniak (1974)):

$$P(a_n) = \frac{\alpha^k}{\alpha^{[n]}} \frac{n!}{1^{m_1} \cdots n^{m_n}} \frac{1}{m_1! \cdots m_n!}. \tag{2.3}$$

Note that $\sum_{j=1}^{n} m_j = k$ and $1^{m_1} \cdots n^{m_n} = n_1 \cdots n_k$. This is the celebrated Ewens sampling formula (Ewens (1972)). Hoppe (1984) shows that the sequence (a_n) is Markov, with marginal distribution given by (2.3).

Clearly, the sequence $(S_n, n \geq 1)$ is not exchangeable. However, we can associate another process to the urn sampling, the process of *colors*, which is exchangeable. If one "paints" the sequence (S_n), generating the colors at random from a diffuse color distribution F_0 (i.e., as i.i.d. draws from F_0, where $F_0(\{x\}) = 0$ for any x), then the resulting sequence of *colors* (X_n) is a Pólya sequence, with $X_1 \sim F_0$ and, for any $n > 1$, $X_{n+1} \mid (X_1, \ldots, X_n) \sim P_n$ as in (2.1). Therefore, by the results of Blackwell and MacQueen (1973), the sequence (X_n) is exchangeable, and its de Finetti measure is a $\mathrm{DP}(\alpha F_0)$. Colored Hoppe's urn provides a natural way of decomposing the joint distribution of $(X_1, \ldots, X_n)$ in terms of the probability of the random partition, generated by $(S_1, \ldots, S_n)$, and the density of the distinct colors. Roughly speaking,

$$p(x_1, \ldots, x_n) = p(s_1, \ldots, s_n) \prod_{j=1}^{d_n} f_0(x_j^*), \tag{2.4}$$

where $x_1^*, \ldots, x_{d_n}^*$ are the distinct values in $(x_1, \ldots, x_n)$, with common density f_0, and the labels $s_1, \ldots, s_n$ identify the random partition generated by $(x_1, \ldots, x_n)$; see Antoniak (1974). In terms of the well-known Chinese restaurant metaphor (see, e.g., Pitman (1996), Section 4), the labels $(S_1, \ldots, S_n)$ generated by Hoppe's urn give the allocation of customers at tables; then, tables are painted at random from the diffuse color distribution F_0.

Random partitions are of interest in several fields, such as combinatorics and genetics. In Bayesian inference, this feature of the Dirichlet process is widely exploited in hierarchical models, to model clustering and dimension reduction. In particular, a Dirichlet process mixture model assumes that

$$\begin{aligned} Y_i \mid \theta_i &\overset{\text{indep}}{\sim} f(y \mid \theta_i), \qquad i = 1, \ldots, n, \\ \theta_i \mid G &\overset{\text{i.i.d.}}{\sim} G, \\ G &\sim \mathrm{DP}(\alpha G_0). \end{aligned}$$

This model induces a random partition of the individual parameters $(\theta_1, \ldots, \theta_n)$ and thus a dimension reduction, or "clustering," of the data. See Quintana and Iglesias (2003) for a comparison with product partition models. Furthermore, the predictive urn scheme facilitates the implementation of MCMC algorithms for simulating from the conditional distributions of interest (MacEachern (1994) and Escobar and West (1995)). Thus, while a drawback in many applications, the discrete nature of the Dirichlet process, or, in other terms, the structure of the predictive rule which characterizes it, is in fact quite appropriate in hierarchical models, and it is one of the main reasons of its recent popularity in an extremely wide range of applied fields. Excellent overviews are given in Dunson (2010) and Teh and Jordan (2010). Some delicate related issues are pointed out by Petrone and Raftery (1997).

Exchangeable partition probability function. As we have seen, an exchangeable sequence (X_n) with a $\mathrm{DP}(\alpha F_0)$ de Finetti measure induces a random partition ρ_n of $\{1, \ldots, n\}$, whose probability law, if F_0 is diffuse, is given by (2.2). Note that the partition probability function (2.2) is a symmetric function of the size of the groups, $(n_1, \ldots, n_k)$. This property holds more generally. Given an exchangeable sequence (X_n), we can define a random partition $(A_1, \ldots, A_k)$ of $\{1, \ldots, n\}$ by letting i and j be in the same group if $X_i = X_j$. Then we have

$$P\big(\rho_n = (A_1, \ldots, A_k)\big) \equiv P\left(\bigcap_{j=1}^{k} (X_i = X_j^* \text{ for all } i \in A_j)\right) = p(n_1, \ldots, n_k) \quad (2.5)$$

for a symmetric function p of $(n_1, \ldots, n_k)$, where n_j is the number of elements in A_j. A partition probability function p so generated is called *exchangeable partition probability function* (EPPF) derived from the sequence (X_n). More formally,

p is defined on the space of sequences $\mathbf{n} = (n_1, n_2, \ldots)$, identifying $(n_1, \ldots, n_k)$ as $\mathbf{n} = (n_1, \ldots, n_k, 0, 0, \ldots)$. Let $\mathbf{n}^{j+}$ be defined from $\mathbf{n}$ by incrementing n_j by 1. Clearly, an EPPF p must satisfy

$$p(1, 0, 0, \ldots) = 1 \quad \text{and} \quad p(\mathbf{n}) = \sum_{j=1}^{k+1} p(\mathbf{n}^{j+}).$$

The concept of EPPF has been introduced by Pitman (1995), based on earlier relevant work by Kingman; see Kingman (1978).

Alternative definitions of the Dirichlet process. Ferguson (1973) gives an alternative definition of the Dirichlet process, as a normalized Gamma process. The idea is that, as the Dirichlet distribution is the joint distribution of a set of independent Gamma variables divided by their sum, the Dirichlet process can be constructed as a Gamma process with independent increments, divided by the sum.

Theorem 4 (Ferguson (1973)). *Let $\Gamma_{(1)} > \Gamma_{(2)} > \cdots$ be the ordered points of a Poisson random measure on $(0, \infty)$ with mean measure $\alpha x^{-1} \exp(-x)\, dx$. Let $p_i = \Gamma_{(i)} / \Gamma$, where $\Gamma = \sum_i \Gamma_{(i)}$, and define*

$$\tilde{F} = \sum_{j=1}^{\infty} p_j \delta_{\hat{X}_j}, \tag{2.6}$$

where the $\hat{X}_j$'s are i.i.d. according to F_0, independently on the $\Gamma_{(i)}$. Then $\tilde{F} \sim \mathrm{DP}(\alpha F_0)$, independently on Γ which has a Gamma$(\alpha, 1)$ *distribution.*

The distribution of the sequence of random weights (p_j) defined in (2.6), with $p_1 > p_2 > \cdots > 0$ and $\sum_{j=1}^{\infty} p_j = 1$ a.s., is called *Poisson–Dirichlet* with parameter (α); see Kingman (1975).

To have a closer intuition of this construction, consider first the finite case. Let $\Gamma_{i,N} \overset{\text{indep}}{\sim} \mathcal{G}(\alpha/N, 1)$, $i = 1, \ldots, N$, where we denote by $\mathcal{G}(a, b)$, or by $\mathcal{G}(\cdot \mid a, b)$, the Gamma distribution with density proportional to $x^{a-1} \exp(-bx)$. Let $\Gamma_N = \sum_i \Gamma_{i,N}$ and $p_{i,N} = \Gamma_{i,N} / \Gamma_N$. Then the random vector $(p_{1,N}, \ldots, p_{N,N})$ has a Dirichlet distribution $D(\alpha/N, \ldots, \alpha/N)$. Note that, for any set A, the number $N(A)$ of $\Gamma_{i,N}$ which lie in A, has a binomial distribution with parameters $N, \mathcal{G}(A \mid \alpha/N, 1)$; and for disjoint sets $A_1, \ldots, A_m$, we have $(N(A_1), \ldots, N(A_m)) \sim$ multinomial$(N, \mathcal{G}(A_1 \mid \alpha/N, 1), \ldots, \mathcal{G}(A_m \mid \alpha/N, 1))$.

Then for large N, $N(A)$ is approximately Poisson$(\alpha \mathcal{G}(A \mid 0, 1))$, where $\mathcal{G}(\cdot \mid 0, 1)$ is the improper gamma distribution with density $x^{-1} \exp(-x)$, and $N(A_1), \ldots, N(A_m)$ are approximately independent. Kingman (1975) motivates interest for the case where α is small and N large in the "heap problem," when a few items are relatively popular, while there is a long tail of items more rarely demanded.

To study the limit case for $N \to \infty$, the devise is to "embed" the vector $(p_{i,N}, \ldots, p_{N,N})$ in a process; more precisely, to regard it as a vector of increments of a Gamma process. Let $(\xi(t), t \geq 0)$ be a Gamma process with $\xi(0) = 0$, such that the increments of ξ on disjoint intervals are independent, and $\xi(t_2) - \xi(t_1) \sim \mathcal{G}(\alpha(t_2 - t_1), 1)$. The process ξ increases only in jumps, and can be constructed as

$$\xi(t) = \sum_i \Gamma_{(i)} \delta_{U_i}(t),$$

where the heights $(\Gamma_{(i)})$ of the jumps are the points, in decreasing order, of a nonhomogeneous Poisson process with mean function $\gamma(\cdot) = \alpha \mathcal{G}(\cdot \mid 0, 1)$ (thus, the number of Γ_i in A has a Poisson$(\alpha \mathcal{G}(A \mid 0, 1))$ distribution, and arrivals in disjoint sets are independent), and the atoms $U_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$, independently on (Γ_i). Now, consider the normalized increments of $\xi(\cdot)$:

$$q_{j,N} = \frac{\xi(j/N) - \xi((j-1)/N)}{\xi(1)}.$$

Because $q_{j,N} \sim \mathcal{G}(\alpha/N, 1)$ and $\xi(1) = \sum_{i=1}^{N} (\xi(j/N) - \xi((j-1)/N))$, the vector of increments $q_{j,N}$ has a Dirichlet distribution, $(q_{1,N}, \ldots, q_{N,N}) \sim D(\alpha/N, \ldots, \alpha/N)$, and we can study its limit behavior as $N \to \infty$ in place of that of $(p_{1,N}, \ldots, p_{N,N})$. It can be shown (Kingman (1975)) that the ordered vector $q_{(1),N} \geq q_{(2),N} \geq \cdots \geq q_{(N),N}$ is such that $q_{(j),N} \to \Gamma_{(j)}/\xi(1)$. Thus, the limiting distribution of $(q_{1,N}, \ldots, q_{N,N})$ is the distribution of the ordered normalized jumps $\Gamma_{(i)}/\xi(1)$ of the Gamma process on the interval $(0, 1)$, a Poisson–Dirichlet distribution with parameter α.

Furthermore, if we define a Gamma process such that

$$\xi(A) = \sum_i \Gamma_{(i)} \delta_{\hat{X}_i},$$

where the $\hat{X}_i$ are i.i.d. according to a distribution F_0, independently on the Γ_i, then $\xi(A) \sim \mathcal{G}(\alpha F_0(A), 1)$ and the increments on disjoint sets are independent. Therefore, the normalized process

$$\tilde{F}(\cdot) = \sum_i \frac{\Gamma_{(i)}}{\Gamma} \delta_{\hat{X}_i}(\cdot),$$

where $\Gamma = \sum_i \Gamma_i$, has Dirichlet finite-dimensional distributions with parameters driven by αF_0, thus it has a DP(αF_0) probability law, as stated in Theorem 4.

Completely random measures. The construction (2.6) of the Dirichlet process is a notable example of a more general construction of nonparametric priors via normalized completely random measures (CRM) (Kingman (1967); see also Regazzini, Lijoi and Prünster (2003), and Lijoi and Prünster (2010) for an excellent review). A completely random measure $m(\cdot)$ on $\mathbb{R}$ is characterized by the property

that for disjoint sets $A_1, \ldots, A_m$, the random variables $m(A_1), \ldots, m(A_m)$ are independent. It can be shown that a CRM (with no determinist components) can be constructed by first constructing a Poisson process Π on the product space $\mathbb{R} \times (0, \infty)$ with mean function $\mu(\cdot)$ (thus, the number of points in a subset B of the product space has a Poisson($\mu(B)$) distribution). Then, define $m(A)$ as the sum of the values of Γ_i for all points $(\hat{X}_i, \Gamma_i)$ of Π for which $\hat{X}_i$ lies in A:

$$m(A) = \sum_i \Gamma_i \delta_{\hat{X}_i}(A).$$

Then $m(\cdot)$ is a completely random measure, whose distribution is determined by $\mu(\cdot)$ (through it is referred as its *Lévy intensity*). A random probability measure can be constructed by normalizing the CRM $m(\cdot)$.

It is usually assumed that the CRM is *homogeneous*, that is

$$\mu(d\Gamma, dx) = \gamma(d\Gamma) \times \alpha(dx), \tag{2.7}$$

that is, the heights (Γ_i) of the jumps of $m(\cdot)$ are independent on the locations $(\hat{X}_i)$. For the Dirichlet process,

$$\mu(d\Gamma, dx) = \mathcal{G}(d\Gamma \mid 0, 1) \times \alpha F_0(dx)$$

(where α is now a positive scalar), which defines a Gamma process.

In this general construction, different random probability measures can be characterized, by different choices of the Lévy measure. Kingman (1975) defines the *σ-stable process*, which has a geometric tail behavior, different from the exponential tail behavior of the Dirichlet process. A construction based on normalized Inverse-Gaussian processes is given by Lijoi, Mena and Prünster (2005). Thibaux and Jordan (2007) show that a *Beta process* (Hjort (1990)) can be constructed by taking an improper Beta$(0, c)$ distribution as the measure γ in (2.7); interestingly, they show that this is the de Finetti measure corresponding to the *Indian Buffet* process (Griffiths and Ghahramani (2006)).

Stick-breaking representation. The Dirichlet process is interestingly related to the problem of species sampling (Fisher, Corbet and Williams (1943); see Pitman (1996)). Consider a population of N distinct species and suppose that individuals of the ith species are trapped according to a homogeneous Poisson process with rate Γ_i; that is, letting $N_i(t)$ be the arrivals of species i, we have $N_i(t) \mid \Gamma_i \sim$ Poisson process with rate Γ_i. Γ_i represents the *abundance* of species i, and it is assumed that $\Gamma_i \overset{\text{indep}}{\sim} \mathcal{G}(\alpha/N, 1)$ for some $\alpha > 0$. Consider the ranked abundances $\Gamma_{(1)} > \cdots > \Gamma_{(N)}$. Fisher already noticed that this model has an interesting behavior for $N \to \infty$. McCloskey (1965) studied the limit model, where (in the terms of Theorem 4) the species abundances Γ_i are generated as a nonhomogeneous Poisson process with rate measure $\alpha\mathcal{G}(0, 1)$. Suppose that the jth species to appear in the sampling from such limit model is a species whose abundance is $\Gamma_{(\pi_j)}$, and denote by $p_j^* = p_{\pi_j}$ the corresponding relative frequency. Then from the properties of

Poisson processes it follows that (p_i^*) is a *size-biased permutation* of the sequence of ranked weights $(p_i = \Gamma_{(i)}/\Gamma)$; that is, for all finite sequences $(i_1, \ldots, i_k)$, the conditional probability of the event $(\pi_1 = i_1, \ldots, \pi_k = i_k)$ given $(p_1, p_2, \ldots)$ is

$$p_{i_1} \frac{p_{i_2}}{1 - p_{i_1}} \cdots \frac{p_{i_k}}{1 - p_{i_1} - \cdots - p_{i_{k-1}}}.$$

McCloskey (1965) shows that, if (p_i^*) is a size-biased permutation of a sequence of random variables $p_1 \geq p_2 \geq \cdots > 0$ with $\sum_i p_i = 1$, then it holds

$$p_j^* = V_j \prod_{i=1}^{j-1} (1 - V_i) \tag{2.8}$$

for a sequence of i.i.d. random variables (V_i) if and only if (p_i) has a Poisson–Dirichlet(α) distribution. In this case the V_i are i.i.d. according to a Beta distribution with parameters $(1, \alpha)$. The distribution of the sequence (p_i^*) so obtained is called *stick-breaking*, since the p_i^* in (2.8) can be interpreted as successive breaks of a stick of unit length, or GEM(α), after Griffiths, Engen and McCloskey, who contributed to its development in genetics and ecology.

From these results, it follows that if $\tilde{F} \sim \mathrm{DP}(\alpha F_0)$, then $\tilde{F}$ is a.s. represented as

$$\tilde{F} = \sum_{i=1}^{\infty} p_i^* \delta_{X_i^*},$$

where the weights (p_i^*) have a stick breaking distribution (2.8) and the X_i^* are i.i.d. according to F_0, independently on (p_i^*).

These results can be usefully regarded in a predictive approach. Consider a Pólya sequence (X_n), with predictive rule: $X_1 \sim F_0$ with F_0 diffuse, and

$$X_{n+1} \mid X_1, \ldots, X_n, K_n = k \sim P_n = \frac{\alpha}{\alpha + n} F_0 + \frac{1}{\alpha + n} \sum_{j=1}^{k} n_j \delta_{X_j^*}, \tag{2.9}$$

where K_n is the number of distinct "species" in the sample $(X_1, \ldots, X_n)$; given $K_n = k$, $(X_1^*, \ldots, X_k^*)$ are the distinct values in the sample, in the order of appearance, and $n_1, \ldots, n_k$ is the vector of their counts. Then, as shown by Blackwell and MacQueen (1973), (X_n) is exchangeable, with a DP(αF_0) de Finetti measure. But it can also be proved (see Theorem 5 and the following discussion in the next section) that the sequence of predictive distributions (P_n) converges to a random measure

$$\tilde{F} = \sum_{j=1}^{\infty} p_j^* \delta_{X_j^*}, \tag{2.10}$$

where X_j^* is the jth distinct value that appears, with $X_j^* \overset{\text{i.i.d.}}{\sim} F_0$, independently on (p_j^*); and the weights (p_j^*) are a size-biased permutation of the (p_i) in (2.6) and

have a stick-breaking(α) distribution. By the asymptotic properties of the predictive distribution discussed in Section 2.1, it follows that, if $\tilde{F} \sim \mathrm{DP}(\alpha F_0)$, with F_0 diffuse, it is a.s. equal to (2.6) and to (2.10). This clarifies the relationship between the discrete representation (2.6) of the Dirichlet process and the, possibly more popular, stick-breaking one (Sethuraman (1994)), given by (2.10), which here arises as the limit of the sequence of predictive distributions.

2.3 Species sampling models

Pitman (1996) defines a class of predictive rules, in the framework of *species sampling*, that generalizes Blackwell and McQueen scheme (2.1). The problem of species sampling is widely studied in ecology, genetics and population dynamics. Suppose that a random sample $X_1, X_2, \ldots$ is drawn from a large population of individuals of various species, and X_i represents the species of the ith individual sampled. The space $\mathcal{X}$ of possible values of X_i is thought as an arbitrary set of tags or colors, used to label the various species. It is assumed that to the jth distinct species to appear in the sample it is assigned a tag X_j^*, where the X_j^* are i.i.d. according to a diffuse distribution F_0. This framework suggests the following predictive rule for the sequence (X_n):

$$\begin{aligned} X_1 &\sim F_0, \\ X_{n+1} \mid X_1, \ldots, X_n, K_n = k &\sim \sum_{j=1}^{k} p_j(\mathbf{N}_n)\delta_{X_j^*} + p_{k+1}(\mathbf{N}_n)F_0, \end{aligned} \tag{2.11}$$

where, as in (2.9), K_n is the number of different species to appear in the first n observations, X_j^* is the jth species to appear, and $\mathbf{N}_n = (N_{1,n}, N_{2,n}, \ldots)$ is the vector of counts of various species observed in the sample $(X_1, \ldots, X_n)$. This means that, given $(X_1, \ldots, X_n)$ such that species j is observed n_j times, the next individual X_{n+1} can be of the previously observed species X_j^*, with probability p_j, $j = 1, \ldots, k$, or it is a new species, randomly sampled from F_0, with probability p_{k+1}. The *predictive weights* $(p_1, \ldots, p_{k+1})$ in (2.11) only depend on the counts $(n_1, \ldots, n_k)$. Blackwell and McQueen's prediction rule (2.1) is the special case where $p_j = n_j/(\alpha + n)$ and $p_{k+1} = \alpha/(\alpha + n)$.

A sequence of r.v.'s (X_n) is called a *species sampling sequence* if it is an exchangeable sequence with prediction rule of the form (2.11) for a diffuse distribution F_0.

Theorem 5 (Pitman (1996), Proposition 11). *Suppose (X_n) is a species sampling sequence, and let P_n denote the predictive distribution of $X_{n+1} \mid X_1, \ldots, X_n$, as displayed in (2.11). Then*

(i) *P_n converges in total variation norm a.s. as $n \to \infty$ to the random distribution*

$$\tilde{F} = \sum_j p_j^* \delta_{X_j^*} + \left(1 - \sum_j p_j^*\right) F_0, \tag{2.12}$$

where p_j^* is the frequency of the jth species to appear, that is

$$p_j^* = \lim \frac{N_{j,n}}{n} \qquad a.s.;$$

(ii) *the X_j^* are i.i.d. according to F_0, independently of the p_j^*;*

(iii) *$(X_1, X_2, \ldots)$ is a sample from $\tilde{F}$, that is, $X_i \mid \tilde{F} \overset{i.i.d.}{\sim} \tilde{F}$.*

Note that the number K_∞ of distinct values in the infinite sequence $(X_1, X_2, \ldots)$ is a.s. equal to $\inf\{k : p_1^* + \cdots + p_k^* = 1\}$. Thus, the meaning of (ii) is that, conditionally on $(p_1^*, p_2^*, \ldots)$ with $K_\infty = k$, the X_j^* are i.i.d. according to F_0 for $1 \le j \le k$.

Theorem 5 is an extension of Blackwell and McQueen's Theorem 3. However, Theorem 5 makes the assumption that (X_n) is exchangeable, which was instead part of the conclusions in Theorem 3. Furthermore, it provides no explicit description of the distribution of $\tilde{F}$ (of the distribution of the weights (p_j^*)). In Theorem 3, instead, it is proved that $\tilde{F} \sim \mathrm{DP}(\alpha F_0)$. There are however further results, and remarkable examples where one can explicitly find the distribution of $\tilde{F}$.

First, we have a sort of reciprocal of Theorem 5: if we start from a sample (X_n) from a random distribution

$$\tilde{F} = \sum_i p_i \delta_{\hat{X}_i} + \left(1 - \sum_i p_i\right) F_0, \tag{2.13}$$

defined for some sequence of random variables (p_i) such that $p_i \ge 0$ and $\sum_i p_i \le 1$ a.s. and $\hat{X}_i$ i.i.d. according to a diffuse distribution F_0, independently of (p_i), then (X_n) is a species sampling sequence. This model, with $\tilde{F}$ defined as above and (X_n) a random sample from $\tilde{F}$, is called a *species sampling model*. The weight p_i in (2.13) is interpreted as the relative frequency of the ith species in some listing of the species present in the population, and $\hat{X}_i$ as the tag assigned to that species. Then a question is what is the relationship between the weights (p_i) and the weights (p_i^*) in (2.12). This question can be easily answered if the species sampling model is *proper*, that is, if $\sum_i p_i = 1$ a.s., so that $\tilde{F}$ is a.s. discrete. Then

$$\tilde{F} = \sum_i p_i \delta_{\hat{X}_i} = \sum_j p_j^* \delta_{X_j^*},$$

where X_j^* and p_j^* are defined as in (2.12), in terms of a sample (X_n) from $\tilde{F}$. Furthermore, if (p_i) is decreasing, the sequence (p_j^*) is a size-biased permutation of (p_j). This was the case of the Dirichlet process, where $\tilde{F} = \sum_i p_i \delta_{\hat{X}_i}$ with decreasing weights having a Poisson–Dirichlet(α) distribution, and then the (p_j^*) are a size-biased permutation of (p_i), with a GEM(α) distribution. Thus, these results extend the representations (2.6) and (2.10) of the Dirichlet process. It can

be shown that a species sampling model is proper if and only if $p_1^* > 0$ a.s., or if and only if $K_n/n \to 0$ a.s.

It remains to give conditions on the predictive weights $(p_j, j = 1, \ldots, k+1)$ in the predictive rule (2.11), such that exchangeability of (X_n) holds. It can be shown that exchangeability holds if and only if the predictive weights are defined in terms of an EPPF.

Theorem 6 (Pitman (1996), Theorem 14). *Let (X_n) be governed by the predictive rule (2.11), with a diffuse distribution F_0. The sequence (X_n) is exchangeable iff the predictive weights can be obtained as*

$$p_j = \frac{p(\mathbf{n}^{j+})}{p(\mathbf{n})} \qquad \text{for } j = 1, \ldots, k+1, p \text{ rovided } p(\mathbf{n}) > 0 \tag{2.14}$$

for a nonnegative, symmetric function p. Then (X_n) is a sample from $\tilde{F}$ as in Theorem 5, and the EPPF of (X_n) is the unique nonnegative symmetric function p such that (2.14) holds and $p(1, 0, 0, \ldots) = 1$.

The theorem is based on two results. It can be proved that, for each pair (p, F_0), where p is an EPPF, and F_0 is a diffuse distribution, there exists a unique distribution for a species sampling sequence (X_n) such that p is the EPPF of (X_n) and F_0 is the distribution of X_1. Furthermore, from formula (2.5) and Bayes rule, it follows that the predictive weights p_j that determine the prediction rule (2.11) of a species sampling sequence (X_n) can be computed in terms of the EPPF p of (X_n), as in (2.14).

Example 1 (Dirichlet process). If (X_n) is a Pólya sequence, with F_0 diffuse, it is easy to check that the predictive weights are obtained from the EPPF given by (2.2). Thus, (X_n) is governed by a species sampling model corresponding to the EPPF (2.2) and the distribution F_0.

Example 2 (Finite Dirichlet). Suppose that the predictive weights in the prediction rule (2.11) are given by

$$p_j = \frac{n_j + \alpha/N}{\alpha + n} \qquad \text{for } j = 1, \ldots, k \quad \text{and} \quad p_{k+1} = \frac{\alpha - k\alpha/N}{\alpha + n} \tag{2.15}$$

for a positive constant α and $k \leq N$. This predictive rule is obtained from the EPPF

$$p_\alpha(n_1, \ldots, n_k, 0, 0, \ldots) = \frac{\prod_{j=1}^{k-1}(\alpha - j\alpha/N)}{(1+\alpha)^{k-1}} \prod_{i=1}^{k} \left(1 + \frac{\alpha}{N}\right)^{[n_i - 1]}$$

for $k \leq N$. Since $p_\alpha(1, 0, 0, \ldots) = 1$ and p_α is symmetric, the sequence (X_n) defined by this predictive rule is exchangeable. Note that the number K_n of distinct species in a sample converges to N as n tends to infinity. In fact, this predictive

rule corresponds to sampling from a population of a finite number N of species, $\tilde{F} = \sum_{j=1}^{N} p_j \delta_{\hat{X}_j}$, where $(p_1, \ldots, p_N) \sim \text{Dirichlet}(\alpha/N, \ldots, \alpha/N)$ and the $\hat{X}_j$ are i.i.d. according to a diffuse F_0, independently on (p_i). For $N \to \infty$, the predictive distributions (2.15) converge to Blackwell and MacQueen's (2.1). In this sense, the prior on $\tilde{F}$ here defined gives a finite approximation of the Dirichlet process, and it is sometimes called *finite Dirichlet prior*; see Ishwaran and James (2001), who also discuss its stick-breaking representation, and Gibbs sampling methods for Bayesian computations. Extensions have been recently studied by Petrone, Guindani and Gelfand (2009).

Example 3 (Two parameters Poisson–Dirichlet process). Suppose that the predictive weights in the prediction rule (2.11) are defined as

$$p_j = \frac{n_j - \theta}{\alpha + n} \qquad \text{for } j = 1, \ldots, k \quad \text{and} \quad p_{k+1} = \frac{\alpha + k\theta}{\alpha + n} \tag{2.16}$$

for real parameters α and θ such that $0 \le \theta < 1$ and $\alpha > -\theta$. This predictive rule is obtained according to (2.14) for the EPPF

$$p_{(\alpha,\theta)}(n_1, \ldots, n_k, 0, 0, \ldots) = \frac{\prod_{j=1}^{k-1}(\alpha + j\theta)}{(1+\alpha)^{n-1}} \prod_{i=1}^{k} (1-\theta)^{[n_i - 1]}. \tag{2.17}$$

Since $p_{(\alpha,\theta)}(1) = 1$ and $p_{(\alpha,\theta)}$ is symmetric, the sequence (X_n) defined by this predictive rule is exchangeable. The predictive rule (2.16) characterizes the *two parameters Poisson–Dirichlet process*, introduced by Perman, Pitman and Yor (1992) and further studied by Pitman (1995) and Pitman and Yor (1997), and sometimes also referred as *Pitman–Yor process*. As appears from (2.14), the Poisson–Dirichlet process allows a more flexible predictive structure than the Dirichlet process, which may be too poor in some problems, depending only on the counts of the different sampled species. Instead, in (2.16) the predictive probability of observing a new species also depends on the number k of distinct species sampled. The distribution of the ranked weights in the population $\tilde{F}$ characterized by (2.16) is a *two parameters* extension of the *Poisson–Dirichlet distribution* that we have for the Dirichlet process, which corresponds to the case $\theta = 0$. By Theorem 5 and the following discussion, the predictive distribution (2.16) converges to a random measure $F = \sum_{j=1}^{\infty} p_j^* \delta_{X_j^*}$, where (p_i^*) is a size-biased permutation of (p_i). It can be shown that (p_i^*) has a stick-breaking representation as in (2.8), where $V_j \overset{\text{indep}}{\sim} \text{Beta}(\alpha + j\theta, 1 - \theta)$.

Pitman (2003) derived general laws, termed *Poisson–Kingman distributions*, for sequences of ranked probability masses (p_i). Gnedin and Pitman (2005) define a class of *Gibbs-type* EPPFs that extends (2.17):

$$p(n_1, \ldots, n_k) = w_{n,k} \prod_{j=1}^{k} (1-\theta)^{[n_j - 1]}$$

(for a suitable array $w_{n,k}$) from which one can define the predictive rule

$$X_{n+1} \mid X_1, \ldots, X_n \sim \frac{w_{n+1,k+1}}{w_{n,k}} F_0 + \frac{w_{n+1,k}}{w_{n,k}} \sum_{j=1}^{k} (n_j - \theta)\delta_{X_j^*}$$

characterizing a class of *Gibbs-type priors*. They give characterization results of Gibbs-type priors $\tilde{F} = \sum_j p_j \delta_{\hat{X}_j}$ in terms of ranked weights (p_j) with a Poisson–Kingman distribution.

For a wide review of these problems, and of nonparametric priors beyond the Dirichlet process, we refer again to Lijoi and Prünster (2010).

3 Mixtures of Markov chains

The previous results give predictive constructions of exchangeable probability laws. In this section, we review some nonparametric prior constructions for Markov exchangeable sequences.

3.1 de Finetti theorem for Markov chains

Diaconis and Freedman (1980) give a de Finetti theorem for Markov chains. We briefly remind their basic results. Let $(X_n, n \geq 0)$ be a stochastic process taking values on a countable set I, whose elements are referred as "states"; with no loss of generality, we can let $I = \{0, 1, 2, \ldots\}$. Two sequences $x = (x_0, \ldots, x_n)$ and $y = (y_0, \ldots, y_n)$ in I^n are *equivalent*, $x \sim y$, if they start from the same state and have the same transitions counts. The sequence (X_n) is partially exchangeable in the sense of Diaconis and Freedman (or, following the terminology of Zabell (1995), *Markov exchangeable*, to distinguish this notion from de Finetti's partial exchangeability), if $x \sim y$ implies $P(X_0 = x_0, \ldots, X_n = x_n) = P(X_0 = y_0, \ldots, X_n = y_n)$. The sequence (X_n) is *recurrent* if $P(X_n = X_0 \text{ for infinitely many } n) = 1$.

Theorem 7 (Diaconis and Freedman (1980), Theorem 7). *Let (X_n) be a recurrent sequence of random variables taking values in a (at most) countable set I. Then (X_n) is Markov exchangeable if and only if it is a mixture of Markov chains. That is, given the initial state x_0, there exists a unique probability measure (prior) μ on the space of transition matrices on I, such that*

$$P(X_1 = x_1, \ldots, X_n = x_n \mid X_0 = x_0) = \int \prod_{i=1}^{n} \pi_{x_{i-1}}(x_i)\, d\mu(\pi),$$

where $\pi_i(j) := \pi_{i,j}$ (i.e., π_i is the ith row of the transition matrix π, considered as a probability measure).

In other words, a recurrent process (X_n) is Markov exchangeable if and only if there exists a random transition matrix Π such that, conditionally on Π, (X_n) is a Markov chain with transition matrix Π. The (prior) distribution of Π is the probability measure μ in the above equation.

Define a x_0-block for the sequence (X_n) as a finite sequence of states that begins at x_0 and contains no further x_0. The proof of the above result is based on the fact that Markov exchangeability and recurrence imply exchangeability of the sequence $(B_n, n \geq 1)$ of the successive x_0-blocks. It is useful to note that the latter property implies that for any measurable transformation $\phi(B)$ (e.g., $\phi(B)$ may be the length of the x_0-block B), the sequence $(\phi(B_n), n \geq 1)$ is also exchangeable. Thus, the above theorem implies a de Finetti theorem for the exchangeable sequence $(\phi(B_n))$.

Following a suggestion in de Finetti (1959), Fortini et al. (2002) give a different characterization of mixtures of Markov chains, based on *successor states*. The properties of the sequences of successors states were already studied by Zabell (1995), in a beautiful note extending the characterization of conjugate Dirichlet priors through Johnson's sufficiency postulate to Markov exchangeable sequences. Fortini et al. (2002) clarify the relationship between Diaconis and Freedman's notion of partial exchangeability and de Finetti's one. The mth successor of state i, $X_{i,m}$, is defined as the value of the process immediately after the mth visit to state i. In order to avoid having rows of finite length, they introduce a "dummy state," denoted by ∂, and, if a state is visited only a finite number of times n, let $X_{i,m}$ equal to ∂ for $m > n$. Thus, the sequence of successors of state i is well defined, as an infinite sequence of random variables with values in $I^* = I \cup \partial$. It can be easily shown that, if the process (X_n) is recurrent and Markov exchangeable, then it is also *strongly recurrent*, meaning that for *any* state $i \in I$, $P(X_n = i \text{ infinitely often} \mid i \text{ is visited}) = 1$. Strong recurrence implies that any state is either never visited or it is visited infinitely often. In this case, the sequence of successors of any state i is either $(\partial, \partial, \ldots)$, if state i is never visited, or it is an infinite I-valued sequence, if state i is visited infinitely often.

Zabell (1995) shows that, if a process (X_n) is recurrent and Markov exchangeable, then the matrix of the successors states is partially exchangeable; that is, its distribution is invariant under permutations within rows. Fortini et al. (2002), Theorem 1, show that the reciprocal implication also holds; that is, a process (X_n) is recurrent and Markov exchangeable if and only if the matrix of the successors states is partially exchangeable. Under this hypothesis the process is a mixture of Markov chains, that is, there exists a stochastic transition matrix Π on I^*, such that, conditionally on Π, (X_n) is a Markov chain with transition matrix Π; furthermore, the prior distribution is uniquely determined (provided the class of transition matrices is suitably defined, see Fortini et al. (2002) for more details). It comes from their results that the prior distribution of Π is the de Finetti measure of the array of successors states.

The above mentioned results are exploited to provide predictive characterizations of priors for Bayesian inference on Markov chains. We discuss some important constructions, based on urn schemes, in the next sections. Intuitively, one can define a process along a sequence of related urns, such that draws from urn i represent the successors of state i; if drawn according to an exchangeable scheme, one can expect to characterize a mixture of Markov chains, together with the prior on the random transition matrix.

3.2 Reinforced urn processes

Muliere, Secchi and Walker (2000) define a class of *reinforced urn processes* (RUPs) that are Markov exchangeable, thus, when recurrent, are conditionally Markov. The urn scheme characterizes the prior on the transition matrix. RUPs have been applied for Bayesian nonparametric inference in different contexts, from survival analysis (Bulla, Muliere and Walker (2009)) to credit risk (Cirillo, Hüsler and Muliere (2010)).

RUPs are informally defined as random walks on a space of Pólya urns. More formally, a RUP is defined by four elements: a countable state space I, a finite set of colors $E = \{c_1, \ldots, c_k\}$, and a law of motion $q : (I \times E) \to S$; finally, to each $x \in I$ it is associated an urn U_x, with known initial composition $\alpha(x) = (\alpha_x(c_1), \ldots, \alpha_x(c_k))$, where $\alpha_x(c) \geq 0$ is the number of balls of color c initially contained in urn U_x, and we let $\alpha_x = \sum_{j=1}^{k} \alpha_x(c_j)$. It is assumed that the law of motion q has the property that, for every $x, y \in I$, there is at most one color $c(x, y) \in E$ such that $q(x, c(x, y)) = y$.

Given these ingredients, a RUP is defined as follows. Fix $X_0 = x_0$ and move to urn U_{x_0}. Pick a ball from U_{x_0} and return it, along with another ball of the same color. If $c \in E$ is the color of the sampled ball, set $X_1 = c$, and move to urn $U_{q(x,c)}$, as determined by the law of motion. If $q(x, c) = y$, say, move to urn U_y, pick a ball, and so on. Thus, balls are drawn from each urn according to a Pólya scheme, and one moves across urns according to the given law of motion. The process of colors (X_n) so defined is called RUP, with the four given elements.

The main property of RUPs is that they are Markov exchangeable. Therefore, a recurrent RUP (X_n) is a mixture of Markov chains. The reinforced urn scheme characterizes the probability law of the sequence (X_n) and therefore the prior μ. Muliere et al. (2000), Theorem 2.16, show that μ is such that the rows of Π are independent, and the xth row is a random probability measure on $(y_1 = q(x, c_1), \ldots, y_k = q(x, c_k))$, with probability masses $(\Pi_x(y_1), \ldots, \Pi_x(y_k))$ which have a Dirichlet distribution with parameters $(\alpha_x(c_1), \ldots, \alpha_x(c_k))$.

An interesting example of RUP gives a characterization of the Beta-Stacy process (Walker and Muliere (1997)), that is widely used as a prior in Bayesian nonparametric survival analysis. Suppose that: $I = \{0, 1, 2, \ldots\}$, the set of colors contains only two colors, white and black say, $E = \{w, b\}$, and the law of motion is such that $q(x, b) = x + 1$ and $q(x, w) = 0$, for all $x \in S$. From the previous results,

when the resulting RUP is recurrent, it is a mixture of Markov chains. Furthermore, letting T_n be the length on the nth 0-block (e.g., for a 0-block $(0, 1, 2, 3, 0)$, $T = 4$), the sequence $(T_n, n \geq 1)$ is exchangeable. Muliere et al. (2000) show that its de Finetti measure is a Beta-Stacy process on I with parameters $\{\alpha_j(w), \alpha_j(b)\}$. Thus, T_n can be interpreted as the survival time for the n-individual and, assuming that individuals are exchangeable, this construction gives a characterization of the Beta-Stacy as a prior on the survival times. These results can be extended to characterize neutral to the right processes (Doksum (1974)).

A restriction of RUPs is that they are defined only for a finite number of colors. This implies that, in each step, the chain can only reach a finite number of states; in other words, each row of the transition matrix has at most k nonzero entries; and the states that are reachable in one step from x have to be fixed a priori. For example, from state x, one can only move to the states $(q(x, c_1), \ldots, q(x, c_k))$, which is a restrictive assumption in many applications. An extension that allows for a countable set of colors is discussed in Fortini and Petrone (2011), and characterizes a prior on the random transition matrix such that the rows are independent Dirichlet processes.

3.3 Urn schemes for Bayesian inference in hidden Markov models

A clever and extremely fruitful urn scheme for Bayesian nonparametric inference in hidden Markov models (HMM) has been proposed by Beal, Ghahramani and Rasmussen (2002). HMMs are widely applied in several fields, from speech recognition to time series analysis. However, one restriction is that the number of latent states has to be known a priori. Beal, Ghahramani and Rasmussen (2002) suggest a model that allows Bayesian inference for HMMs without bounding a priori the number of states, which is therefore referred as *infinite HMM*.

The urn scheme which is the basis of the infinite HMM is defined as follows. Consider first an *oracle* Hoppe's urn, which initially contains γ black ball. The process of the colors' labels generated by Hoppe's sampling from the oracle urn is denoted by $(S_n^{(o)}, n \geq 1)$. When a new color with label i is sampled from the oracle urn, we create a Hoppe's urn labeled as U_i, which initially contains α black balls, and define a processes (S_n) generated by recursively sampling from these urns as follows.

We start with a draw from the oracle urn; since initially it contains only black balls, a new color with label 1 is generated, and we let $S_1^{(o)} = 1$ and $S_0 = 1$. Then, we create a Hoppe's urn U_1 and pick a ball from it. Being necessarily black, a new color should be generated, and to this aim we enquire the oracle urn. That is, we pick a ball from the oracle urn, and if it is labelled 1, we set $S_2^{(o)} = 1$ and $S_1 = 1$; if black, a new color with label 2 is generated, and we set $S_2^{(o)} = 2$ and $S_1 = 2$. Then we move to urn U_{S_1}, and so on.

Thus after n draws, given $(S_1 = s_1, \ldots, S_n = i, M = m, S_1^{(o)} = s_1^{(o)}, \ldots, S_m^{(o)} = s_m^{(o)})$, where M denotes the random number of draws from the oracle urn, we generate S_{n+1} as follows:

- with probability $t_{i,j}/(\alpha + t_i)$, $S_{n+1} = j$, for $j = 1, \ldots, d_m$, where $t_{i,j}$ are the transitions from i to j in $(s_1, \ldots, s_n)$ (i.e., the number of balls of label j in urn U_i), $t_i = \sum_j t_{i,j}$ and $d_m = \max(s_1^{(o)}, \ldots, s_m^{(o)})$ is the number of colors already generated from the oracle urn;
- with probability $\alpha/(\alpha + t_i)$, a black ball is sampled from U_i, thus a new draw is generated from the oracle urn:

$$S_{m+1}^{(o)} \mid M = m, S_1^{(o)} = s_1^{(o)}, \ldots, S_m^{(o)} = s_m^{(o)} \sim \frac{\gamma}{\gamma + m}\delta_{d_m+1} + \sum_{j=1}^{d_m} \frac{m_j}{\gamma + m}\delta_j,$$

where m_j is the number of j in $(s_1^{(o)}, \ldots, s_m^{(o)})$; and we let $S_{n+1} = S_{m+1}^{(o)}$.

Thus,

$$\begin{aligned} S_{n+1} \mid S_1 = s_1, \ldots, S_n = i, M = m, S_1^{(o)} = s_1^{(o)}, \ldots, S_M^{(o)} = s_m^{(o)} \\ \sim \sum_{j=1}^{d_m} \left(\frac{t_{i,j}}{\alpha + t_i} + \frac{\alpha}{\alpha + t_i} \frac{m_j}{\gamma + m} \right) \delta_j + \left(\frac{\alpha}{\alpha + t_i} \frac{\gamma}{\gamma + m} \right) \delta_{d_m+1}. \end{aligned}$$

The process (S_n) is not Markov exchangeable. However, if we "paint" the process $(S_i^{(o)})$ with colors ξ_j^*, i.i.d. from a diffuse color distribution F_0, the resulting process (X_n) defined by the colored (S_n) is Markov exchangeable and recurrent (see Fortini and Petrone (2011)). Thus, it is a mixture of Markov chains, for which the urn construction characterizes the mixing measure. More precisely, there exists a discrete random probability measure p_0 such that, conditionally on p_0, (X_n) is a mixture of Markov chains; precisely

- $(X_n) \mid p_0, \Pi$ is a Markov chain, with state space corresponding to the support $I(p_0)$ of p_0, transition matrix Π on $I(p_0)$ and initial distribution p_0:
- $\Pi \mid p_0$ is a transition matrix on the support $I(p_0)$ of p_0, whose rows Π_i are exchangeable, with $(\Pi_i, i \in I(p_0)) \mid p_0 \overset{\text{i.i.d.}}{\sim} \mathrm{DP}(\alpha p_0)$
- and $p_0 \sim \mathrm{DP}(\gamma F_0)$.

In other words, (X_n) is conditionally Markov and the prior on the rows of the random transition matrix is a *hierarchical Dirichlet process* (Teh et al. (2006)). Interesting developments of the infinite HMM, to encourage stronger state persistence, are discussed in Fox et al. (2011).

An area of research that seems still open is how to define natural priors for Bayesian inference on Markov chains with given properties. A direction of research is given by Diaconis and Rolles (2006), who introduce a conjugate prior for

the transition matrix of a *reversible* Markov chain, through a random walk with reinforcement on a graph. They show that the prior can be characterized in a predictive way, along the lines of Johnson's characterization of the Dirichlet conjugate prior (Zabell (1982, 1995)). Bacallado (2011) has recently extended Diaconis and Rolles construction, to the case of reversible Markov chains of order r, and to variable-order Markov chains.

4 Developments and final remarks

A problem widely studied in the recent years, for Bayesian nonparametric inference with dependence structures more complex than exchangeability, is the construction of dependent random measures. For example, consider partially exchangeable data, modeled as

$$\begin{aligned} Y_{j,1}, \ldots, Y_{j,n} \mid \theta_{j_1}, \ldots, \theta_{j,n} &\sim \prod_{i=1}^{n} f(y_{j,i} \mid \theta_{j,i}), \qquad j = 1, \ldots, k, \\ \theta_{j,1}, \theta_{j,2}, \ldots \mid G_j &\overset{\text{i.i.d.}}{\sim} G_j, \\ \mathbf{G} = (G_1, \ldots, G_k) &\sim \mu. \end{aligned}$$

Heterogeneity is modeled inside each of k groups, by allowing individual parameters $\theta_{j,i}, i \geq 1$ inside group j. Such parameters are regarded as a random sample from a group specific latent distribution G_j. Clustering inside group j can be modeled assuming a Dirichlet process prior for the random distribution G_j. To borrow strength across groups, it is desirable to assign a prior on the vector of random distributions $\mathbf{G} = (G_1, \ldots, G_k)$ such that the components G_j are dependent. Early work on this problem is due to Cifarelli and Regazzini (1978), who proposed a mixture of products of Dirichlet processes; that is, $G_j \mid \lambda \overset{\text{i.i.d.}}{\sim} \text{DP}(\alpha G_0(\cdot \mid \lambda))$, with $\lambda \sim H(\lambda)$. See Muliere and Petrone (1993) for an application in a regression context. A recent development is the *hierarchical Dirichlet Process* (Teh et al. (2006)), that assumes that $G_j \mid G_0 \overset{\text{i.i.d.}}{\sim} \text{DP}(\alpha G_0)$, where G_0 is itself a Dirichlet process. The clever choice of a Dirichlet process as the base measure G_0 allows to model common clusters across groups, since, because of the discrete nature of the Dirichlet process, all the random distributions G_j have the same support, given by the atoms of G_0. An important application of the hierarchical Dirichlet process as a prior in hidden Markov models has been reviewed in the previous section.

In a regression context, MacEachern (1999) and (2001) suggested a general construction of *dependent Dirichlet processes* (DDPs). Here, the groups are indexed by the values $x_1, \ldots, x_k$ of a covariate x. Assume that $(\theta_{x_1}, \ldots, \theta_{x_m}) \mid (G_{x_1}, \ldots, G_{x_m}) \sim \prod_i G_{x_i}$. McEachern suggested a general class of priors for $\mathbf{G} = (G_{x_1}, \ldots, G_{x_m})$, such that the G_{x_i} are dependent and marginally $G_{x_i} \sim \text{DP}$, by exploiting the stick-breaking discrete representation of the Dirichlet process.

Following MacEachern's idea, many ways for constructing DDP priors have been proposed in the recent literature. The so called *single-p* DDP assumes that $G_{x_i} = \sum_j p_j \delta_{\theta_j^*(x_i)}$, where the common weights (p_j) have a GEM(α) prior, and the atoms $(\theta_j^*(x_1), \ldots, \theta_j^*(x_k))$, $j = 1, 2, \ldots$, are i.i.d. from a joint distribution G_0 on $\mathbb{R}^k$, therefore inducing a dependence among the random distributions G_{x_i}. In other applications, for example, in the analysis of temporal or spatial data, the object of inference is a vector (or, more generally, a stochastic process) $\theta = (\theta_x, x \in \mathcal{X})$, where x denotes a temporal or spatial coordinate. In this case, one can assume that $\theta \mid \tilde{\mathbf{G}} \sim \tilde{\mathbf{G}}$, where $\tilde{\mathbf{G}}$ is a random probability measure on $\mathbb{R}^{\mathcal{X}}$. In particular, if $\tilde{\mathbf{G}} \sim \mathrm{DP}(\alpha \mathbf{G}_0)$, then the random marginals G_x of $\mathbf{G}$ have a single-p DDP prior. More general constructions of *multiple-p* DDPs have been proposed, by linking the stick breaking weights. We refer to Dunson (2010) for an excellent overview and references. Recent proposals aim at giving a general framework to model the dependence structure among the stick-breaking weights, and study general properties of the resulting priors (Barrientos, Jara and Quintana (2011)). Griffin and Steel (2011) propose a flexible class of time-dependent nonparametric priors for Bayesian nonparametric modelling of time series, whose marginals have a general stick-breaking form.

These prior constructions largely extend the availability of Bayesian nonparametric methods besides exchangeability, to highly complex depended data. However, the structure of the predictive rules are in general analytically complicated, and computations can consequently be highly demanding. A predictive approach leading to clearer predictive assumptions appears therefore appealing also in constructing dependent random measures. A bivariate Dirichlet process based on dependent urn schemes has been proposed by Walker and Muliere (2003); developments for time-dependent random measures are suggested by Caron, Davy and Doucet (2007). The central focus on prediction in the learning process has possibly been the basic reason of the impressive developments of Bayesian nonparametrics methods in the machine learning community in the recent years. The already quoted hierarchical Dirichlet process and the infinite HMM are just remarkable examples of the extremely active research in this field. A very fruitful construction of a predictive scheme for infinite latent features problems is the *Indian buffet* process (Griffiths and Ghahramani (2006)). Here, exchangeable objects or individuals are described through a potentially infinite array of features, resulting in an underlying random binary matrix with exchangeable rows and an unbounded number of columns. The Indian buffet characterizes the prior on this random matrix through the predictive rule, such as the Chinese restaurant characterizes the Dirichlet process. The de Finetti measure of the Indian buffet has been later found by Thibaux and Jordan (2007). Binary arrays with exchangeability structures are of interest in many contexts, such as in social network studies; see, for example, Roy and Teh (2009), who propose a *Mondrian process* as a prior on random binary matrices for Bayesian inference for relational data. These works relate with theory in Aldous

(1981). This is just to mention some interesting lines of developments. Our review is certainly far from being exhaustive. This is a very active area of research, and further fruitful interaction across theoretical, applied, statistical, machine learning, genetics literature can certainly be envisaged.

We have focussed on predictive characterizations of nonparametric priors, and did not discuss the inferential aspects, updating rules and computational issues. The recent volume by Hjort et al. (2010) provides a rich reference.

All the constructions that we have briefly reviewed in this note characterize priors that are a.s. discrete. A problem is whether it is possible to give a predictive characterization of nonparametric priors with support on a class of absolutely continuous distributions. Dirichlet process mixture models are commonly used as nonparametric priors for continuous data, for example, in density estimation problems. Petrone and Veronese (2010) discuss a general framework where Dirichlet process mixture models are interpreted as a smoothing of discrete nonparametric priors. However, the predictive structure of these models is complicated. Gaussian processes offer another powerful tool for Bayesian nonparametric inference; however, their predictive characterization does not seem to be fully explored. Predictive constructions of absolutely continuous nonparametric priors seems to remain an open problem.

Acknowledgments

We would like to thank Alexandra Schmidt for her helpful suggestions and support. This work has been partially supported by the Italian Ministry of University and Research, Grant 2008MK3AFZ, and by Bocconi University research grants.

References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis* **11**, 581–598. MR0637937

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2**, 1152–1174. MR0365969

Arellano-Valle, R. B. and Bolfarine, H. (1995). On some characterizations of the t-distribution. *Statistics and Probability Letters* **25**, 79–85. MR1364821

Arellano-Valle, R. B., Bolfarine, H. and Iglesias, P. L. (1994). A predictivistic interpretation to the multivariate t-distribution. *Test* **3**, 221–236. MR1365732

Bacallado, S. (2011). Bayesian analysis of variable-order, reversible Markov chains. *The Annals of Statistics* **39**, 838–864. MR2816340

Barrientos, A. F., Jara, A. and Quintana, F. (2011). On the support of MacEachern's dependent Dirichlet processes. Technical report, Dept. Statistics, Pontificia Universidad Católica de Chile.

Beal, M. J., Ghahramani, Z. and Rasmussen, C. E. (2002). The infinite hidden Markov model. In *Advances in Neural Information Processing Systems* **14**, 577–584. Cambridge, MA: MIT Press.

Berti, P. and Rigo, P. (1997). A Glivenko–Cantelli theorem for exchangeable random variables. *Statistics and Probability Letters* **32**, 385–391. MR1602215

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics* **1**, 353–355. MR0362614

Bulla, P., Muliere, P. and Walker, S. (2009). A Bayesian nonparametric estimator of a multivariate survival function. *Journal of Statistical Planning and Inference* **139**, 3639–3648. MR2549111

Caron, F., Davy, M. and Doucet, A. (2007). Generalized Pólya urn for time-varying Dirichlet process mixtures. In *23rd Conference on Uncertainty in Artificial Intelligence (UAI'2007)*, Canada.

Cifarelli, D. M. and Regazzini, E. (1978). Problemi statistici nonparametrici in condizioni di scambiabilità parziale e impiego di medie associative. *Quaderni Istituto di Matematica Finanziaria*, *Università di Torino Ser. III* **12**, 1–36. English translation available at http://www.unibocconi.it/wps/allegatiCTP/CR-Scamb-parz[1].20080528.135739.pdf.

Cifarelli, D. M. and Regazzini, E. (1996). De Finetti's contribution to probability and statistics. *Statistical Science* **11**, 253–282. MR1445983

Cirillo, P., Hüsler, J. and Muliere, P. (2010). A nonparametric urn-based approach to interacting failing systems with an application to credit risk modeling. *International Journal of Theoretical and Applied Finance* **13**, 1223–1240. MR2748505

Dawid, A. P. (1978). Extendebility of spherical matrix distributions. *Journal of Multivariate Analysis* **8**, 559–566. MR0520963

de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* **7**, 1–68. English translation: Foresight, its logical laws, its subjective sources. In *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, eds.) (1964) 97–156. New York: Wiley. MR1508036

de Finetti, B. (1959). La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista. In *Atti corso CIME su Induzione e Statistica Varenna* 1–115. Roma: Cremonese. English translation: *Probability, Induction and Statistics* (1972) 147–227. New York: Wiley.

Diaconis, P., Eaton, M. L. and Lauritzen, S. L. (1992). Finite De Finetti theorems in linear models and multivariate analysis. *Scandinavian Journal of Statistics* **19**, 289–315. MR1211786

Diaconis, P. and Freedman, D. (1980). de Finetti's theorem for Markov chains. *The Annals of Probability* **8**, 115–130. MR0556418

Diaconis, P. and Freedman, D. (1984). Partial exchangeability and sufficiency. In *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics. Applications and New Directions* (J. K. Gosh and J. Roy, eds.) 205–236. Calcutta: Indian Statistical Instutute. MR0786142

Diaconis, P. and Rolles, S. W. W. (2006). Bayesian analysis for reversible Markov Chains. *The Annals of Statistics* **34**, 1270–1292. MR2278358

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics* **7**, 269–281. MR0520238

Doksum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability* **2**, 183–201. MR0373081

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics: Principles and Practice* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.). Cambridge, UK: Cambridge Univ. Press. MR2730665

Eaton, M. L., Fortini, S. and Regazzini, E. (1993). Spherical symmetry: An elementary justification. *Journal of the Italian Statistical Society* **2**, 1–16.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588. MR1340510

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112. MR0325177

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230. MR0350949

Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12**, 42–58.

Fortini, S., Ladelli, L. and Regazzini, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhya, Ser. A* **62**, 86–109. MR1769738

Fortini, S. and Petrone, S. (2011). Hierarchical reinforced urn processes. Technical report, Bocconi Univ., Milano.

Fortini, S., Ladelli, L., Petris, G. and Regazzini, E. (2002). On mixtures of distributions of Markov chains. *Stochastic Processes and Their Applications* **100**, 147–165. MR1919611

Freedman, D. (1963). Invariants under mixing that generalize de Finetti's theorem: Continuous time parameter. *The Annals of Mathematical Statistics* **34**, 1194–1216. MR0189111

Fox, E. B., Sudderth, E. B., Jordan, M. I. and Willsky, A. S. (2011). A Sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics* **5**, 1020–1056.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. New York: Springer-Verlag. MR1992245

Gnedin, A. V. and Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **325**, 83–102. MR2160320

Griffin, J. E. and Steel, M. (2011). Stick-breaking autoregressive processes. *Journal of Econometrics* **162**, 383–396.

Griffiths, T. L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian Buffet Process. In *Advances in Neural Information Processing Systems (NIPS)* (Y. Weiss, B. Schölkopf and J. Platt, eds.), 475–482. Cambridge, MA: MIT Press.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta process in models for life history data. *The Annals of Statistics* **18**, 1259–1294. MR1062708

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge, UK: Cambridge Univ. Press. MR2722987

Hoppe, F. M. (1984). Pólya-like urns and the Ewens's sampling formula. *Journal Mathematical Biology* **20**, 91–94. MR0758915

Iglesias, P. L., Loschi, R. H., Pereira, C. A. B. and Wechsler, S. (2009). A note on extendibility and predictivistic inference in finite populations. *Brazilian Journal of Probability and Statistics* **23**, 216–226. MR2575434

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173. MR1952729

Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. New York: Springer. MR2161313

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics* **21**, 59–78. MR0210185

Kingman, J. F. C. (1972). On random sequences with spherical symmetry. *Biometrika* **59**, 492–493. MR0343420

Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society, Ser. B* **37**, 1–22. MR0368264

Kingman, J. F. C. (1978). The representation of partition structures. *Journal of the London Mathematical Society* **18**, 374–380. MR0509954

Lijoi, A., Mena, R. H. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *Journal of the American Statistical Association* **100**, 1278–1291. MR2236441

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.) 80–136. Cambridge, UK: Cambridge Univ. Press. MR2730661

Loschi, R. H., Iglesias, P. I. and Arellano-Valle, R. B. (2003). Predictivistic characterizations of multivariate student-t models. *Journal of Multivariate Analysis* **85**, 10–23. MR1978174

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* **23**, 727–741. MR1293996

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

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. Crete: International Society for Bayesian Analysis.

McCloskey, J. W. (1965). A model for the distribution of individuals by species in an environment. Ph.D. thesis, Michigan State Univ. MR2615013

Muliere, P. and Petrone, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: Parametric and nonparametric models. *Journal of the Italian Statistical Society* **2**, 349–364.

Muliere, P., Secchi, P. and Walker, S. G. (2000). Urn schemes and reinforced random walks. *Stochastic Processes and Their Applications* **88**, 59–78. MR1761692

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95–110. MR2082149

Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92**, 21–39. MR1156448

Petrone, S., Guindani, M. and Gelfand, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society, Ser. B* **71**, 755–782. MR2750094

Petrone, S. and Raftery, A. E. (1997). A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics and Probability Letters* **36**, 69–83. MR1491076

Petrone, S. and Veronese, P. (2010). Feller operators and mixture priors in Bayesian nonparametrics. *Statistica Sinica* **20**, 379–404. MR2640700

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158. MR1337249

Pitman, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory* (T. S. Ferguson et al., eds.). *Lecture Notes—Monograph Series* **30**, 245–267. Hayward, CA: IMS. MR1481784

Pitman, J. (2003). Poisson–Kingman partitions. In: *Statistics and Science: A Festschrift for Terry Speed* (D. R. Goldstein, ed.). *IMS Lecture Notes—Monograph Series* **40**, 1–34. Beachwood: IMS. MR2004330

Pitman, J. and Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900. MR1434129

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Ser. B* **65**, 557–574. MR1983764

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics* **31**, 560–585. MR1983542

Roy, D. M. and Teh, Y. W. (2009). The Mondrian Process. In *Advances in Neural Information Processing Systems (NIPS)* (D. Koller, Y. Bengio, D. Schuurmans, L. Bottou and A. Culotta, eds.) **21**, 1377–1384. NIPS.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650. MR1309433

Smith, A. F. M. (1981). On random sequences with centered spherical symmetry. *Journal of the Royal Statistics Society, Ser. B* **43**, 208–209. MR0626767

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications In *Bayesian Nonparametrics* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.) 158–207. Cambridge, UK: Cambridge Univ. Press. MR2730663

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581. MR2279480

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Journal of Machine Learning Research—Proceedings Track* **2** 564–571. Proceedings of AISTATS, San Juan, Puerto Rico, 2007.

Walker, S. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *The Annals of Statistics* **25**, 1762–1780. MR1463574

Walker, S. G. and Muliere, P. (2003). A bivariate Dirichlet process. *Statistics and Probability Letters* **64**, 1–7. MR1995803

Zabell, S. L. (1982). W. E. Johnson's "sufficientness" postulate. *The Annals of Statistics* **10**, 1090–1099. MR0673645

Zabell, S. L. (1995). Characterizing Markov exchangeable sequences. *Journal of Theoretical Probability* **8**, 175–178. MR1308676

Department of Decision Sciences
Bocconi University
Via Roentgen 1
20136 Milano
Italy
E-mail: sandra.fortini@unibocconi.it
sonia.petrone@unibocconi.it