Bayesian Inference Using an Extension of the Dirichlet Process

Nils Lid Hjort
University of Oslo

Andrea Ongaro
Università di Milano-Bicocca

– October 2003 –

Abstract

A family of nonparametric prior distributions which extends the Dirichlet process is introduced and studied. Such family is first constructed by normalising suitable compound Poisson processes. An alternative derivation shows that such priors admit a simple representation as discrete random probability measures with symmetric Dirichlet weights independent of i.i.d. locations. The latter representation proves useful in deriving manageable expressions for the posterior and predictive distributions. A number of Bayesian nonparametric estimators based on the family are discussed. Furthermore, an analysis of the characteristics of a sample drawn from the family demonstrates its potential as a second stage prior in hierarchical Bayesian clustering models.

1. Background and introduction

The Dirichlet process ([3], [4]) still plays a central role in Bayesian nonparametric statistics, both as a special case of many larger families of priors on distributions for the data and as element of more complex hierarchical models. See [10] and [6] for recent reviews.

Our aims in this note are to exhibit a new useful generalisation of the Dirichlet process and then derive and discuss various results and nonparametric Bayesian inference methods based on this generalisation. In particular we shall point out the effective gain reached in terms of flexibility by such generalisation, showing at the same time its considerable tractability.

We start out giving the essentials of the Gamma and Dirichlet processes, also establishing some notation.

A Gamma-distributed Y with parameters (a, b) has density proportional to $y^{\alpha-1}e^{-by}$ and its mean and variance are a/b and a/b^2. Consider independent $Y_1, \ldots, Y_n$ where $Y_i \sim \text{Gamma}(a_i, b)$. Then $S = \sum_{i=1}^m Y_i$ is Gamma distributed with parameters $(\sum_{i=1}^m a_i, b)$, and furthermore the random probability vector $(Y_1/S, \ldots, Y_m/S)$ is Dirichlet with parameter $(a_1, \ldots, a_m)$. These are the keys to the existence of (and constructions of) Gamma and Dirichlet processes over any interval, or in fact any measurable space. At this stage we confine ourselves to the unit interval $[0, 1]$. Let F_0 be any distribution function on $[0, 1]$ and let a and b positive. We then say thet Z is a Gamma process parameters (aF_0, b) if its increments are independent and of the form $\mathrm{d}Z(t) \sim \text{Gamma}(a\mathrm{d}F_0(t), b)$. In particular $Z(t) \sim \text{Gamma}(aF_0(t), b)$ with

$$\mathrm{E}Z(t) = aF_0(t)/b \qquad \text{and} \qquad \text{Var}\, Z(t) = aF_0(t)/b^2. \tag{1.1}$$

The random distribution function $F(t) = Z(t)/Z(1)$ over $[0, 1]$ becomes a Dirichlet process with parameter aF_0 (the scale parameter b becoming immaterial for F). Its first moments are

$$\mathrm{E}F(t) = F_0(t) \qquad \text{and} \qquad \text{Var}\, F(t) = (a+1)^{-1}F_0(t)(1 - F_0(t)). \tag{1.2}$$

The contents of the paper are as follows. In Section 2 we introduce a family of compound Poisson processes on $[0, 1]$ and show that the Gamma process can be reached as limiting case. This motivates (Section 3.1) the definition of a generalised Dirichlet process (GDP) as normalisation of such compound processes. Section 3.2 presents a different representation which extends the GDP to an arbitrary sample space. Such a representation is in terms of a random number of symmetric Dirichlet distributed weights attached to i.i.d. locations. It is used in Section 4 to derive the marginal distribution for data and the posterior process.

Estimators of mean and variance parameters including the predictive distribution are discussed in Section 5. Properties of a sample drawn from the GDP, with particular emphasis on the structure of possible configurations of ties in the data, are studied in Section 6. The derived results are shown to imply an added flexibility of the GDP family with respect to the Dirichlet process, which is of particular relevance in applications to hierarchical settings used to model clusterings of items.

2. The Gamma process as a limit of compound Poisson processes

Let $M(\cdot)$ be a Poisson process with parameter $\lambda F_0(\cdot)$. Consider the compound process

$$Z_\lambda(t) = \sum_{i=1}^{M(t)} G_i \qquad \text{for } 0 \le t \le 1, \tag{2.1}$$

where the G_i's are independent Gamma$(a/\lambda, b)$ variables and independent of M. Note that although Z_λ forms a Gamma process conditional on any path of M, it is not a Gamma process marginally. We have

$$\mathrm{E}Z_\lambda(t) = aF_0(t)/b \qquad \text{and} \qquad \mathrm{Var}\, Z_\lambda(t) = aF_0(t)/b^2[1 + a/\lambda],$$

showing in comparison with (1.1) that Z_λ like Z is centred at aF_0/b, but exhibiting somewhat more variation than that of a pure Gamma process. However, the Gamma process structure is reached as λ grows.

LEMMA 1. *Consider the compound Poisson process Z_λ of (2.1), governed by parameters λF_0 for M and $(a/\lambda, b)$ for the Gamma summands G_i. Then, as λ grows to infinity, the Z_λ process converges in distribution to that of a pure gamma process Z with parameters (aF_0, b).*

PROOF: The Laplace transform of $Z_\lambda(t)$ is in general terms

$$\mathrm{E}\exp\{-\theta Z_\lambda(t)\} = \mathrm{E}L_0(\theta)^{M(t)} = \exp\{-\lambda F_0(t)(1 - L_0(\theta))\},$$

with $L_0(\theta) = \mathrm{E}\exp(-\theta G_i)$ being the Laplace transform of the G_is. In the present case, $L_0(\theta) = \exp\{-(a/\lambda)\log(1 + \theta/b)\}$, which is seen to lead to

$$\mathrm{E}\exp\{-\theta Z_\lambda(t)\} \to \exp\{-aF_0(t)\log(1 + \theta/b)\} \qquad \text{as } \lambda \to \infty.$$

This means that $Z_\lambda(t) \stackrel{d}{\to} Z(t)$. In the same way one shows that a finite set of increments of Z_λ converges in distribution to the corresponding set of increments of Z, basically using that both have independent increments. Tightness of the Z_λ system follows from monotonicity. This secures convergence of Z_λ to Z in the space of all right-continuous functions on $[0, 1]$ with left-hand limits, equipped with the Skorohod topology. $\square$

3. Generalised Dirichlet processes

The above result motivates the study of a random distribution function more general than the Dirichlet process. A supplementary representation then leads to a further generalisation.

3.1. The normalised compound Poisson process

Consider

$$F_\lambda(t) = \frac{Z_\lambda(t)}{Z_\lambda(1)} = \frac{\sum_{i \le M(t)} G_i}{\sum_{i \le M(1)} G_i} \qquad \text{on } [0,1]. \tag{3.1}$$

This necessarily constitutes a generalisation of the Dirichlet process, since this results from sending λ to infinity by the Lemma above. Note that the scale parameter b becomes immaterial for the ratio process. We therefore simply take $b = 1$, and use G_i's which are Gamma$(a/\lambda, 1)$. The case of having no Poisson events at all, i.e. $M(1) = 0$, has probability $\exp(-\lambda)$, which will be very small in our intended applications. Nevertheless, to avoid misspecifications in formula (3.1) we choose to condition the Poisson process $M(\cdot)$ on the event $M(1) > 0$. We are then led to the following definition.

DEFINITION 1. *Let G_i's be independent* Gamma$(a/\lambda, 1)$ *variables, independent of the process $M'(\cdot)$, which is distributed as $M(\cdot) \,|\, M(1) > 0$, where $M(\cdot)$ is a Poisson process with parameter λF_0. Then the process*

$$F_\lambda(t) = \frac{\sum_{i=1}^{M'(t)} G_i}{\sum_{i=1}^{M'(1)} G_i} \qquad \text{on } [0,1] \tag{3.2}$$

is called a GENERALISED DIRICHLET PROCESS (GDP) *with parameters (a, F_0, λ).*

We first derive expressions for the mean and the variance of the GDP, to compare with the well-known formulae (1.2) for the Dirichlet (a, F_0). In the following we write M_1 for the truncated Poisson (λ)-distributed variable $M'(1)$.

PROPOSITION 1. *For the generalised Dirichlet process $F_\lambda(t)$, one has*

$$\mathrm{E}F_\lambda(t) = F_0(t) \qquad \text{and} \qquad \mathrm{Var}\, F_\lambda(t) = \mathrm{E}\left[\frac{1 + a/\lambda}{1 + aM_1/\lambda}\right] F_0(t)(1 - F_0(t)). \tag{3.3}$$

PROOF: Conditional on the full path of $M'(\cdot)$, $Z_\lambda(\cdot)$ is a Gamma process, which means that $F_\lambda(\cdot)$ becomes an ordinary Dirichlet. Therefore,

$$\mathrm{E}[F_\lambda(t) \,|\, M'] = \frac{M'(t)a/\lambda}{M_1 a/\lambda} = \frac{M'(t)}{M_1}$$

and

$$\mathrm{Var}(F_\lambda(t) \,|\, M') = \frac{1}{M_1 a/\lambda + 1} \frac{M'(t)}{M_1} \left(1 - \frac{M'(t)}{M_1}\right).$$

Next observe that $M'(t)$ conditional on M_1 is a binomial $(M_1, F_0(t))$. It follows that $\mathrm{E}[M'(t)/M_1 \mid M_1] = F_0(t)$, leading to $\mathrm{E}F_\lambda(t) = F_0(t)$.

The variance of $F_\lambda(t)$ can be written as $I+II$, where the first term is $\mathrm{E}\,\mathrm{Var}(F_\lambda(t) \mid M')$ and the second is $\mathrm{Var}\,\mathrm{E}[F_\lambda(t) \mid M']$. Via conditioning on M_1, one finds

$$I = \mathrm{E}\frac{1}{M_1 a/\lambda + 1}\frac{M_1 - 1}{M_1}F_0(t)(1 - F_0(t)),$$

while $II = \mathrm{Var}\,M'(t)/M_1$. This is again computed via conditioning on M_1, and the result is $\mathrm{E}1/M_1 F_0(t)(1 - F_0(t))$. Adding I and II gives the factor

$$\mathrm{E}\left[\frac{1}{1 + aM_1/\lambda}\left(1 - \frac{1}{M_1}\right) + \frac{1}{M_1}\right] = \mathrm{E}\frac{1 + a/\lambda}{1 + aM_1/\lambda}$$

times $F_0(t)(1 - F_0(t))$, proving the variance claim. □

REMARK 1. The variable M_1/λ has

$$\text{mean } \frac{1}{1 - \exp(-\lambda)} \quad \text{and} \quad \text{variance } \frac{1}{\lambda(1 - \exp(-\lambda))} - \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))^2}$$

and so goes to 1 as λ increases. The variance factor $k_\lambda = (1 + a/\lambda)\mathrm{E}1/(1 + aM_1/\lambda)$ can be shown, by using Jensen inequality, to be bigger than the $1/(1+a)$ factor appearing in (1.2), which is its limit as λ grows. This agrees with the previously observed feature of the generalised Gamma process being somewhat more variable than the pure Gamma process. It is easy to compute k_λ for given λ and a via the Poisson probabilities. And a Taylor approximation gives

$$k_\lambda = (1 + a/\lambda)\,\mathrm{E}\frac{1}{1 + aM_1/\lambda} \doteq \frac{1}{1 + a}\left(1 + \frac{a}{\lambda}\right)\left(1 + \frac{a^2}{(1 + a)^2\lambda}\right).$$

REMARK 2. It is clear from Definition 1 that, conditionally on M', the F_λ forms a Dirichlet process (with a finite rather than an infinite number of jumps). So the distribution of $F_\lambda(t)$ is a Beta conditional on M', but not marginally. Specifically,

$$\Pr\{F_\lambda(t) \le y\} = \mathrm{E}\Pr\{\mathrm{Beta}\,(M'(t)a/\lambda, (M_1 - M'(t))a/\lambda) \le y\}, \tag{3.4}$$

where $M'(t)$ and M_1 are as in Definition 1. This may again be alternatively computed via conditioning on M_1, involving a binomial distribution for $M'(t)$. Thus there are representations of the (3.4) distribution in terms of infinite sums, easily computed for given (a, F_0, λ) and t. It is also easy to evaluate by simulation.

3.2. An alternative representation

Another way of understanding the generalised Dirichlet process is as follows. The random probability measure works by distributing random probabilities $\beta_i = G_i/G$ to random positions $\xi_{(i)}$ for $i = 1, \ldots, M_1$, writing $G = \sum_{i \leq M_1} G_i$, where $\xi_{(1)} < \cdots < \xi_{(M_1)}$ are the locations for events chosen by the process M', which is distributed as $M \,|\, \{M(1) > 0\}$, where M is a Poisson process λF_0. By a well-known result the locations of the events of the Poisson process behave as the ordering of a sample $\xi_1, \ldots, \xi_{M_1}$ chosen independently from the distribution F_0. It is easy then to see that this property continues to hold for the process M'.

We may now rearrange terms in $\sum_{i=1}^{M_1} \beta_i \delta_{\xi_{(i)}}$ (where δ_ξ denotes the probability measure concentrated in position ξ), using the symmetry of the distribution for $(\beta_1, \ldots, \beta_{M_1})$. Hence F_λ admits the representation

$$F_\lambda(A) = \sum_{i=1}^{M_1} \beta_i \delta_{\xi_i}(A), \tag{3.5}$$

where the ξ_i's are i.i.d. from F_0 and independent of M_1 and of the β_i's, which form, conditionally on M, a Dirichlet distribution with M_1 components and parameters $(a/\lambda, \ldots, a/\lambda)$, hereafter denoted by $\mathrm{Dir}_{M_1}(a/\lambda, \ldots, a/\lambda)$. A nice aspect of the (3.5) representation is that only the truncated Poisson distributed variable M_1 is involved, not the full path of the process M'.

Here $F_\lambda(A)$ is the same as $P_\lambda(A)$, where P_λ is the probability measure determined by the cumulative distribution function F_λ. Some comments on expression (3.5) follow.

REMARK 3. Definition 1 is restricted to consider the unit interval as sample space. On the contrary, representation (3.5) is clearly well defined for an arbitrary sample space, extending therefore the definition of the GDP.

In the following, we shall indicate with P_λ the extended GDP defined on an arbitrary sample space $(\mathcal{X}, \mathcal{A})$ and with P_0 the common distributions of the ξ_i's.

REMARK 4. Expression (3.5) displays a particularly transparent structure which makes the process both easy to interpret and analytically tractable. In particular, it provides a clear understanding of the characteristics – total number of points, their locations and weights – of the discrete random probability measure defined by the GDP. Furthermore, it allows the derivation of relevant quantities, like the marginal distribution of the data and the posterior distribution of the process.

A key role in such representation is played by the exchangeable Dirichlet distribution of the random weights β_i's, which makes them very tractable. In contrast, the

distribution of the weights in the Sethuraman [9] representation of the Dirichlet process looks somewhat more involved:

$$P(A) = \sum_{i=1}^{\infty} \beta_i' \delta_{\xi_i}(A), \tag{3.6}$$

where $\beta_i' = \theta_i \prod_{j<i}(1-\theta_j)$ and θ_j are independent from a Beta distribution with parameter $(1, a)$.

REMARK 5. Representation (3.5) makes clear that the GDP is a special case of a fairly broad class of discrete random probability measures considered in [7], where a number of properties of the class are derived. In particular, results on the support of the members of the class are given. From these results it is immediate to see that the GDP maintains the same large support of the Dirichlet process, so that it can effectively be considered a genuine nonparametric prior. More precisely, one can prove that, under the topology of pointwise convergence, the support of the GDP is formed by all the probability measures absolutely continuous with respect to P_0. If one instead considers the topology of convergence in distribution, then the support is given by all the probability measures whose support is contained in the support of P_0.

REMARK 6. The GDP process is fairly simple to simulate: by (3.5) it suffices to generate random variates from P_0 and from the Poisson and the Gamma distribution.

4. Marginals distributions and the posterior process

Suppose $X_1, \ldots, X_n$ are independent observations chosen from the randomly selected P_λ. In other words,

$$\Pr\{X_1 \in A_1, \ldots, X_n \in A_n \,|\, P_\lambda\} = P_\lambda(A_1) \cdots P_\lambda(A_n)$$

for all measurable sets $A_1, \ldots, A_n$. We wish to study their marginal distributions and the distribution of P_λ given the data.

4.1. Marginal distributions for data

Their marginal distribution can be expressed as the mean of the above expression over the distribution of P_λ, and, in particular,

$$\Pr\{X_1 \in A\} = \mathrm{E} P_\lambda(A) = P_0(A), \tag{4.1}$$

by arguments of Proposition 1 of Section 3. Thus, the so-called predictive distribution of a single X_i is the base measure P_0 itself.

The simultaneous marginal distribution of two or more data points is more cumbersome. We illustrate with $n = 2$ and $n = 3$. First, using (3.5), one finds

$$\begin{aligned}
&\Pr\{X_1 \in A_1, X_2 \in A_2 \,|\, M_1 = m\} \\
&= \mathrm{E} \sum_{i=1}^{m} \sum_{j=1}^{m} \beta_i \beta_j I\{\xi_i \in A_1, \xi_j \in A_2\} \\
&= \sum_{i,j} \mathrm{E}\beta_i \beta_j \Pr\{\xi_i \in A_1, \xi_j \in A_2\} \\
&= \sum_{i} \frac{\tau(\tau+1)}{m\tau(m\tau+1)} P_0(A_1 \cap A_2) + \sum_{i \neq j} \frac{\tau^2}{m\tau(m\tau+1)} P_0(A_1) P_0(A_2),
\end{aligned}$$

writing $\tau = a/\lambda$. This implies

$$\Pr\{X_1 \in A_1, X_2 \in A_2\} = k_\lambda P_0(A_1 \cap A_2) + (1 - k_\lambda) P_0(A_1) P_0(A_2),$$

where k_λ is defined as in Remark 1. A consequence of this is that given $X_1 = x_1$, then $X_2 = x_1$ with probability k_λ while with remaining probability $1 - k_\lambda$ is drawn from P_0.

Similarly, but requiring more algebraic work, the $n = 3$ case can be tended to. One needs to sort triples of indexes (i, j, k) into those with three different elements, those with two, and those with all indexes equal. The result is

$$\begin{aligned}
&\Pr\{X_1 \in A_1, X_2 \in A_2, X_3 \in A_3 \,|\, M_1 = m\} \\
&= m \frac{\tau(\tau+1)(\tau+2)}{m\tau(m\tau+1)(m\tau+2)} P_0(A_1 \cap A_2 \cap A_3) \\
&\quad + m(m-1) \frac{\tau^2(\tau+1)}{m\tau(m\tau+1)(m\tau+2)} \{P_0(A_1 \cap A_2) P_0(A_3) \\
&\qquad + P_0(A_1 \cap A_3) P_0(A_2) + P_0(A_2 \cap A_3) P_0(A_1)\} \\
&\qquad + m(m-1)(m-2) \frac{\tau^3}{m\tau(m\tau+1)(m\tau+2)} P_0(A_1) P_0(A_2) P_0(A_3).
\end{aligned}$$

Summing over all m gives a distribution of a mixture type, allowing a certain probability for three distinct data points from P_0, and other probabilities for various configuration of ties.

4.2. The posterior process

To carry out Bayesian inference, we need the posterior distribution of P_λ given a sample $X_1, \ldots, X_n$. In the following, this posterior distribution will be derived starting from representation (3.5) along the lines of [7].

The GDP does not form a conjugate class of prior distributions, as it can be easily verified. The next theorem establishes the structure of a larger conjugate class of priors.

THEOREM. *Consider the following random probability measure P defined on the arbitrary sample space $(\mathcal{X}, \mathcal{A})$:*

$$P(A) = \sum_{i=1}^{r} \gamma_i \delta_{x_i}(A) + \sum_{j=1}^{M} \beta_j \delta_{\xi_j}(A) \qquad A \in \mathcal{A}, \tag{4.2}$$

where $x_i \in \mathcal{X}$, $i = 1, \ldots, r$, are fixed distinct constants, $r \geq 0$ is a fixed integer, $M \geq 0$ is an integer valued random variable. Moreover,

$$(\gamma_1, \ldots, \gamma_r, \beta_1, \ldots, \beta_M) \,|\, M \sim Dir_{r+M}\left(k_1, \ldots, k_r, \frac{a}{\lambda}, \ldots, \frac{a}{\lambda}\right),$$

the ξ_i's are independent from P_0 and independent of M, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$ and P_0 is a diffuse probability measure (i.e. it gives zero probability to each single point) on $(\mathcal{X}, \mathcal{A})$. Furthermore, assume that $M + r \geq 1$ with probability one. Let $p(m)$ be the probability distribution of M and let X be an observation from P. Then, the posterior distribution of the process has the following form.

1. *If $x \neq x_i$, $i = 1, \ldots, r$, then $P \,|\, \{X = x\}$ is distributed as*

$$\sum_{i=1}^{r} \gamma_i' \delta_{x_i} + \gamma_{r+1}' \delta_x + \sum_{j=1}^{M'} \beta_j' \delta_{\xi_j}$$

where

$$\Pr\{M' = m\} \propto p(m+1) \frac{m+1}{(m+1)\frac{a}{\lambda} + \sum_{i=1}^{r} k_i} \qquad m = 0, 1, \ldots$$

$$(\gamma_1', \ldots, \gamma_r', \beta_1', \ldots, \beta_{M'}') \,|\, M' \sim \mathrm{Dir}_{r+1+M'}\left(k_1, \ldots, k_r, 1 + \frac{a}{\lambda}, \frac{a}{\lambda}, \ldots, \frac{a}{\lambda}\right)$$

and the ξ_i's are independent from P_0 and independent of M', $\boldsymbol{\gamma}'$, $\boldsymbol{\beta}'$.

2. *If $x = x_\ell$ for some integer $1 \leq \ell \leq r$, then $P \,|\, \{X = x\}$ is distributed as*

$$\sum_{i=1}^{r} \gamma_i'' \delta_{x_i} + \sum_{j=1}^{M''} \beta_j'' \delta_{\xi_j}$$

where

$$\Pr\{M'' = m\} \propto p(m) \frac{1}{m\frac{a}{\lambda} + \sum_{i=1}^{r} k_i} \qquad m = 0, 1, \ldots$$

$$(\gamma_1'', \ldots, \gamma_r'', \beta_1'', \ldots, \beta_{M''}'') \,|\, M'' \sim \mathrm{Dir}_{r+M''}\left(k_1, \ldots, k_\ell + 1, \ldots, k_r, \frac{a}{\lambda}, \ldots, \frac{a}{\lambda}\right)$$

and the ξ_i's are independent from P_0 and independent of M'', $\boldsymbol{\gamma}''$, $\boldsymbol{\beta}''$.

It follows that the class of prior distributions defined in the theorem is a conjugate class. As a technical point, notice that case 1) must be considered only if M is not degenerate at zero, since otherwise $\Pr\{X \neq x_i,\, i = 1, \ldots, r\} = 0$. The proof of the theorem requires two preliminary lemmas. In the first lemma we shall show how an observation X from P can be expressed in terms of the random elements which define P.

LEMMA 2. *Let* $\boldsymbol{\alpha} = (\boldsymbol{\gamma}, \boldsymbol{\beta}) = (\gamma_1, \ldots, \gamma_r, \beta_1, \ldots, \beta_M)$ *and* $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_M)$. *Let us define the random variable* T_i *for* $i = 1, \ldots, r + M$ *as*

$$T_i = \begin{cases} x_i & 1 \leq i \leq r \\ \xi_{i-r} & r+1 \leq i \leq r+M. \end{cases}$$

Let us then introduce a further integer valued random variable I *such that*

$$\Pr\{I = i \,|\, M, \boldsymbol{\alpha}, \boldsymbol{\xi}\} = \begin{cases} \alpha_i & 1 \leq i \leq r+M \\ 0 & i > r+M. \end{cases}$$

Then T_I *is an observation from* P, *that is the distribution of* T_I *given* P *is* P.

PROOF. Let us compute the conditional distribution of T_I given P:

$$\begin{aligned} \Pr\{T_I \in A \,|\, P\} &= \Pr\{T_I \in A \,|\, M, \boldsymbol{\alpha}, \boldsymbol{\xi}\} \\ &= \sum_{i=1}^{r+M} \Pr\{T_I \in A \,|\, M, \boldsymbol{\alpha}, \boldsymbol{\xi}, I = i\} \Pr\{I = i \,|\, M, \boldsymbol{\alpha}, \boldsymbol{\xi}\} \\ &= \sum_{i=1}^{r+M} \delta_{T_i}(A)\alpha_i = P(A). \end{aligned}$$

This is what was needed to prove. □

LEMMA 3. *The conditional distribution of* $M, \boldsymbol{\alpha}, \boldsymbol{\xi}$ *given* $T_I = x$, *where* $\boldsymbol{\alpha}, \boldsymbol{\xi}, T_I$ *are defined as in Lemma 2, is given by*

1. *if* $x \neq x_i$, $i = 1, \ldots, r$, *then*

$$\Pr\{M = m \,|\, T_I = x\} \propto p(m) \frac{m}{ma/\lambda + \sum_{i=1}^{r} k_i}$$

and

$$\boldsymbol{\alpha}, \boldsymbol{\xi} \,|\, \{T_I = x, M = m\} \sim \boldsymbol{\alpha}'_L, \boldsymbol{\xi}'_L,$$

where L *is an integer valued random variable such that*

$$\Pr\{L = \ell\} = \frac{1}{m} \quad \text{for } \ell = 1, \ldots, m,$$

$\boldsymbol{\alpha}'_L, \boldsymbol{\xi}'_L$ *are conditionally independent given* $L = \ell$ *with*

$$\boldsymbol{\alpha}'_L \,|\, L = \ell \sim \mathrm{Dir}_{r+m}\left(k_1, \ldots, k_r, \frac{a}{\lambda}, \ldots, \frac{a}{\lambda}, \frac{a}{\lambda} + 1, \frac{a}{\lambda}, \ldots, \frac{a}{\lambda}\right),$$

where the value $\frac{a}{\lambda} + 1$ *occurs in position* $r + \ell$ *and* $\boldsymbol{\xi}'_L \,|\, \{L = \ell\}$ *is a vector of* m *independent random variables whose* ℓ*th component is degenerate at* x *while the others are distributed as* P_0*;*

2. *if* $x = x_\ell$ *for some* $1 \le \ell \le r$*, then*

$$\Pr\{M = m \,|\, T_I = x\} \propto \frac{p(m)}{ma/\lambda + \sum_{i=1}^r k_i},$$

$$\boldsymbol{\alpha} \,|\, \{T_I = x, M = m\} \sim \mathrm{Dir}_{r+m}\left(k_1, \ldots, k_\ell + 1, \ldots, k_r, \frac{a}{\lambda}, \ldots, \frac{a}{\lambda}\right)$$

and $\boldsymbol{\xi} \,|\, \{\boldsymbol{\alpha}, T_I = x, M = m\}$ *is a vector of* m *independent random variables distributed as* P_0*.*

Notice that we need to consider case 1) only if M is not degenerate at zero.

PROOF. Let us consider first the distribution of $M \,|\, \{T_I = x\}$. It is enough to show that $\Pr\{M = m \,|\, T_I = x\}$ as defined in the Lemma satisfies the equality

$$\int_A \Pr\{M = m \,|\, T_I = x\} \,\mathrm{d}P_{T_I}(x) = \Pr\{T_I \in A, M = m\} \tag{4.3}$$

where $A \in \mathcal{A}$ and P_{T_I} is the marginal distribution of T_I.

The right hand side (rhs) of (4.3) is equal to

$$\begin{aligned}\Pr\{T_I \in A \,|\, M = m\}p(m) &= \mathrm{E}[P(A) \,|\, M = m]\, p(m) \\ &= p(m)\left(\frac{ma/\lambda}{ma/\lambda + \sum k_i} P_0(A) + \sum_{i=1}^r \frac{k_i}{m\frac{a}{\lambda} + \sum k_i}\delta_{x_i}(A)\right).\end{aligned}$$

The left hand side (lhs) can be written as

$$\int_{A\setminus\{x_1,\ldots,x_r\}} \Pr\{M = m \,|\, T_I = x\} c \,\mathrm{d}P_0(x) + \sum_{x_i \in A} \Pr\{M = m \,|\, T_I = x_i\} \frac{k_i}{\sum k_i}(1-c), \tag{4.4}$$

where $c = \mathrm{E}[M(a/\lambda)/(M(a/\lambda) + \sum k_i)$. Expression (4.4) is easily seen to be equal to the rhs of (4.3).

Let us consider now the conditional distribution of $\boldsymbol{\alpha}, \boldsymbol{\xi}$ given T_I and M. The results stated in the Lemma is proven once we show that the following equality holds

$$\begin{aligned}&\int_A \Pr\{\boldsymbol{\alpha} \in B, \xi_j \in A_j,\, j = 1, \ldots, M \,|\, T_I = x, M = m\} \,\mathrm{d}P_{T_I \,|\, M=m}(x) \\ &\quad = \Pr\{\boldsymbol{\alpha} \in B, \xi_j \in A_j,\, j = 1, \ldots, M, T_I \in A \,|\, M = m\}\end{aligned} \tag{4.5}$$

where $P_{T_I\,|\,M=m}$ is the conditional distribution of T_I given $M=m$ and B, A, A_j are suitable measurable sets. Let us compute first the rhs of (4.5). We have:

$$\begin{aligned}\Pr\{T_I \in A, I=i \,|\, \boldsymbol{\alpha},\boldsymbol{\xi}, M=m\} &= \Pr\{T_I \in A \,|\, I=i,\boldsymbol{\alpha},\boldsymbol{\xi},M=m\} \\ &\qquad \Pr\{I=i\,|\,\boldsymbol{\alpha},\boldsymbol{\xi},M=m\} \\ &= \delta_{T_i}(A)\alpha_i \quad \text{for } i=1,\ldots,r+m.\end{aligned}$$

The rhs is therefore equal to

$$\begin{aligned}&\sum_{i=1}^{r+m} \Pr\{\boldsymbol{\alpha}\in B, \xi_j\in A_j,\, j=1,\ldots,M, T_I\in A, I=i\,|\,M=m\} \\ &\quad= \sum_{i=1}^{r+m} \mathrm{E}\left[I_B(\boldsymbol{\alpha})\delta_{T_i}(A)\alpha_i \prod_{j=1}^{M} I_{A_j}(\xi_j)\,|\,M=m\right] \\ &\quad= \sum_{i=1}^{r} \mathrm{E}\left[\alpha_i I_B(\boldsymbol{\alpha})\,|\,M=m\right]\delta_{x_i}(A)\prod_{j=1}^{m}P_0(A_j) + \\ &\qquad + \sum_{i=r+1}^{r+m} \mathrm{E}\left[\alpha_i I_B(\boldsymbol{\alpha})\,|\,M=m\right] P_0(A_{i-r}\cap A)\prod_{j\neq i-r} P_0(A_j). \qquad (4.6)\end{aligned}$$

Let us now compute the lhs of (4.5). This is equal to $I+II$, where

$$I = \int_{A\setminus\{x_1,\ldots,x_r\}} \Pr\{\boldsymbol{\alpha}\in B,\xi_j\in A_j,\, j=1,\ldots,M\,|\,T_I=x, M=m\}\frac{ma/\lambda}{ma/\lambda+\sum k_i}\,\mathrm{d}P_0(x) \qquad (4.7)$$

and

$$II = \sum_{x_i\in A} \Pr\{\boldsymbol{\alpha}\in B,\xi_j\in A_j,\, j=1,\ldots,M\,|\,T_I=x_i, M=m\}\frac{k_i}{ma/\lambda+\sum k_i}. \qquad (4.8)$$

Let us see that I is equal to the second sum in expression (4.6). The equality of II with the first sum in (4.6) can be proven in a similar way, completing therefore the proof. We find that I can be written as

$$\begin{aligned}&\int_{A\setminus\{x_1,\ldots,x_r\}} \left[\sum_{\ell=1}^{m}\frac{\mathrm{E}[\alpha_{\ell+r}I_B(\boldsymbol{\alpha})\,|\,M=m]}{\mathrm{E}[\alpha_{\ell+r}\,|\,M=m]}\left(\prod_{\substack{j=1,\ldots,m\\ j\neq\ell}} P(A_j)\right) I_{A_\ell}(x)\frac{1}{m}\right] \\ &\frac{m\frac{a}{\lambda}}{m\frac{a}{\lambda}+\sum_{k_i}}\,\mathrm{d}P_0(x) \\ &\qquad = \sum_{\ell=1}^{m}\mathrm{E}[\alpha_{\ell+r}I_B(\boldsymbol{\alpha})\,|\,M=m]\left(\prod_{\substack{j=1,\ldots,m\\ j\neq\ell}} P(A_j)\right)\int_A I_{A_\ell}(x)\,\mathrm{d}P_0(x)\end{aligned}$$

which coincides with the second sum in (4.6). □

PROOF OF THE THEOREM. By Lemma 2, the distribution of $P \mid X$ can be derived by computing the conditional distribution of $M, \boldsymbol{\alpha}, \boldsymbol{\xi}$ given T_I. The latter distribution is provided by Lemma 3. In order to see the equivalence between the distributional results of Lemma 3 and the representation given in the theorem in case 1) (the other case being straightforward), it is enough to notice the following two facts:

1. the m components of the mixture distribution of $\boldsymbol{\alpha}, \boldsymbol{\xi} \mid \{M = m, T_I = x\}$ all lead to the same representation for the posterior distribution of the process. This is related to the symmetrical distribution of the β_i's and the ξ_i's which makes immaterial which of the ξ_i's becomes degenerate at x;

2. the random variable M' in the theorem is obtained by subtracting 1 to $M \mid \{T_I = x\}$. □

The posterior distribution of the process based on n observations can be derived by applying recursively the Theorem. As a consequence, we obtain the following corollary.

COROLLARY. *Let P be a random probability measure given by:*

$$P(A) = \sum_{i=1}^{M} \beta_i \delta_{\xi_i}(A) \tag{4.9}$$

where $\Pr\{M = m\} = p(m)$, $m = 1, 2, \ldots$ *with* $\Pr\{M > m\} > 0 \ \forall m$,

$$\boldsymbol{\beta} \mid M \sim \mathrm{Dir}_M\left(\frac{a}{\lambda}, \ldots, \frac{a}{\lambda}\right)$$

the ξ_i's are independent from P_0 and independent of M, $\boldsymbol{\beta}$ and P_0 is a diffuse probability measure. Suppose we have a sample of n observations $X_1, \ldots, X_n$ from P; suppose furthermore that the sample has r_n distinct values $v_1, \ldots, v_{r_n}$, $1 \leq r_n \leq n$, each of them repeated n_ℓ times, $\ell = 1, \ldots, r_n$. Then, P given $X_1, \ldots, X_n$ is distributed as

$$\sum_{i=1}^{r_n} \gamma_i \delta_{v_i} + \sum_{j=1}^{M_n^\star} \beta'_j \delta_{\xi_j} \tag{4.10}$$

where

$$\Pr\{M_n^\star = m\} \propto p(m + r_n) \frac{(m+1)\cdots(m+r_n-1)}{[(m+r_n)\frac{a}{\lambda}+1]\cdots[(m+r_n)\frac{a}{\lambda}+n-1]} \quad m = 0, 1, \ldots,$$

$$(\boldsymbol{\gamma}, \boldsymbol{\beta}') \mid M_n^\star \sim \mathrm{Dir}_{r_n + M_n^\star}\left(\frac{a}{\lambda} + n_1, \ldots, \frac{a}{\lambda} + n_{r_n}, \frac{a}{\lambda}, \ldots, \frac{a}{\lambda}\right)$$

and the ξ_j's are independent from P_0 and independent of $M_n^\star$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}'$.

PROOF. By induction, using the Theorem. □

The Corollary gives the posterior distribution for a class of priors slightly more general than the GDP. The generalisation consists in letting M be an arbitrary unbounded integer valued random variable, i.e. such that it satisfies the condition $\Pr\{M > m\} > 0$ for all m. This condition guarantees that there is no upper bound on the number of possible distinct observations.

The GDP is obtained by choosing $M = M_1$, i.e.

$$p(m) = \frac{\lambda^m}{m!} \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))};$$

in this case, the random variable $M_n^\star$ in the posterior process has probability distribution

$$\Pr\{M_n^\star = m\} \propto \frac{\lambda^m}{m!} \frac{1}{\{(m + r_n)(a/\lambda)\}^{[n]}}$$

where $x^{[n]} = x(x+1)\cdots(x+n-1)$.

Roughly speaking, the posterior process (4.10) is obtained by adding to the prior process some extra non random points which coincide with the observed values. Its most relevant feature is that it retains a Dirichlet distribution for the random weights; furthermore, it preserves the symmetry of the distribution of the weights associated to the random points ξ_i's. As a consequence, the posterior process is still analytically manageable and easy to simulate.

5. Nonparametric Bayes estimators

In this section we apply results above to derive Bayesian inference methods in a framework where the prior for the underline distribution is taken to be our generalised Dirichlet process. Throughout this section, we shall assume that P_0 is a diffuse probability measure and that h is a measurable function such that $\int h(x)\mathrm{d}P_0(x)$ and $\int h^2(x)\mathrm{d}P_0(x)$ exist finite. Results derived hereafter can be obtained, after some algebraic manipulation, by computing the appropriate expectation of representations (4.10) and (3.5.)

5.1. Estimating a mean parameter

In a nonparametric framework with $X_1, \ldots, X_n$ coming from an unknown distribution P, consider the problem of estimating $\vartheta = \int h(x)\,\mathrm{d}P(x) = \mathrm{E}_P h(X)$. A Bayesian ap-

proach is to give P the GDP prior (a, P_0, λ), and calculate the posterior mean and variance.

Before observing the data we have

$$\vartheta = \sum_{i=1}^{M_1} h(\xi_i)\beta_i$$

and

$$\begin{aligned} \mathrm{E}\,\vartheta &= \int h(x)\,\mathrm{d}P_0(x) = E_{P_0}[h(X)] \\ \mathrm{Var}\ \vartheta &= k_\lambda \int (h(x) - E_{P_0}[h(X)])^2\ \mathrm{d}P_0(x) = k_\lambda\,\mathrm{Var}_{P_0}(h(X)). \end{aligned}$$

A posteriori, we obtain

$$\vartheta \,|\, \mathrm{data} \ \sim \ \sum_{i=1}^{r_n} \gamma_i h(v_i) + \sum_{j=1}^{M_n^\star} \beta_j' h(\xi_j)$$

and the Bayes estimator of ϑ is

$$\mathrm{E}[\vartheta \,|\, \mathrm{data}] \ = \ \bar{q}_n \mathrm{E}_{H_n}[h(X)] + (1 - \bar{q}_n)\mathrm{E}_{P_0}[h(X)] \tag{5.1}$$

where

$$\bar{q}_n = \mathrm{E}[q_n], \quad q_n = \frac{r_n(a/\lambda) + n}{(M_n^\star + r_n)(a/\lambda) + n} \quad \text{and} \quad H_n = \sum_{i=1}^{r_n} \frac{n_i + (a/\lambda)}{n + r_n(a/\lambda)}\delta_{v_i}$$

is a "modified" empirical distribution function giving, for each i, weight proportional to $n_i + a/\lambda$ to the observed value v_i.

The posterior variance of ϑ can be written as

$$\begin{aligned} &\mathrm{Var}_{P_0}(h(X))\left(1 + \frac{a}{\lambda}\right)\mathrm{E}\left[\frac{1 - q_n}{1 + k_n}\right] + \mathrm{Var}_{H_n}(h(X))\mathrm{E}\left[\frac{q_n}{1 + k_n}\right] + \\ &\quad + (\mathrm{E}_{P_0}h(X) - \mathrm{E}_{H_n}h(X))^2\left((1 + r_n a/\lambda + n)\mathrm{E}\left[\frac{q_n}{1 + k_n}\right] - \bar{q}_n^2\right) \end{aligned}$$

where $k_n = (M_n^\star + r_n)(a/\lambda) + n$.

5.2. Estimating the distribution function

The above may be applied with an indicator function for h, and results in the non-parametric Bayesian estimator

$$\begin{aligned} \mathrm{E}[P_\lambda(A) \,|\, \mathrm{data}] &= \Pr\{X_{n+1} \in A \,|\, \mathrm{data}\} \\ &= H_n(A)\bar{q}_n + (1 - \bar{q}_n)P_0(A). \end{aligned}$$

Also,

$$\begin{aligned}\mathrm{Var}\{P_\lambda(A)\,|\,\mathrm{data}\} &= P_0(A)(1-P_0(A))\left(1+\frac{a}{\lambda}\right)\mathrm{E}\left[\frac{1-q_n}{1+k_n}\right] \\ &\quad + H_n(A)(1-H_n(A))\mathrm{E}\left[\frac{q_n}{1+k_n}\right] \\ &\quad + (P_0(A)-H_n(A))^2\left((1+r_n a/\lambda+n)\mathrm{E}\left[\frac{q_n}{1+k_n}\right]-\bar{q}_n^2\right).\end{aligned}$$

Let us examine the form of the predictive distribution $\Pr\{X_{n+1}\in A\,|\,\mathrm{data}\}$, assuming, for the time being, the enlarged GDP considered in the corollary.

The quantity $\bar{q}_n$ represents the a posteriori probability that a new observation equals one of the previously observed values. It determines how likely is that a new value is sampled; therefore it also controls the total number of distinct values in a sample. Notice that it depends on the distribution of M in (4.9), so that different choices for such distribution lead to different behaviour for the random number of distinct values in a sample. Further details on such aspect will be given at the end of Section 6.

Conditionally on the event that the new observation coincides with one of the previous values, the distribution H_n determines the probability that the new observation is equal to the various observed distinct values. Such distribution is unaffected by M and depends on the prior specifications only through the parameter $\tau = a/\lambda$. The probability attached to each distinct values v_i is an increasing linear function of the number n_i of times that such a value has been observed. In particular, if the frequencies n_is are all equal then H_n gives the same probability $1/r_n$ to the r_n distinct values.

The parameter τ determines the extent to which a value which has been observed more times than another is more likely to be observed again. More precisely, the bigger is τ the more similar are the probabilities that H_n attach to each distinct value, and this holds fro any value of the frequencies n_is. This can be seen by observing that the probability attached to each distinct value v_i is an increasing function of τ if n_i is below the average frequency n/r_n, is constant and equal to $1/r_n$ if $n_i = n/r_n$ and is decreasing for $n_i < n/r_n$. At the two opposite extremes we have the $\tau = 0$ case, which corresponds to the pure Dirichlet process, where the probabilities are proportional to the n_i and the $\tau \to \infty$ situation where all distinct values receive the same probability $1/r_n$.

Let us now consider more closely the behaviour of the probability $\bar{q}_n$ for the strict GDP with parameter (a, P_0, λ).

Next proposition compares $\bar{q}_n$ for the GDP with finite λ, with the corresponding value $a/(a+n)$ of a pure Dirichlet process.

PROPOSITION 2. *Let $X_1, \ldots, X_{n+1}$ be a sample from P_λ, where P_λ is a GDP(a, P_0, λ). Then*

$$\Pr\{\cup_{i=1}^{n}\{X_{n+1} = X_i\} \,|\, X_1, \ldots, X_n\} = \bar{q}_n \geq \frac{r_n(a/\lambda) + n}{r_n(a/\lambda) + n + a} > n/(a+n).$$

PROOF. By applying the Jensen inequality, one has that $\bar{q}_n$ is greater or equal to

$$\frac{r_n(a/\lambda) + n}{(\mathrm{E}M_n^\star + r_n)\,(a/\lambda) + n}.$$

The result then follows easily if we can prove that $\mathrm{E}M_n^\star \leq \lambda$. To this end, let N be a Poisson random variable with mean λ. It is easy then to check that the likelihood ratio $\Pr\{N = m\}/\Pr\{M_n^\star = m\}$ is an increasing function of m. This implies $\lambda = \mathrm{E}N \geq \mathrm{E}M_n^\star$. $\square$

Notice that this implies in particular that for $n \to \infty$, $\bar{q}_n$ tends to one, so that the predictive distribution tends to concentrate on the observed values.

Consider then, for any fixed finite λ, the behaviour of $\bar{q}_n$ as a function of a. It is easy to check that, in analogy with the pure Dirichlet process case, $\bar{q}_n$ tends to one for $a \to 0$. This corresponds to the so-called noninformative prior distribution, which produces as an estimate of P the empirical distribution function.

On the other hand, when $a \to \infty$, $\bar{q}_n$ does not converge to zero as it happens for the Dirichlet process, but, by Proposition 2, is greater than $r_n/(r_n + \lambda.)$ This behaviour is related to the fact that for large a the Dirichlet process tends to concentrate at the degenerate random probability measure P_0; on the contrary, the GDP converges for any finite λ to

$$\sum_{i=1}^{M_1} \frac{1}{M_1}\delta_{\xi_i},$$

which is a sort of empirical distribution function based on a sample of size M_1. In this extreme case the Bayesian estimator of P will give the same probability to each of the r_n distinct observed values.

5.3. Estimating the variance

Consider the parameter

$$\sigma^2 = \mathrm{Var}_{P_\lambda}\, h(X) = \int h^2(x)\,\mathrm{d}P_\lambda(x) - \left(\int h(x)\,\mathrm{d}P_\lambda(x)\right)^2.$$

The prior estimator of σ^2 is

$$\mathrm{E}\sigma^2 = (1 - k_\lambda)\,\mathrm{Var}_{P_0}(h(X)),$$

whereas the posterior estimator of σ^2 can be written as

$$\begin{aligned}\mathrm{E}[\sigma^2 \,|\, \text{data}] \;=\; & \mathrm{E}\left[\frac{k_n}{1+k_n}\left(\mathrm{Var}_{P_0}(h(X))\left(\frac{k_n - \frac{a}{\lambda}}{k_n}\right)(1-q_n) + \right.\right. \\ & \left.\left. + \; q_n(\mathrm{Var}_{H_n}(h(X))q_n + (1-q_n)\mathrm{E}_{H_n}[h(X) - \mathrm{E}_{P_0}h(X)]^2)\right)\right],\end{aligned}$$

which can be interpreted as a linear combination of three estimators of the variance.

6. Sample properties

We investigate here the structure of the distribution of a random sample of n observations from a GDP prior. In particular we shall focus on the probabilities of various configurations of ties among the observations.

This is of interest, for example, in hierarchical Bayesian models where a nonparametric prior is placed on the distribution of the parameters and ties in the parameters determine clusters in the observations (cf. [2], [5]). A drawback in the use of the Dirichlet process as a second stage nonparametric prior is that it strongly favours clusters with unequal sizes (see [8]). One aim of this section will be to examine from this perspective the effective gain which can be achieved by using the GDP prior.

The probabilities of specific configurations of ties can be obtained starting from the probabilities $\bar{q}_{n,r_n} = \bar{q}_{n,r_n}(\lambda, a)$ of sampling a previously observed value given that we have observed r_n distinct values on n observations. For example, noticing that $\Pr(X_2 = X_1 \,|\, X_1) = \bar{q}_{1,1}$ and $\Pr(X_3 = X_2 = X_1 \,|\, X_1 = X_2) = \bar{q}_{2,1}$ we have $\Pr(X_2 = X_1) = \bar{q}_{1,1}$ and $\Pr(X_3 = X_2 = X_1) = \bar{q}_{2,1}\,\bar{q}_{1,1}$. By induction we obtain

$$\Pr(X_1 = X_2 = \cdots = X_n) = \prod_{i=1}^{n-1} \bar{q}_{i,1}.$$

The case of different repeated observations can be treated in a similar way, on the basis of expression (5.1). For example, we have

$$\begin{aligned}&\Pr(X_1 = X_2 = X_3 \neq X_4 = X_5) = \\ &\quad = \Pr(X_5 = X_4 \neq X_3 = X_2 = X_1 \,|\, X_4 \neq X_3 = X_2 = X_1) \times \\ &\qquad \Pr(X_4 \neq X_3 = X_2 = X_1 \,|\, X_3 = X_2 = X_1)\Pr(X_3 = X_2 = X_1) \\ &\quad = \bar{q}_{4,2}\,\frac{1 + a/\lambda}{2 + 2a/\lambda}\,(1 - \bar{q}_{3,1})\,\bar{q}_{2,1}\,\bar{q}_{1,1}.\end{aligned}$$

If we now want to know the probability that a sample of five observations contains two distinct values, one of them repeated three times and the other twice, irrespectively of their order of appearance, we just have to multiply the previous expression for the appropriate combinatorial factor which is in this case $\binom{5}{3}$. This is because, thanks to the exchangeability of the observations, each of the $\binom{5}{3}$ different sequences of X_is leading to the same structure of ties has the same probability.

Along the lines followed by [1] for the Dirichlet process, it is then possible to extend the above discussion obtaining a general formula for the probabilities of different configurations of ties. However the resulting formula is relatively cumbersome.

We shall instead follows here a slightly different approach which leads to a simpler expression. The basic idea used in the proof of the next proposition is to derive all relevant probabilities conditionally on the total number of points M in the prior distribution.

In the following we shall use the reparametrisation $(\tau = a/\lambda, \lambda = \lambda)$, as it is more appropriate to describe sample properties. In such a reparametrisation the Dirichlet process with strength parameter a is obtained by letting λ going to infinity and τ going to zero subject to the constraint $\lambda\tau = a$.

PROPOSITION 3. *Let $X_1, \ldots, X_n$ be a sample from P defined in (4.9) and let $\mathcal{U}(n_1, \ldots, n_k)$ be an unordered configuration of ties, that is $\mathcal{U}(n_1, \ldots, n_k)$ is the set of vectors $\boldsymbol{x} \in \mathcal{X}^n$ which have k distinct values, $1 \leq k \leq n$, such that one of them is repeated n_1 times, a second one n_2 times, ..., and the k-th one n_k times. Then we have*

$$\begin{aligned} &\Pr(X_1, \ldots, X_n \in \mathcal{U}(n_1, \ldots, n_k)) = \\ &\quad = \tau^{k-1} \binom{n}{n_1 \ldots n_k} \mathrm{E}\left[\frac{(M-k+1)^{[k-1]}}{(1+M\tau)^{[n-1]}}\right] \prod_{i=1}^{n} \frac{1}{m_i!} \prod_{j=1}^{k} (1+\tau)^{[n_j-1]}, \end{aligned} \tag{6.1}$$

where m_i is the number of n_js equal to i and $x^{[n]} = x(x+1)\cdots(x+n-1)$, with the convention that $x^{[n]} = 0$ if $x \leq 0$ and $x^{[n]} = 1$ if $x > 0$ and $n = 0$.

PROOF. We shall compute $Pr(X_1, \ldots, X_n \in \mathcal{U}(n_1, \ldots, n_k) \,|\, M)$. The result then follows by taking expectation with respect to M.

An examination of the proof of the Theorem and the Corollary shows that, in the notation of the Corollary, P given $X_1, \ldots, X_n$ and M is distributed as

$$\sum_{i=1}^{r_n} \gamma_i \delta_{v_i} + \sum_{j=1}^{M-r_n} \beta'_j \delta_{\xi_j}$$

where

$$(\boldsymbol{\gamma}, \boldsymbol{\beta}') \,|\, M \sim \mathrm{Dir}_M\,(\tau + n_1, \ldots, \tau + n_{r_n}, \tau, \ldots, \tau)$$

and the ξ_j's are independent from P_0 and independent of M, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}'$. Notice furthermore that $r_n \leq M$ a.s.. It follows that, for $1 \leq j \leq r_n$,

$$\Pr(X_{n+1} = v_j \,|\, X_1, \ldots, X_n, M) = \frac{n_j + \tau}{M\tau + n}. \tag{6.2}$$

and

$$\Pr(X_{n+1} \notin \{X_1, \ldots, X_n\} \,|\, X_1, \ldots, X_n, M) = \frac{(M - r_n)\tau}{M\tau + n}. \tag{6.3}$$

We can now compute the probability of a specific sequence of X_is belonging to the configuration $\mathcal{U}(n_1, \ldots, n_k)$, namely the probability that, conditionally on M, the first k X_is are all different, the following $(n_1 - 1)$ X_is are all equal to X_1, the subsequent $(n_2 - 1)$ X_is are equal to X_2 and so on. Following the same argument employed in the discussion above Proposition 3 and using (6.2) and (6.3), such a probability can be shown to be equal to zero if $k > M$ and otherwise to

$$\prod_{i=1}^{k-1} \frac{(M-i)\tau}{M\tau + i} \prod_{i=1}^{n_1 - 1} \frac{i + \tau}{M\tau + k + i - 1} \prod_{i=1}^{n_2 - 1} \frac{i + \tau}{M\tau + k + n_1 + i - 2} \cdots \tag{6.4}$$
$$\cdots \prod_{i=1}^{n_k - 1} \frac{i + \tau}{M\tau + n_1 + \cdots, +n_{k-1} + i}.$$

A combinatorial argument then shows that there are $\binom{n}{n_1 \ldots n_k} \prod_{i=1}^{n} \frac{1}{m_i!}$ different sequences of X_is belonging to $\mathcal{U}(n_1, \ldots, n_k)$. The result then follows after some simple algebraic manipulations noticing that all such sequences have the same probability. □

Expression (6.1) can be further elaborated to get a better insight into the relation between the parameters (τ, λ) and the structure of configurations of ties.

We shall study the joint distribution of an ordered configuration of ties and the random number of distinct observations. Let us first give a precise definition of ordered configuration of ties. Let $\Pi_n = \{C_1, \ldots, C_k\}$ be a partition of $N_n = \{1, \ldots, n\}$, that is an unordered collection of k disjoint non-empty subsets of N_n such that $\cup_{i=1}^{k} C_i = N_n$. To any such a partition we can associate an ordered configuration of ties $\mathcal{O}(\Pi_n)$ by setting

$$\mathcal{O}(\Pi_n) = \Big\{ \boldsymbol{x} \in \mathcal{X}^n \,:\, x_i = x_j \;\Leftrightarrow\; i, j \in C_l \text{ for some } l = 1, \ldots, k \Big\}.$$

Notice that, denoting by n_j the cardinality of C_j and by m_i the number of n_js equal to i, there are $\binom{n}{n_1 \dots n_k} \prod_{i=1}^{n} \frac{1}{m_i!}$ ordered configurations belonging to the same unordered configuration $\mathcal{U}(n_1, \dots, n_k)$. Furthermore, each of these ordered configuration receives the same probability. A slight variation of the argument used in the proof of Proposition 3 gives

$$\Pr(X_1, \dots, X_n \in \mathcal{O}(\Pi_n) \,|\, M) = \tau^{k-1} \frac{(M-k+1)^{[k-1]}}{(1+M\tau)^{[n-1]}} \prod_{j=1}^{k} (1+\tau)^{[n_j-1]}. \tag{6.5}$$

Denote by K_n the random number of distinct values in n observations. Then expression (6.5) together with a simple combinatorial computation yields

$$\begin{aligned} &\Pr(X_1, \dots, X_n \in \mathcal{O}(\Pi_n) \,|\, K_n = k, M) = \\ &\quad = \Pr(X_1, \dots, X_n \in \mathcal{O}(\Pi_n) \,|\, K_n = k) = \frac{\prod_{j=1}^{k} (1+\tau)^{[n_j-1]}}{W(\tau, k)}, \end{aligned} \tag{6.6}$$

where

$$W(\tau, k) = \sum_{\Delta_k} \binom{n}{n_1 \dots n_k} \prod_{i=1}^{n} \frac{1}{m_i!} \prod_{j=1}^{k} (1+\tau)^{[n_j-1]}$$

and $\Delta_k = \{n_1, \dots, n_k : 1 \le n_1 \le \cdots \le n_k,\, n_1 + \cdots + n_k = n\}$. Furthermore, for $1 \le k \le \min\{n, M\}$, we have

$$\Pr(K_n = k \,|\, M) = W(\tau, k)\ \tau^{k-1}\ \frac{(M-k+1)^{[k-1]}}{(1+M\tau)^{[n-1]}}\,. \tag{6.7}$$

It follows that, conditionally on the number of distinct observations K_n, the distribution of the configurations of ties depends only on the parameter τ. Inspection of the distribution reveals that τ controls how likely unequal sizes n_is of the groups of repeated observations are. More precisely, in agreement with the interpretation of τ discussed in Section 5.2, one has that the bigger is τ the more likely are configurations with balanced group sizes. This can be seen as follows.

Consider the ratio between the probability of an ordered configuration with k unbalanced group sizes such as $(n-k+1, 1, \dots, 1)$ and the probability of an arbitrary ordered configuration with group sizes $(n_1, \dots, n_k)$. By formula (6.6) this is equal to

$$\frac{(1+\tau)^{[n-k]}}{(1+\tau)^{[n_1-1]} \cdots (1+\tau)^{[n_k-1]}}, \tag{6.8}$$

which is a non-increasing continuous function of τ ranging from $\binom{n-k}{n_1-1 \dots n_k-1}$ when $\tau \to 0$ (the Dirichlet process case) to 1 when $\tau \to +\infty$.

To appreciate the great diversity of the probabilities obtained under different τ values, even for small sample sizes, choose $n = 10$ and $k = 2$. Then the Dirichlet process will give to any ordered configuration with $n_1 = 9$ and $n_2 = 1$, $\binom{8}{4} = 70$ times the probability given to any balanced configuration with $n_1 = 5$ and $n_2 = 5$, against a ratio of 25.2 for $\tau = 1$, of 5.67 for $\tau = 5$ and of 2.28 for $\tau = 20$.

More generally, let us say that an ordered configuration is more unbalanced than another one, if the former can be obtained from the latter by moving elements of groups with lower sample sizes to groups with higher sample sizes. Then it can be shown that the ratio of the probability of the more unbalanced configuration to the probability of the other one is a decreasing function of τ tending to one as τ goes to infinity. The latter is a general phenomenon: when $\tau \to \infty$ all ordered configurations tends to receive the same probability.

Let us now consider the distribution of K_n. This distribution as well as its mean can be recovered by expression (6.7). Incidentally, notice that they are affected by the distribution of M, as suggested by the discussion in Section 5.2. In the attempt of obtaining a simpler expression, we shall consider here a different approach to the derivation of $\mathrm{E}[K_n]$.

Let $D_1 \equiv 1$ and, for $i = 2, 3, \ldots$, $D_i = 1$ if $X_i \notin \{X_1, \ldots, X_{i-1}\}$ and $D_i = 0$ otherwise. Notice that $K_n = D_1 + \cdots + D_n$. From expression (6.3) one has that

$$D_{i+1} \,|\, D_1, \ldots, D_i,\, M \sim D_{i+1} \,|\, K_i, M \sim \mathrm{Be}\left(\frac{(M - K_i)\tau}{M\tau + i}\right),$$

where $\mathrm{Be}(p)$ denotes the Bernoulli distribution with success probability p. It follows that

$$\mathrm{E}[K_{i+1} \,|\, K_i, M] = K_i + \left(\frac{(M - K_i)\tau}{M\tau + i}\right).$$

Taking expectation conditionally on M one obtains

$$\mathrm{E}[K_{i+1} \,|\, M] = \mathrm{E}[K_i \,|\, M]\left(1 - \frac{\tau}{M\tau + i}\right) + \frac{M\tau}{M\tau + i}.$$

This recursive relation can be explicitly solved yielding the following formula:

$$\mathrm{E}[K_n \,|\, M] = \sum_{j=1}^{n} \frac{M\tau}{M\tau + j - 1} \frac{\Big((M-1)\tau + j\Big)^{[n-j]}}{(M\tau + j)^{[n-j]}}. \tag{6.9}$$

In particular for $\tau = 1$ we have the simple expression $\mathrm{E}[K_n \,|\, M] = nM/(M + n - 1)$, where n and M play a symmetric role.

To complete the discussion of the sample distribution it is left to give the distribution of the distinct observations given an ordered configuration of ties. By using formula (5.1) it is possible to show that, conditionally on $\{X_1, \ldots, X_n \in \mathcal{O}(\Pi_n)\}$, the k distinct observations among the first n ones are independent and identically distributed according to P_0. As a consequence the same distribution for the distinct observations holds if we condition only on K_n.

The results derived in this section have strong implications in the following hierarchical Bayesian setting. Conditionally on the parameters $\theta_1, \theta_2, \ldots$, let the observations $X_1, X_2, \ldots$ be independent with each X_i having distribution of the form $F(\cdot, \theta_i)$. Furthermore, let the parameters $\theta_1, \theta_2, \ldots$ be a random sample from a nonparametric prior P. This setting can be adopted to model clusters among the observations: repeated values of the θ_is produce a cluster among the corresponding observations and the number of distinct values among the θ_is determine the number of clusters among the observations. See [2] and [5] for more details.

The common practice in the literature is to choose a Dirichlet process as a nonparametric prior P. This implies that we can only model the mean number of distinct clusters which depends on the parameter a. On the contrary the structure of group sizes is fixed, corresponding to the value zero for the parameter τ and it strongly favours unequal cluster sizes (see [8] for a discussion of undesirable effects of such a feature).

If we instead choose a GDP prior for P we can model both aspects. The parameter τ will regulate the degree of unbalancedness among the cluster sizes. The parameter λ can be chosen to control the number of clusters: for any given τ, $\mathrm{E}[K_n]$ is an increasing function of λ ranging from 1 to n. This can be proved by noticing that $\mathrm{E}[K_n \mid M]$ given in formula (6.9) is an increasing function of M and M has a monotone likelihood ratio.

Notice also that λ controls the expected maximum number of distinct observations. This can be seen as follows. Clearly, M is the maximum number of distinct values possibly present in the observations. Therefore $\lambda/(1 - \exp(-\lambda))$, which is increasing in λ, gives the expected maximum number of distinct observations.

References

[1] Antoniak, C.E. (1974). Mixture of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

[2] Escobar, M.D., West, M. (1998). Computing nonparametric hierarchical models. In: Dey, D., Muller, P., Sinha, D. (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics, Lecture Notes in Statistics 133, Springer, New York.

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

[4] Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.

[5] MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In: Dey, D., Muller, P., Sinha, D. (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics, Lecture Notes in Statistics 133, Springer, New York.

[6] Hjort, N.L. (2003). Topics in nonparametric Bayesian statistics [with discussion]. In *Highly Structured Stochastic Systems* (eds. P.J. Green, N.L. Hjort and S. Richardson), Oxford University Press, 455–478.

[7] Ongaro, A. and Cattaneo, C. (2002). Discrete random probability measures: a general framework for Bayesian inference. *Quaderno di Dipartimento n.6*, Dipartimento di Statistica, Università Milano-Bicocca.

[8] 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.

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

[10] Walker, S.G., Damien, P., Laud, P.W. and Smith A.F.M. (1998). Bayesian nonparametric inference for random distributions and related functions [with discussion]. *Journal of the Royal Statistical Society* **B 61**, 485–527.