On the Distribution of Jumps of the Dirichlet Process

J.-M. Rolin*

January 1993

Abstract

In this paper, a new and very easily tractable description of the Dirichlet process with diffuse parameter is presented. This description permits to simulate the posterior distribution of the nonparametric Bayesian Statistical Analysis of an i.i.d. sample.

This text presents research results of the Belgian programme on interuniversity poles of attraction initiated by the Belgian state, Prime Minister's Office, Science Policy Programming. The scientific responsability is assumed by its author.

Keywords: Dirichlet Process, Nonparametric Bayesian Statistics.

A.M.S.-M.O.S. Primary 62F15
Secondary 62G99 60G57 60C05 60G42 60G50

* University of Louvain, Belgium. Institut de Statistique and CORE.

1. Introduction.

In his famous paper, Ferguson (1973, section 4) gives a complete description of the distribution of the jumps of the Dirichlet process and of the positions where these jumps occur. However, the distribution of the jumps, derived from the distribution of the jumps of the Gamma process is rather complicated and unmanageable either for direct computation or for simulation of the Dirichlet process. This distribution is characterized by giving a description of the (finite dimensional) distribution of the n largest jumps for any n.

In this paper, we will show that a properly chosen infinite permutation of these jumps permits one to obtain a very nice and easily tractable distribution of these jumps, when the parameter of the Dirichlet process is a diffuse finite measure. Thanks to this distribution and using the results obtained in a companion paper (Rolin, (1992)) describing the posterior distribution of the Dirichlet process when observing an i.i.d. sample, simulation of this distribution is very simple.

2. Ferguson's Representation of the Dirichlet Process.

In this section, we briefly recall the useful properties of the Dirichlet process and the representation of this process as a purely atomic measure given by Ferguson (1973, section 4). Let us consider an abstract probability space $(\Omega, \mathcal{M}, P)$ and a Standard Borel space $(S, \mathcal{S})$. Let F be a random probability measure on $(S, \mathcal{S})$, i.e. an application $F : \Omega \times \mathcal{S} \to (I\!R^+, I\!B^+)$, denoted $F(B)(\omega)$, such that, for any $B \in \mathcal{S}, F(B) = F(B)(\cdot)$ is a positive random variable and for any $\omega \in \Omega, F(\cdot)(\omega)$ is a probability measure on $(S, \mathcal{S})$. Recall the definition of the Dirichlet process.

Definition: *A* random probability measure F on $(S, \mathcal{S})$ is a Dirichlet process of parameter $\alpha\mu$ where $\alpha \in I\!R_0^+$ and μ is a probability measure on $(S, \mathcal{S})$, to be written as $F \sim \mathcal{D}i(\alpha\mu)$, if for any measurable (nontrivial) partition $\mathcal{P} = \{B_\ell : 1 \leq \ell \leq k\}$ of $S, \{F(B_\ell) : 1 \leq \ell \leq k\}$ has the Dirichlet distribution $Di\big[\{\alpha\mu(B_\ell) : 1 \leq \ell \leq k\}\big]$ (see Wilks (1962)). ∎

In particular, $F(B)$ has a Beta distribution, denoted as

$$F(B) \ \sim \ B[\alpha\mu(B), \alpha\mu(B^c)] \qquad \forall B \in \mathcal{S}. \tag{2.1}$$

Following Ferguson (1973, section 4), under the assumption that μ is a diffuse prob-

ability measure on $(S, \mathcal{S})$, i.e., $\mu[\{x\}] = \mu(x) = 0 \ \ \forall x \in S$,

$$F = \sum_{1 \le j < \infty} P_j \varepsilon_{V_j} \tag{2.2}$$

where ε_x denotes the point mass at x. Now

$$V_j : (\Omega, \mathcal{M}, P) \ \rightarrow \ (S, \mathcal{S}) \qquad 1 \le j < \infty$$

are i.i.d. random variables distributed according to μ.

$$P_j : (\Omega, \mathcal{M}, P) \ \rightarrow \ (\mathbb{R}^+, \mathcal{B}^+) \qquad 1 \le j < \infty$$

are positive random variables such that

$$\sum_{1 \le j < \infty} P_j = 1 \qquad \text{and} \qquad P_{j+1} < P_j \qquad \forall j \in \mathbb{N}_0 \tag{2.3}$$

independent of the random variables $\{V_j : 1 \le j < \infty\}$, i.e., using the notations of Florens, Mouchart and Rolin (1990),

$$P_1^\infty \ \underline{\parallel} \ V_1^\infty. \tag{2.4}$$

To describe the distribution of $\{P_j : 1 \le j < \infty\}$, let

$$J_j : (\Omega, \mathcal{M}, P) \ \rightarrow \ (\mathbb{R}^+, \mathcal{B}^+) \qquad 1 \le j < \infty$$

be positive random variables such that

$$J_j \ > \ J_{j+1} \qquad \forall j \in \mathbb{N}_0 \tag{2.5}$$

$$P[J_1 \le t] = e^{-\alpha \int_t^\infty \frac{e^{-u}}{u} du} \tag{2.6}$$

$$P[J_{j+1} \le t | J_1^j] = e^{-\alpha \int_t^{J_j} \frac{e^{-u}}{u} du} \qquad \forall t \in [0, J_j] \ \ \forall j \in \mathbb{N}_0 \tag{2.7}$$

then

$$P_j \;=\; (\sum_{1\le i<\infty} J_i)^{-1} J_j. \tag{2.8}$$

Note that the distribution of $\{P_j : 1 \le j < \infty\}$ only depends on α while the distribution of $\{V_j : 1 \le j < \infty\}$ only depends on μ. From this definition, it is clearly difficult to compute the distribution of $\{P_j : 1 \le j \le k\}$ even if we known that $\sum_{1\le j<\infty} J_j \sim \Gamma(1,\alpha)$, i.e. has a Gamma distribution of shape parameter α. Note also that (2.2) is more that an equality in distribution. It is an a.s. equality because $\{P_j : 1 \le j < \infty\}$ and $\{V_j : 1 \le j < \infty\}$ may be measurably defined as functions of F under the Standard Borel space assumption.

To end this section, we recall the posterior distribution of the Dirichlet process. Let

$$X_i : (\Omega, \mathcal{M}) \;\to\; (S, \mathcal{S}) \qquad 1 \le i < \infty$$

be i.i.d. random variables distributed according to F, i.e., using the notations of Florens, Mouchart, Rolin (1990),

$$\underset{1\le i<\infty}{\|} X_i|F \tag{2.9}$$

$$X_i|F \;\sim\; F \tag{2.10}$$

i.e. $P[X_i \in B|F] \;=\; F(B) \;\; \forall B \in \mathcal{S}$.
Then, $\forall n \in I\!N$,

$$F|X_1^n \;\sim\; \mathcal{D}i(\alpha\mu + n\hat{F}_n) \tag{2.11}$$

where $\hat{F}_n$ is the empirical distribution (probability measure), i.e.

$$\hat{F}_n \;=\; \frac{1}{n} \sum_{1\le i\le n} \varepsilon_{X_i} \tag{2.12}$$

In particular,

$$E[F(B)|X_1^n] \;=\; \frac{\alpha}{\alpha+n}\mu(B) + \frac{n}{\alpha+n}\hat{F}_n(B). \tag{2.13}$$

The main idea of this paper is to look at (2.13) when n goes to infinity. For the left hand side, by the martingale convergence theorem

$$E[F(B)|X_1^n] \;\to\; E[F(B)|X_1^\infty] \qquad \text{a.s.}$$

Now, since F is identified by X_1 because $F(B) \;=\; P[X_1 \in B|F]$,

$$E[F(B)|X_1^\infty] \;=\; F(B) \qquad \text{a.s.}$$

(see e.g. Florens, Mouchart, Rolin (1990), Theorem 9.3.12).

In the next sections, we will show that

$$\hat{F}_n(B) \;\to\; \sum_{1\le k<\infty} G_k \varepsilon_{Z_k}(B) \qquad \text{a.s. } \forall B \in \mathcal{S}. \tag{2.14}$$

Due to the separability of $\mathcal{S}$, this will provide

$$F \;=\; \sum_{1\le k<\infty} G_k \varepsilon_{Z_k} \qquad \text{a.s.} \tag{2.15}$$

In this equality, $\{Z_k : 1 \le k < \infty\}$ are the distinct values of $\{X_i : 1 \le i < \infty\}$ in the order of their appearance, and $G_k, 1 \le k < \infty$, is the limit of the frequencies of the X_i's that are equal to Z_k. Now the distribution of $\{G_k : 1 \le k < \infty\}$ and of $\{Z_k : 1 \le k < \infty\}$ are simple to describe. The techniques used are related to generalized Polya urn schemes. They have been partly used by Blackwell and MacQueen (1973). Some distributional results have also been obtained by Yamato (1984a) and b)) but in neglecting the order of appearance of the distinct values. In the a.s. equality (2.15), a peculiar fact however is that it seems difficult to define $\{G_k : 1 \le k < \infty\}$ and $\{Z_k : 1 \le k < \infty\}$ as functions of F.

3. Properties of an InÆnite Sample from the Dirichlet Process.

Let $X_i : (\Omega, \mathcal{M}) \to (S, \mathcal{S}), 1 \leq i < \infty$, be a sequence of random variables such that

$$\underset{1 \leq i < \infty}{\|} X_i | F \tag{3.1}$$

$$X_i | F \ \sim \ F \qquad \forall i \in I\!N_0 \tag{3.2}$$

We define random variables describing the configuration of this sample. Let $Y_1 \equiv 1$ and

$$Y_i \ = \ 1_{\bigcap\limits_{1 \leq j \leq i-1} \{X_i \neq X_j\}} \qquad \forall i \geq 2 \tag{3.3}$$

(Y_i is an indicator equal to one if and only if X_i is a new distinct value in the sample). Next, let

$$M_n \ = \ \sum_{1 \leq i \leq n} Y_i \qquad \forall n \in I\!N_0 \tag{3.4}$$

be the number of distinct values in X_1^n and let

$$I_k \ = \ \inf\{n : M_n = k\} \qquad \forall k \in I\!N_0 \tag{3.5}$$

be the index of the k^{th} distinct value in X_1^∞.

Now,

$$Z_k \ = \ X_{I_k} \qquad \forall k \in I\!N_0 \tag{3.6}$$

is the k^{th} distinct value in X_1^∞. For the equality structure of the sample, we define

$$C_n \ = \ \sum_{1 \leq k \leq n} k 1_{\{X_n = Z_k\}} \ \ \forall n \in I\!N_0. \tag{3.7}$$

Note the following identities,

$$X_n \;=\; Z_{C_n} \tag{3.8}$$

$$M_n \;=\; \sup_{1\le i\le n} C_i. \tag{3.9}$$

Let us denote by $\mathcal{M}_1^n, \mathcal{C}_1^n, \mathcal{X}_1^n, \mathcal{Z}_1^k$ the filtrations generated by the processes M_n, C_n, X_n, Z_k. Clearly,

$$\mathcal{M}_1^n \subset \mathcal{C}_1^n \subset \mathcal{X}_1^n \tag{3.10}$$

and $\{I_k : 1 \le k < \infty\}$ is a strictly increasing sequence of $\mathcal{M}_1^n$-stopping times such that $I_k \ge k, \qquad \forall k \in I\!N_0$, since

$$\{I_k \le n\} \;=\; \{M_n \ge k\}. \tag{3.11}$$

The σ-fields associated to these stopping times are defined as

$$\mathcal{C}_1^{I_k} \;=\; \underset{k\le n<\infty}{\cup} \mathcal{C}_1^n \cap \{I_k = n\} \tag{3.12}$$

$$\mathcal{X}_1^{I_k} \;=\; \underset{k\le n<\infty}{\cup} \mathcal{X}_1^n \cap \{I_k = n\} \tag{3.13}$$

and we also define

$$\mathcal{Z}_1^{M_n} \;=\; \underset{1\le k\le n}{\cup} \mathcal{Z}_1^k \cap \{M_n = k\}. \tag{3.14}$$

It is a simple set theory exercice to show that

$$\mathcal{X}_1^{I_k} \;=\; \mathcal{Z}_1^k \vee \mathcal{C}_1^{I_k} \tag{3.15}$$

$$\mathcal{X}_1^n \;=\; \mathcal{Z}_1^{M_n} \vee \mathcal{C}_1^n \tag{3.16}$$

and consequently

$$\mathcal{X}_1^\infty \;=\; \mathcal{Z}_1^\infty \vee \mathcal{C}_1^\infty \tag{3.17}$$

Finally, we introduce the number of values equal to Z_k in the sample $\{X_i : 1 \leq i \leq n\}$, i.e.

$$D_{n,k} \;=\; \sum_{1\leq i\leq n} 1_{\{C_i=k\}}. \tag{3.18}$$

now the empirical distribution (2.12) may be rewritten as

$$\hat{F}_n \;=\; \frac{1}{n} \sum_{1\leq k\leq M_n} D_{n,k}\varepsilon_{Z_k} \tag{3.19}$$

where all Z_k's are distinct and all $D_{n,k}$'s are greater or equal to 1.

It remains to look at the asymptotic behaviour of $M_n, D_{n,k}, Z_k$ as n goes to infinity. The only tool is provided by the following Polya urn scheme (see Blackwell and MacQueen, (1973)),

$$P[X_{n+1} \in B|X_1^n] \;=\; \frac{\alpha}{\alpha+n}\mu(B) + \sum_{1\leq k\leq M_n} \frac{D_{n,k}}{\alpha+n}\varepsilon_{Z_k}(B) \tag{3.20}$$

Under the following assumption, that will be maintained throughout this paper,

Assumption: μ is a diffuse probability measure,

it can be deduced that

$$P[C_{n+1} = k|\mathcal{X}_1^n] \;=\; \frac{D_{n,k}}{\alpha+n}1_{\{M_n\geq k\}} + \frac{\alpha}{\alpha+n}1_{\{M_n=k-1\}} \tag{3.21}$$

3.1. Distribution of the Jump Positions.

In this section, we will show that $\mathcal{Z}_1^\infty \perp\!\!\!\perp \mathcal{C}_1^\infty$ and that $\{Z_k : 1 \le k < \infty\}$ are i.i.d. with distribution μ. This result is well known (see e.g. Antoniak (1974)) but we will however prove it because to the best of our knowledge no rigorous mathematical proof is available in the literature.

Proposition 3.1. Under the assumption, $\forall k \in \mathbb{N}_0$

(i) $Z_{k+1} \perp\!\!\!\perp (\mathcal{Z}_1^k \vee \mathcal{C}_1^{I_{k+1}})$

(ii) $P[Z_{k+1} \in B] \;=\; \mu(B)$ ∎

As a straightforward corollary, (see e.g. Florens, Mouchart, Rolin (1990), Theorem 7.6.3):

Corollary 3.2. Under the assumption,

(i) $\underset{1\le k<\infty}{\perp\!\!\!\perp} Z_k$

(ii) $\mathcal{Z}_k \;\sim\; \mu.$ ∎

Proof of proposition 3.1.

Let $A \in \mathcal{Z}_1^k$ and $C \in \mathcal{C}_1^n$ with $n \ge k$. Then, using (3.6)

$$\begin{aligned} & P\big[\{Z_{k+1} \in B\} \cap A \cap C \cap \{C_{n+1} = r\} \cap \{I_{k+1} = n+1\}\big] \\ &= P\big[\{X_{n+1} \in B\} \cap A \cap \{M_n = k\} \cap C \cap \{M_{n+1} = k+1\}\big] 1_{\{r=k+1\}} \\ &= E\big[P[X_{n+1} \in B, M_{n+1} = M_n + 1 | \mathcal{X}_1^n] 1_{A\cap\{M_n=k\}\cap C}\big] 1_{\{r=k+1\}} \end{aligned}$$

since $A \cap \{M_n = k\} \in \mathcal{Z}_1^{M_n} \subset \mathcal{X}_1^n$ by (3.14) and (3.16). Now, by (3.20)

$$P\big[X_{n+1} \in B, M_{n+1} = M_n + 1 | \mathcal{X}_1^n\big] \;=\; \frac{\alpha}{\alpha+n}\mu(B).$$

Hence

$$\begin{aligned} & P\big[\{Z_{k+1} \in B\} \cap A \cap C \cap \{C_{n+1} = r\} \cap \{I_{k+1} = n+1\}\big] \\ &= \mu(B)\frac{\alpha}{\alpha+n} P\big[A \cap \{M_n = k\} \cap C\big] 1_{\{r=k+1\}} \\ &= \mu(B) P\big[A \cap C \cap \{C_{n+1} = r\} \cap \{I_{k+1} = n+1\}\big]. \end{aligned}$$

By the monotone class theorem, this finishes the proof because
$C \cap \{C_{n+1} = r\} \cap \{I_{k+1} = n+1\}$ is a generating set of $\mathcal{C}_1^{I_{k+1}}$ by (3.12). ∎

Proposition 3.3. Under the assumption,

$$\mathcal{Z}_1^\infty \underline{\parallel} \mathcal{C}_1^\infty.$$

■

Proof: We first show, by an induction argument, that

$$\mathcal{Z}_1^k \underline{\parallel} \mathcal{C}_1^{I_k} \qquad \forall k \in I\!N_0. \tag{3.22}$$

This is true for $k = 1$, because $I_1 = 1$ and $C_1 = 1$. Now, from (3.21), it is easily seen that, $C_{n+1} \underline{\parallel} \mathcal{X}_1^n | \mathcal{C}_1^n$ and by Theorem 6.2.2 of Florens, Mouchart, Rolin (1990), this is equivalent to

$$\mathcal{C}_1^\infty \underline{\parallel} \mathcal{X}_1^n | \mathcal{C}_1^n \qquad \forall n \in I\!N_0. \tag{3.23}$$

Now, n may be replaced by I_k, since, $\forall A \in \mathcal{X}_1^\infty$

$$P(A|\mathcal{X}_1^{I_k}) = \sum_{k \le n < \infty} P(A|\mathcal{X}_1^n) 1_{\{I_k = n\}}.$$

Therefore, since $\mathcal{Z}_1^k \subset \mathcal{X}_1^{I_k}$ by (3.15)

$$\mathcal{C}_1^\infty \underline{\parallel} \mathcal{Z}_1^k | \mathcal{C}_1^{I_k} \quad \forall k \in I\!N_0. \tag{3.24}$$

By Theorem 2.2.10 of Florens, Mouchart, Rolin (1990), (3.22) and (3.24) are equivalent to

$$\mathcal{Z}_1^k \underline{\parallel} \mathcal{C}_1^\infty. \tag{3.25}$$

In particular, this implies $\mathcal{Z}_1^k \underline{\parallel} \mathcal{C}_1^{I_{k+1}}$. Therefore, by Corollary 7.6.4 of Florens, Mouchart, Rolin (1990) this, along with Proposition 3.1 (i), is equivalent to

$$\mathcal{Z}_{k+1} \underline{\parallel} \mathcal{Z}_1^k \underline{\parallel} \mathcal{C}_1^{I_{k+1}} \tag{3.26}$$

which implies $\mathcal{Z}_1^{k+1} \underline{\parallel} \mathcal{C}_1^{I_{k+1}}$. This completes the induction argument. Therefore (3.22) is true for all $k \in I\!N_0$ and this implies (3.25) and a monotone class argument finishes the proof. ■

The following corollary may be useful for predictive computations for a finite sample.

Corollary 3.4. Under the assumption

$$\mathcal{Z}_1^{M_n} \underline{\parallel} \mathcal{C}_1^n | M_n \qquad \forall n \in I\!N_0.$$

■

Proof: From (3.14), $\mathcal{Z}_1^{M_n} \subset \mathcal{Z}_1^\infty \vee \sigma(M_n)$. But $\sigma(M_n) \subset \mathcal{C}_1^n$. Therefore by Corollary 2.2.11 of Florens, Mouchart, Rolin (1990),

$$\mathcal{Z}_1^{M_n} \underline{\parallel} \mathcal{C}_1^\infty | M_n.$$

■

3.2. Asymptotic Behaviour of the Number of Distinct Values.

From (3.21), we deduce that, $\forall k \leq n+1$,

$$P[M_{n+1} = k | \mathcal{X}_1^n] \;=\; \frac{n}{\alpha+n} 1_{\{M_n = k\}} + \frac{\alpha}{\alpha+n} 1_{\{M_n = k-1\}}. \tag{3.27}$$

Therefore

$$P[M_{n+1} = k] \;=\; \frac{n}{\alpha+n} P[M_n = k] \;+\; \frac{\alpha}{\alpha+n} P[M_n = k-1]. \tag{3.28}$$

By an induction argument on n, it has been proved (see Antoniak (1974) section 4) that, $\forall n \in I\!N_0$

$$P[M_n = k] \;=\; \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^k s_{n,k} \qquad \forall 1 \leq k \leq n \tag{3.29}$$

where $s_{n,k}$ are known as the absolute values of Stirling numbers of the first kind that are determined by

$$\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \;=\; \prod_{0 \leq i \leq n-1} (\alpha+i) \;=\; \sum_{1 \leq k \leq n} \alpha^k s_{n,k}. \tag{3.30}$$

However, to describe the asymptotic behaviour of M_n, it is preferable to use another argument. Indeed, from (3.27),

$$P[Y_i = 1|\mathcal{X}_1^{i-1}] \;=\; P[M_i = M_{i-1} + 1|\mathcal{X}_1^{i-1}] \;=\; \frac{\alpha}{\alpha + i - 1}. \tag{3.31}$$

Therefore, M_n is the sum of n independent indicator random variables. This entails the following proposition.

Proposition 3.5. Under the assumption,

$$\frac{M_n}{\ln n} \;\to\; \alpha \qquad \text{a.s. and in } L_2 \text{ as } n \to \infty.$$

■

Proof: Let $a_i = \ln\,(\alpha + i - 1)$. Then,

$$\sum_{2 \le i < \infty} E[\{\frac{Y_i - E(Y_i)}{a_i}\}^2] \;=\; \sum_{2 \le i < \infty} \frac{1}{a_i^2} \frac{\alpha(i-1)}{(\alpha + i - 1)^2} \;\le\; \alpha \sum_{2 \le i < \infty} \frac{1}{(\alpha + i - 1)a_i^2} < \infty.$$

Therefore, $\displaystyle\sum_{2 \le i < \infty} \frac{Y_i - E(Y_i)}{a_i} < \infty$ a.s. and by Kronecker's lemma

$$\frac{1}{a_n}(M_n - E(M_n)) \to 0 \qquad \text{a.s.}$$

But,

$$E(M_n) \;=\; \sum_{1 \le i \le n} \frac{\alpha}{\alpha + i - 1} \;=\; \alpha\{\psi(\alpha + n) - \psi(\alpha)\} \tag{3.32}$$

where ψ is the digamma function (i.e. $\psi(\alpha) = \dfrac{d}{d\alpha}\,\ln\Gamma(\alpha)$). Since $\dfrac{\psi(\alpha)}{\ln\alpha} \to 1$ as $\alpha \to \infty$ (see, e.g., Abramovitz and Stegun (1964)), $\dfrac{1}{a_n}E(M_n) \to \alpha$. Therefore $\dfrac{M_n}{\ln(\alpha + n - 1)} \to \alpha$ a.s. or equivalently $\dfrac{M_n}{\ln n} \to \alpha$ a.s. Convergence in L_2 is straightforward. ■

It was already known that $\dfrac{E(M_n)}{\ln n} \to \alpha$ as $n \to \infty$ and that $M_n \to \infty$ a.s. (see Antoniak (1974) section 4). Thus, Proposition 3.5 not only improves known results but also has some important practical consequences.

On one hand, from a sampling statistical inference point of view, we already have noticed that the distribution of the Z_k's only depends on μ, is independent of

the C_i's whose distribution only depends on α. It will be shown in the next section, that M_n is a sufficient statistic for α in $\mathcal{C}_1^n$. Therefore, in the presence of a sample of size $n, (\mathcal{X}_1^n)$, from a Dirichlet distribution, $(\mathcal{D}i(\alpha\mu))$, a.s. consistant estimates of α and μ will be provided by $\dfrac{M_n}{\ln n}$ and $\dfrac{1}{M_n}\displaystyle\sum_{1\le k\le M_n}\varepsilon_{Z_k}$. Finite sample properties of these estimates will be derived from Corollary 3.2 and 3.4 and from (3.29) and (3.31). Asymptotic properties may be derived from the following result the proof of which, based on characteristic functions, is left to the reader.

Proposition 3.6. Under the assumption,

$$(\alpha\ \ln n)^{-\frac{1}{2}}(M_n - \alpha\ \ln n)\ \rightarrow\ \mathcal{Z}$$

in distribution as $n \to \infty$ where $\mathcal{Z}$ has a standard normal distribution. ∎

On the other hand, from a Bayesian statistical inference point of view, proposition 3.5 solves the delicate problem of the specification of α. Indeed, if one is facing a sample of size n of distinct values ($M_n = n$) (the more frequent case), he will consistently choose α to be equal to $\dfrac{n}{\ln n}$. It may be seen from (2.13), that this choice will preserve the asymptotic properties of the posterior distribution.

3.3. Distributions of the Configurations.

The aim of this section is to derive from (3.21) the distribution of the equality structure of the sample of size n, i.e., the distribution of $\{C_i : 1 \le i \le n\}$ and the distribution of $\{D_{n,\ell} : 1 \le \ell \le k\}$ for $k \le n$.

Some results are already available in the literature (see, e.g., Antoniak (1974) section 4 and Yamato (1984). But these results forget the order of arrival of the observations and as a consequence are not at all easily tractable. The way we define $D_{n,\ell}$ in (3.18), i.e. the number of observations equal to the ℓ^{th} distinct value of the sample, retain some order of appearance and this will furnish easily tractable distributions.

Proposition 3.7. Under the assumption, $\quad \forall\ 2 \le n < \infty$

$$P\left[\bigcap_{2\le i\le n}\{C_i = \ell_i\}\right] = \frac{\Gamma(\alpha)\alpha^{m_n}}{\Gamma(\alpha+n)}\left(\prod_{1\le k\le m_n}(d_{n,k}-1)!\right)\mathbf{1}_{\bigcap_{2\le i\le n}\{\ell_i \le m_{i-1}+1\}}$$

where $m_i = \sup_{1\le j\le i}\ell_j, d_{n,k} = \displaystyle\sum_{1\le i\le n} 1_{\{\ell_i = k\}}$ and $\ell_1 = 1$. ∎

Proof: From (3.21), it is easily seen that

$$P[\bigcap_{2\le i\le n}\{C_i=\ell_i\}] \;=\; P[\bigcap_{2\le i\le n-1}\{C_i=\ell_i\}]a_n \;=\; \prod_{2\le i\le n} a_i.$$

where

$$\begin{aligned} a_i &= \frac{d_{i-1,\ell_i}}{\alpha+i-1}\mathbf{1}_{\{\ell_i\le m_{i-1}\}} + \frac{\alpha}{\alpha+i-1}\mathbf{1}_{\{\ell_i=m_{i-1}+1\}} \\ &= (\alpha+i-1)^{-1} d_{i-1,\ell_i}^{\mathbf{1}_{\{\ell_i\le m_{i-1}\}}} \alpha^{\mathbf{1}_{\{\ell_i=m_{i-1}+1\}}} \mathbf{1}_{\{\ell_i\le m_{i-1}+1\}} \end{aligned}$$

The proof is completed by noticing that

$$\sum_{2\le i\le n} \mathbf{1}_{\{\ell_i=m_{i-1}+1\}} = m_n - 1 \text{ and that}$$

$$\prod_{2\le i\le n} d_{i-1,\ell_i}^{\mathbf{1}_{\{\ell_i\le m_{i-1}\}}} = \prod_{1\le k\le m_n} (d_{n,k}-1)!$$

under the constraints $\ell_i \le m_{i-1}+1, \quad \forall\ 2\le i\le n.$

■

Let us remark that by (3.29), it follows from Proposition 3.7., that

$$\begin{aligned} &P\left[\bigcap_{2\le i\le n}\{C_i=\ell_i\}\middle| M_n=k\right] \\ &= s_{n,k}^{-1}\left(\prod_{1\le\ell\le k}(d_{n,\ell}-1)!\right)\mathbf{1}_{\bigcap_{2\le i\le n}\{\ell_i\le m_{i-1}+1\}\cap\{m_n=k\}} \end{aligned}$$

This shows, as it was announced in the preceding section, that M_n is a sufficient statistic for α.

Let us also note that, besides of the constraints, the expression of the probabilities only depends on m_n and $d_{n,k}, 1\le k\le m_n$. This leads to the following proposition.

Proposition 3.8. Under the assumption, $\forall\ n\in\mathbb{N}_0, \quad \forall\ k\le n, \quad \forall\ d_\ell\ge 1,$ $1\le\ell\le k$ such that $\sum_{1\le\ell\le k} d_\ell = n$

$$P\left[\bigcap_{1\le\ell\le k}\{D_{n,\ell}=d_\ell\}\right] = \frac{\Gamma(\alpha)\alpha^k\Gamma(n)}{\Gamma(\alpha+n)}\prod_{1\le\ell\le k-1}(n-s_\ell)^{-1}$$

where $s_\ell = \sum_{1\le r\le\ell} d_r, \quad 1\le\ell\le k-1.$

■

Proof: Note that under the conditions on the d_ℓ's, $\underset{1\leq\ell\leq k}{\cap}\{D_{n,k}=d_\ell\}\subset\{M_n=k\}$. Now, with the notations of proposition 3.7., the number of choices of $\{\ell_i;1\leq i\leq n\}$ satisfying the constraints and $d_{n,\ell}=d_\ell, 1\leq\ell\leq k$, is given by

$$C_{n-1}^{d_1-1}C_{n-d_1-1}^{d_2-1}C_{n-d_1-d_2-1}^{d_3-1}\cdots C_{d_k-1}^{d_k-1}$$

$$=(n-1)!\{\prod_{1\leq\ell\leq k}(d_\ell-1)!\}^{-1}\{\prod_{1\leq\ell\leq k-1}(n-s_\ell)\}^{-1}$$

■

Proposition 3.8 serves as an initial value to obtain, in an induction argument, the expressions of the probabilities when $\sum_{1\leq\ell\leq k} d_\ell < n$. These are given in the next proposition. We will however omit the proof of this proposition. This proof is very simple in nature since it is based on an induction argument on n, but it is quite tedious to write. The main difficulty was to find the expression of the probabilities!

Proposition 3.9. Under the assumption, $\forall\ n\in\mathbb{N}_0,\quad \forall\ k\leq n,\quad \forall\ d_\ell\geq 1$, $1\leq\ell\leq k$, such that $\sum_{1\leq\ell\leq k} d_\ell\leq n$,

$$P[\underset{1\leq\ell\leq k}{\cap}\{D_{n,\ell}=d_\ell\}] \;=\; \frac{\alpha^k\Gamma(n)\Gamma(\alpha+n-s_k)}{\Gamma(\alpha+n)\Gamma(n+1-s_k)}\cdot\prod_{1\leq\ell\leq k-1}(n-s_\ell)^{-1}$$

where $s_\ell=\sum_{1\leq r\leq\ell} d_r,\quad 1\leq\ell\leq k$

■

Let us remark that the result of proposition 3.9 does not completely describe the distribution of $\{D_{n,\ell}:1\leq\ell\leq k\}$ since the restrictions on the d_ℓ's imply that $M_n\geq k$. However if one of the d_ℓ's is equal to zero, all subsequent d_ℓ's must be equal to zero and the expression of the probabilities are given by the expression of the proposition 3.8. with k replaced by the index of the last d_ℓ that is greater than zero.

3.4. Asymptotic Behaviour of the Frequencies

In this section, we look at the asymptotic behaviour of the $D_{n,k}$'s and show that it induces a representation of F.

Proposition 3.10. Under the assumption, $\forall\ k \in \mathbb{N}_0$

$$\frac{D_{n,k}}{n} \longrightarrow G_k \qquad \text{a.s. and in } L_p \quad \forall\ p \in (0,\infty)$$

as $n \to \infty$. Moreover, $\forall\ k \in \mathbb{N}_0$,

$$E[G_k] = E[(\alpha + I_k)^{-1}] = \frac{1}{\alpha}\left(\frac{\alpha}{\alpha+1}\right)^k \tag{3.33}$$

■

Proof: It suffices to prove that, $\forall\ k \in \mathbb{N}_0$, $\dfrac{D_{n,k}}{\alpha+n}$ is a positive bounded submartingale (martingale if $k = 1$) (see, e.g., Dellacherie – Meyer (1980) or Neveu (1972)). Indeed,

$$\begin{aligned}
&E[D_{n+1,k} \mid C_1^n] \\
&\quad = D_{n,k}P[C_{n+1} \neq k \mid C_1^n] + (D_{n,k}+1)P[C_{n+1} = k \mid C_1^n] \\
&\quad = D_{n,k} + P[C_{n+1} = k \mid C_1^n] \\
&\quad = D_{n,k} + \frac{D_{n,k}}{\alpha+n}\mathbf{1}_{\{M_n \geq k\}} + \frac{\alpha}{\alpha+n}\mathbf{1}_{\{M_n = k-1\}} \\
&\quad = \frac{\alpha+n+1}{\alpha+n}D_{n,k} + \frac{\alpha}{\alpha+n}\mathbf{1}_{\{M_n = k-1\}}
\end{aligned}$$

by (3.21) and because $\{M_n \geq k\} = \{D_{n,k} > 0\}$
Therefore,

$$\begin{aligned}
E\left[\frac{D_{n+1,k}}{\alpha+n+1} \mid C_1^n\right] &= \frac{D_{n,k}}{\alpha+n} + \frac{\alpha}{(\alpha+n)(\alpha+n+1)}\mathbf{1}_{\{M_n=k-1\}} \\
&\geq \frac{D_{n,k}}{\alpha+n} \qquad (= \text{ if } k=1)
\end{aligned}$$

Hence,

$$E\left[\frac{D_{n,1}}{\alpha+n}\right] = E\left[\frac{D_{1,1}}{\alpha+1}\right] = \frac{1}{\alpha+1}$$

and $\forall\ k > 1$, since $D_{k-1,k} = 0$

$$E\left[\frac{D_{n,k}}{\alpha+n}\right] = E\left[\sum_{k \leq r \leq n} \frac{\alpha}{(\alpha+r)(\alpha+r-1)}\mathbf{1}_{\{M_{r-1}=k-1\}}\right]$$

Passing through the limit gives:

$$E[G_k] = \sum_{k \le r < \infty} \frac{\alpha}{(\alpha + r)(\alpha + r - 1)} P[M_{r-1} = k - 1]$$

Now, let us compute the distribution of I_k. Clearly

$$\begin{aligned} P[I_k = n] &= P[M_n = k, M_{n-1} = k - 1] \\ &= \frac{\alpha}{\alpha + n - 1} P[M_{n-1} = k - 1] \end{aligned}$$

Therefore,

$$E[G_k] = \sum_{k \le r < \infty} \frac{1}{\alpha + r} P[I_k = r] = E[(\alpha + I_k)^{-1}]$$

and, by (3.29), $\quad \forall\ k \in \mathbb{N}_0, \quad \forall\ n \ge k$

$$P[I_k = n] = \frac{\Gamma(\alpha)\alpha^k}{\Gamma(\alpha + n)} s_{n-1,k-1} \tag{3.34}$$

This implies that

$$\sum_{k \le n < \infty} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} s_{n-1,k-1} = \frac{1}{\alpha^k}$$

Hence, we deduce that

$$E[(\alpha + I_k)^{-1}] = \frac{1}{\alpha} \left(\frac{\alpha}{\alpha + 1} \right)^k \tag{3.35}$$

■

It may be noticed, from (3.34), that if $\alpha \le 1, E(I_k) = \infty \quad \forall\ k > 1$, while if $\alpha > 1, E(I_k) = (\alpha - 1) \left[\left(\frac{\alpha}{\alpha - 1} \right)^k - 1 \right] \quad \forall\ k \in \mathbb{N}_0$. This sugests a modification of the Dirichlet process behaviour according to whether $\alpha \le 1$ or $\alpha > 1$. This has been reflected in Cifarelli–Reggazini (1990) where they give different expressions of the density of the mean of the Dirichlet process for $\alpha \le 1$ and $\alpha > 1$.

Thanks to proposition 3.10, we may define the following random purely atomic measure on $(S, \mathcal{S})$,

$$G = \sum_{1 \le k < \infty} G_k \mathcal{E}_{Z_k} \tag{3.36}$$

and we have the following proposition.

Proposition 3.11. Under the assumption, G is a probability measure on $(S, \mathcal{S})$ and, as $n \to \infty$,

$$\sup_{B \in \mathcal{S}} \mid \hat{F}_n(B) - G(B) \mid \to 0 \text{ a.s.}$$

■

Proof: Clearly, $\quad \forall\ k \leq n,$

$$\sum_{1 \leq \ell \leq k} \frac{D_{n,\ell}}{n} \leq \sum_{1 \leq \ell \leq n} \frac{D_{n,\ell}}{n} = 1$$

Therefore,

$$\sum_{1 \leq k < \infty} G_k \leq 1 \text{ a.s.}$$

But,

$$E[\sum_{1 \leq k < \infty} G_k] = \sum_{1 \leq k < \infty} \frac{1}{\alpha} \left(\frac{\alpha}{\alpha + 1} \right)^k = 1.$$

Hence $\sum_{1 \leq k < \infty} G_k = 1$ a.s.. Now it is easy from this to show, that for any $r \in \mathbb{N}_0$,

$$\limsup_n \sup_{B \in \mathcal{S}} \mid \hat{F}_n(B) - G(B) \mid \leq \sum_{r+1 \leq k < \infty} G_k$$

■

Finally, we obtain the following representation of the Dirichlet Process.

Proposition 3.12. If $F \sim \mathcal{D}i(\alpha\mu)$ with μ diffuse probability on $(S, \mathcal{S})$, then

$$F = \sum_{1 \leq k < \infty} G_k \mathcal{E}_{Z_k} \qquad \text{a.s.}$$

where

(i) $\quad \mathcal{Z}_1^\infty \perp\!\!\!\perp \mathcal{G}_1^\infty$

(ii) $\quad \perp\!\!\!\perp_{1 \leq k < \infty} Z_k$

(iii) $\quad Z_k \sim \mu$

■

Proof: By (2.13),

$$E[F(B) \mid \mathcal{X}_1^n] = \frac{\alpha}{\alpha+n}\mu(B) + \frac{n}{\alpha+n}\hat{F}_n(B)$$

Hence by the martingale convergence theorem and Proposition 3.11,

$$E[F(B) \mid \mathcal{X}_1^\infty] = G(B) \qquad \text{a.s.} \qquad \forall\ B \in \mathcal{S}$$

But, since F is identified by $\mathcal{X}_1^\infty$, because $F(B) = P[X_1 \in B \mid F]$, (see e.g. Florens, Mouchart, Rolin (1990), Theorem 9.3.12)

$$E[F(B) \mid \mathcal{X}_1^\infty] = F(B) \qquad \text{a.s.}$$

Therefore, $F(B) = G(B)$ a.s. $\forall\ B \in \mathcal{S}$ and since $\mathcal{S}$ is separable, i.e., is generated by a countable subcollection, $F(B) = G(B) \quad \forall\ B \in \mathcal{S}$ a.s. .
Now, clearly, $\mathcal{G}_1^\infty \subset \overline{\mathcal{C}_1^\infty}$, i.e., $\mathcal{C}_1^\infty$ completed by the null sets of $\mathcal{X}_1^\infty$. Therefore the rest of the proof relies on Corollary 3.2. and Proposition 3.3.

■

The last section will be devoted to the description of the finite dimensional distributions of the jump sizes process $\{G_k : 1 \leq k < \infty\}$. But before proceeding to that section, let us provide a seemingly easier argument for the results obtained in this section. Note first that

$$E[F(Z_k) \mid \mathcal{X}_1^n]\mathbf{1}_{\{M_n \geq k\}} = \frac{D_{n,k}}{\alpha+n}\mathbf{1}_{\{M_n \geq k\}} \tag{3.37}$$

and consequently,

$$E[\sum_{1 \leq k \leq M_n} F(Z_k) \mid \mathcal{X}_1^n] = \frac{n}{\alpha+n} \tag{3.38}$$

Since $M_n \to \infty$ a.s., we obtain

$$\begin{aligned} F(Z_k) &= \lim_{n\to\infty} E[F(Z_k) \mid \mathcal{X}_1^n] \\ &= \lim_{n\to\infty} E[F(Z_k) \mid \mathcal{X}_1^n]\mathbf{1}_{\{M_n \geq k\}} \\ &= \lim_{n\to\infty} \frac{D_{n,k}}{\alpha+n}\mathbf{1}_{\{M_n \geq k\}} = \lim_{n\to\infty} \frac{D_{n,k}}{\alpha+n} \end{aligned}$$

Therefore,

$$G_k = F(Z_k) \text{ a.s.} \qquad \forall\ k \in \mathbb{N}_0 \tag{3.39}$$

and from (3.38), taking the expectation and afterwards the limit, we obtain:

$$\sum_{1\leq k<\infty} F(Z_k) = 1 \text{ a.s.} \tag{3.40}$$

3.5. Distribution of the Jump Sizes

The task to be undertaken in this section, is the passing to the limit in the expressions of the probabilities given in proposition 3.9.

We begin by making a change of variable by considering the number of observations smaller or equal to the ℓ^{th} distinct value, i.e.

$$S_{n,\ell} = \sum_{1\leq j\leq \ell} D_{n,j} = \sum_{1\leq i\leq n} \mathbf{1}_{\{C_i\leq \ell\}} \quad \forall\ 1\leq \ell\leq k \tag{3.41}$$

so that, if $S_{n,0} = 0$,

$$D_{n,\ell} = S_{n,\ell} - S_{n,\ell-1} \qquad \forall\ 1\leq \ell\leq k \tag{3.42}$$

Therefore, if $s = (s_1, s_2, \ldots, s_k)$ and

$$P\left[\bigcap_{1\leq \ell\leq k} \{S_{n,\ell} = s_\ell\}\right] = p_{n,k}(s), \tag{3.43}$$

$$\text{for } 1\leq s_1 < s_2 \ldots < s_k \leq n,$$

$$p_{n,k}(s) = \frac{\alpha^k \Gamma(n)\Gamma(\alpha+n-s_k)}{\Gamma(\alpha+n)\Gamma(n+1-s_k)} \prod_{1\leq \ell\leq k-1} (n-s_\ell)^{-1} \tag{3.44}$$

Note that the support of $p_{n,k}(s)$ is strictly contained in the set $\{1\leq s_1\leq s_2\ldots\leq s_k\leq n\}$ but strictly includes the set $\{1\leq s_1<s_2\ldots<s_k\leq n\}$ for which the expressions are given. But if $S_n = (S_{n,1}, S_{n,2}, \ldots, S_{n,k})$, by the remark following Proposition 3.9., for any bounded continuous function on $\mathbb{R}^k$, we have

$$\mid E[f(\frac{1}{n}S_n)] - \sum_{1\leq s_1<s_2\cdots s_k\leq n} f(\frac{1}{n}s)p_{n,k}(s) \mid \leq \sup_{\mathbb{R}^k} \mid f \mid \cdot P\left[M_n < k\right] \tag{3.45}$$

which tends to zero as $n\to\infty$, since $M_n\to\infty$ a.s.. Now if we define, for $x\in\mathbb{R}^k$

$$f_{n,k}(x) = n^k p_{n,k}([nx])\mathbf{1}_{\{1\leq [nx_1]<[nx_2]<\ldots<[nx_k]\leq 1\}} \tag{3.46}$$

(where $[y]$ stands for the largest integer smaller or equal to y.), then

$$\int_{I\!R^k} f(x) f_{n,k}(x) dx$$
$$= \sum_{1 \le s_1 < s_2 \cdots < s_k \le 1} p_{n,k}(s) n^k \int_{\underset{1 \le \ell \le k}{\times}} [\frac{s_\ell}{n}, \frac{s_\ell}{n} + \frac{1}{n}[f(x) dx \tag{3.47}$$

Therefore, by the continuity of f,

$$\lim_{n \to \infty} E[f(\frac{1}{n} S_n)] - \int_{I\!R^k} f(x) f_{n,k}(x) dx \ = \ 0 \tag{3.48}$$

But, by Proposition 3.10, if we define

$$H_k \ = \ \sum_{1 \le \ell \le k} G_\ell \tag{3.49}$$
$$\lim_{n \to \infty} E[f(\frac{1}{n} S_n)] \ = \ E[f(H_1, H_2, \cdots, H_k)]$$

Hence,

$$E[f(H_1, H_2, \cdots H_k)]$$
$$= \lim_{n \to \infty} \int_{I\!R^k} f(x) f_{n,k}(x) dx. \tag{3.50}$$

It suffices now to compute the limit as $n \to \infty$ of $f_{n,k}(x)$ to obtain the next proposition.

Proposition 3.13. Under the assumption, $\forall k \in I\!N_0$,

$$P[H_1 \in dh_1, H_2 \in dh_2, \cdots H_k \in dh_k]$$
$$= \ f_k(h_1, h_2, \cdots h_k) dh_1 dh_2 \cdots dh_k$$

where

$$f_k(h_1, h_2, \cdots h_k) \ = \ \alpha^k (1 - h_k)^{\alpha - 1} \prod_{1 \le \ell \le k-1} (1 - h_\ell)^{-1} \ 1_{\{0 < h_1 < h_2 \cdots < h_k < 1\}}$$

■

Proof: First of all, $\forall h \in \{0 \leq h_1 \leq h_2 \cdots \leq h_k \leq 1\}^c$, it is easy to see that for n sufficiently large, $f_{n,k}(h) = 0$. Now for $h \in \{0 < h_1 < h_2 \cdots < h_k < 1\}$, if $s_\ell = [nh_\ell]$ for $1 \leq \ell \leq k$, as soon as $n \geq \max_{1 \leq \ell \leq k} (h_\ell - h_{\ell-1})^{-1}$ where $h_0 = 0, 1 \leq s_1 < s_2 < \cdots < s_k < n$ and therefore,

$$\begin{aligned} f_{n,k}(h) &= \frac{\alpha^k \Gamma(\alpha + n - s_k)\Gamma(n)}{\Gamma(n + 1 - s_k)\Gamma(\alpha + n)} n^k \prod_{1 \leq \ell \leq k-1} (n - s_\ell)^{-1} \\ &= \alpha^k \frac{\Gamma(\alpha + n - s_k)}{n^{\alpha-1}\Gamma(1 + n - s_k)} \frac{n^\alpha \Gamma(n)}{\Gamma(\alpha + n)} \prod_{1 \leq \ell \leq k-1} (1 - \frac{s_\ell}{n})^{-1} \end{aligned}$$

Now, clearly, $\frac{s_\ell}{n} \uparrow h_\ell$ as $n \to \infty$, and if we recall that $\lim_{x \to \infty} \frac{\Gamma(\alpha + x)}{\Gamma(\beta + x)} x^{\beta - \alpha} = 1$ (see Abramovitz and Stegun (1964)), we obtain,

$$\lim_{n \to \infty} f_{n,k}(h) = \alpha^k (1 - h_k)^{\alpha - 1} \prod_{1 \leq \ell \leq k-1} (1 - h_\ell)^{-1}$$

because $n - s_k = n(1 - \frac{s_k}{n}) \geq n(1 - h_k) \to \infty$ as $n \to \infty$. Therefore $f_{n,k}(h) \to f_k(h)$ a.e. with respect to the Lebesgue measure. But, as $n \to \infty, \int_{I\!R^k} f_{n,k}(h)dh = 1 - P(M_n < k) \to 1$ and clearly, $f_k(h)$ is a consistant set of densities because $f_{k-1}(h_1, h_2, \cdots h_{k-1}) = \int_{I\!R} f_k(h_1, \cdots, h_k)dh_k$ and $f_1(h_1) = \alpha(1 - h_1)^{\alpha-1}$. So $\int_{I\!R^k} f_k(h)dh = 1$. Therefore, by Scheffé's theorem on the convergence of densities, $f_{n,k}(h) \to f_k(h)$ in L_1 and $E[f(H_1, H_2, \cdots, H_k)] = \int_{I\!R^k} f(h) f_k(h)dh$.

■

A simple change of variables in the result obtained in Proposition 3.13, will give the density of $(G_1, G_2, \cdots, G_k) \qquad \forall k \in I\!N_0$ because

$$G_k = H_k - H_{k-1} \qquad \forall k \in I\!N_0 \tag{3.51}$$

(if we set $H_0 = 0$).

Remark that in the proof of Proposition 3.13, we obtain the following partial result

$$G_1 \sim B(1, \alpha). \tag{3.52}$$

This can be easily extended in a very nice and simple description of the distribution of $(G_1, G_2, \cdots, G_k)$ for any k. This is our main result.

Proposition 3.14. If

$$U_k \;=\; \frac{G_k}{1-H_{k-1}} \qquad\qquad \forall k \in I\!N_0$$

which can be inverted in, either

$$G_k \;=\; U_k \prod_{1\le \ell \le k-1} (1-U_\ell) \qquad\qquad \forall k \in I\!N_0$$

or

$$H_k \;=\; 1 - \prod_{1\le \ell \le k} (1-U_\ell) \;\; \forall k \in I\!N_0$$

then

(i) $\underset{1\le k<\infty}{\perp\!\!\!\perp} \; U_k$

(ii) $U_k \;\sim\; B(1,\alpha)$

i.e. $U_k, k \in I\!N_0$, is an i.i.d. infinite sample from the $B(1,\alpha)$ distribution.

■

Proof: From Proposition (3.13), the density of G_k conditionnally on $(G_1, G_2, \cdots, G_{k-1})$ is given by

$$f(g|G_1, G_2, \cdots, G_{k-1}) \;=\; \frac{f_k(H_1, H_2, \cdots, H_{k-1}, H_{k-1}+g)}{f_{k-1}(H_1, H_2, \cdots, H_{k-1})}$$
$$=\; \alpha \frac{(1-H_{k-1}-g)^{\alpha-1}}{(1-H_{k-1})^{\alpha}} \;=\; \alpha(1-\frac{g}{1-H_{k-1}})^{\alpha-1}\;\frac{1}{1-H_{k-1}}$$

This shows that $(U_k|G_1, G_2, \cdots, G_{k-1}) \sim B(1,\alpha)$ and that $U_k \underset{}{\perp\!\!\!\perp} (G_1, G_2, \cdots, G_{k-1})$. The rest of the proof relies on theorem 7.6.3 of Florens, Mouchart, Rolin (1990).

■

Proposition 3.14 provides a so nice description of the distribution of the jumps sizes of the Dirichlet process, that it has a lot of implications which can be easily obtained by simple direct computation. We will restrict ourselfes to a few of these implications that we have arbitrarily considered as the most interesting.

First of all, the distribution of $(G_1, G_2, \cdots, G_k)$ is easily characterized, for any k, by its moments.

Corollary 3.15. $\forall k \in I\!N_0,\ \forall a_\ell \geq 0, 1 \leq \ell \leq k$

$$E[\prod_{1\leq \ell\leq k} G_\ell^{a_\ell}] \ = \ \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha+b_1)} \prod_{1\leq \ell\leq k} \frac{\Gamma(a_\ell+1)}{\alpha+b_\ell} \tag{3.53}$$

where $b_\ell = \sum_{\ell\leq r\leq k} a_r, \quad 1 \leq \ell \leq k$

In particular,

$$E[G_k^a] \ = \ \frac{\Gamma(\alpha)\Gamma(a+1)}{\Gamma(\alpha+a)} \ (\frac{\alpha}{\alpha+a})^k \tag{3.54}$$

■

Proof: Indeed, if $b_{k+1} = 0$

$$\prod_{1\leq \ell\leq k} G_\ell^{a_\ell} \ = \ \prod_{1\leq \ell\leq k} U_\ell^{a_\ell}(1-U_\ell)^{b_{\ell+1}}$$

Therefore

$$\begin{aligned} E[\prod_{1\leq \ell\leq k} G_\ell^{a_\ell}] \ &= \ \prod_{1\leq \ell\leq k} \alpha \frac{\Gamma(a_{\ell+1})\Gamma(\alpha+b_{\ell+1})}{\Gamma(\alpha+b_{\ell+1}+a_\ell+1)} \\ &= \ \alpha^k \prod_{1\leq \ell\leq k} \frac{\Gamma(a_\ell+1)}{\alpha+b_\ell} \cdot \frac{\Gamma(\alpha)}{\Gamma(\alpha+b_1)}. \end{aligned}$$

■

If the characterization of the distribution of G_k by (3.54) is simple, the density of G_k is however more complex as seen as in the next corollary.

Corollary 3.16. $\forall k \in I\!N_0$

$$\frac{d}{dh} P[H_k \leq h] \ = \ \frac{\alpha^k}{(k-1)!} \{\ell n(\frac{1}{1-h})\}^{k-1} (1-h)^{\alpha-1} 1_{[0,1]}(h) \tag{3.55}$$

$$\begin{aligned} \frac{d}{dg} P[G_k \leq g] \ &= \ \int_g^1 \frac{\alpha^k}{(k-2)!} \{\ell n(\frac{1}{s})\}^{k-2} \frac{(s-g)^{\alpha-1}}{s} ds \\ &= \ \alpha^k \Gamma(\alpha) \sum_{k-1\leq r<\infty} \frac{s_{r,k-1}}{\Gamma(\alpha+r)} (1-g)^{\alpha+r-1} \end{aligned} \tag{3.56}$$

This implies the peculiar identity

$$P[G_k > g] \;=\; E[(1-g)^{\alpha+I_{k-1}}]. \tag{3.57}$$

■

Proof: Integrating the densities given in Proposition 3.13 on the first $(\ell-1)$ coordinates gives

$$f_k(h_\ell, h_{\ell+1}, \cdots, h_k) \;= \frac{\alpha^k (1-h_k)^{\alpha-1}}{(\ell-1)!} \{\ell n(\frac{1}{1-h_\ell})\}^{\ell-1} \prod_{\ell \le j \le k-1} (1-h_j)$$
$$\text{on } \{0 < h_\ell < h_{\ell+1} \cdots < h_k < 1\}$$

Taking $\ell = k$ and $\ell = k-1$, and remembering that $H_k = G_k + H_{k-1}$, (3.55) and (3.56) are easily obtained. The second formula comes from the identity derived from the definition of $s_{n,k}$ (3.30)

$$\frac{1}{k!}\{\ell n(\frac{1}{1-z})\}^k \;= \sum_{k \le n < \infty} s_{n,k} \frac{z^n}{n!}$$

and (3.57) is derived from (3.34). ■

Identity (3.57) as well as (3.33) suggest a deep association between G_k and I_k. We have indeed the following result.

Proposition 3.17. $\forall k \in I\!N_0$
(i) $G_k \;\underline{\|}\; \mathcal{X}_1^{I_k} | I_k$
(ii) $G_k | I_k \;\sim\; B(1, \alpha + I_k - 1)$

Proof: As seen before (3.13), (3.6) and (3.39), $\forall a \in I\!R_0^+$

$$\begin{aligned} E[G_k^a | \mathcal{X}_1^{I_k}] \;&=\; E[F(X_{I_k})^a | \mathcal{X}_1^{I_k}] \\ &= \sum_{k \le n < \infty} E[F(X_{I_k})^a | \mathcal{X}_1^n] 1_{\{I_k = n\}} \\ &= \sum_{k \le n < \infty} E[F(X_n)^a | \mathcal{X}_1^n] 1_{\{I_k = n\}} \end{aligned}$$

Using (2.11) and (2.1) and the fact that $\hat{F}_n(X_n) = \frac{1}{n}$ on $\{I_k = n\}$, gives that $F(X_n) | \mathcal{X}_1^n \sim B(1, \alpha + n - 1)$ on $\{I_k = n\}$. Therefore,

$$E[G_k^a|\mathcal{X}_1^{I_k}] = \sum_{k\le n<\infty} \int_0^1 g^a(\alpha+n-1)(1-g)^{\alpha+n-2}dg \quad 1_{\{I_k=n\}}$$
$$= \int_0^1 g^a(\alpha+I_k-1)(1-g)^{\alpha+I_k-2}dg.$$

■

Using the obtained description of the Dirichlet process, we may give another proof of the identity obtained in Rolin (1992) (Proposition 4.3).

Corollary 3.18. Let $F \sim \mathcal{D}i(\alpha\mu)$ where μ is diffuse. Then for any non negative measurable function f on $(S\times[0,1], \mathcal{S}\otimes\mathcal{B}\cap[0,1])$ with $f(x,0)=0\ \forall x\in S$,

(3.58)
$$E\big[\sum_{x\in S} f[x,F(x)]\big]$$
$$= E\big[\sum_{1\le k<\infty} f(Z_k,G_k)\big]$$
$$= \int_{S\times[0,1]} f(x,u)\frac{\alpha(1-u)^{\alpha-1}}{u}\mu(dx)du.$$

■

Proof: Let $f(x,u)=f(x)u^a$ for $a\in\mathbb{R}_0^+$. Then,

$$E\big[\sum_{1\le k<\infty} f(Z_k)G_k^a\big]$$

$$= \int_S f(x)\mu(dx)\cdot\sum_{1\le k<\infty}\frac{\Gamma(\alpha)\Gamma(a+1)}{\Gamma(\alpha+a)}\big(\frac{\alpha}{\alpha+a}\big)^k$$

$$= \int_S f(x)\mu(dx)\ \frac{\Gamma(\alpha)\Gamma(a+1)}{\Gamma(\alpha+a)}\frac{\alpha}{a}$$

$$= \int_S f(x)\mu(dx)\alpha\frac{\Gamma(\alpha)\Gamma(a)}{\Gamma(\alpha+a)}$$

$$= \int_{S\times[0,1]} f(x)u^a\frac{\alpha(1-u)^{\alpha-1}}{u}\mu(dx)du.$$

It is then true for all measurable f by the Stone-Weierstrass and monotone class theorems. ∎

In the last section, we will show how the posterior distribution in nonparametric Bayesian statistical analysis may easily be simulated.

4. Simulation of the Dirichlet Process.

We first recall the result obtained in Rolin (1992). Let

$$F \sim \mathcal{D}i(\alpha\mu) \tag{4.1}$$

where μ is a diffuse probability on $(S, \mathcal{S})$ and $\{Y_i : 1 \leq i \leq n\}$ an i.i.d. sample from F, i.e.,

$$\underset{1\leq i\leq n}{\|} Y_i|F \tag{4.2}$$

$$Y_i|F \sim F. \tag{4.3}$$

Then, the posterior distribution of F is:

$$F|Y_1^n \sim \mathcal{D}i(\alpha\mu + n\hat{F}_n) \tag{4.4}$$

where

$$\hat{F}_n = \frac{1}{n}\sum_{1\leq i\leq n} \varepsilon_{Y_i}. \tag{4.5}$$

If K_n is the number of distinct values in $\{Y_i : 1 \leq i \leq n\}$ (see (3.3) and (3.4)), $\{T_k : 1 \leq k \leq K_n\}$ is the set of ordered distinct values (see (3.5) and (3.6)) and $K_{n,k}$ is the number of Y_i's equal to $T_k, 1 \leq k \leq K_n$, (see (3.7) and (3.18)), then

$$\hat{F}_n = \frac{1}{n}\sum_{1\leq k\leq K_n} K_{n,k}\varepsilon_{T_k}. \tag{4.6}$$

Now, according to Rolin (1992), if we define:

$$P_n \;=\; \sum_{1\le k\le K_n} F(T_k) \tag{4.7}$$

$$F_d(T_k) \;=\; P_n^{-1}F(T_k) \tag{4.8}$$

$$F_d \;=\; \sum_{1\le k\le K_n} F_d(T_k)\varepsilon_{T_k} \tag{4.9}$$

we can decompose F as

$$F \;=\; P_nF_d + (1-P_n)F_c \tag{4.10}$$

and we have

$$P_n \;\underline{\|}\; F_d \;\underline{\|}\; F_c|Y_1^n \tag{4.11}$$

$$P_n|Y_1^n \;\sim\; B(n,\alpha) \tag{4.12}$$

$$F_d|Y_1^n \;\sim\; \mathcal{D}i(n\hat{F}_n) \tag{4.13}$$

$$F_c|Y_1^n \;\sim\; \mathcal{D}i(\alpha\mu). \tag{4.14}$$

Now, (4.13) implies that

$$\{F_d(T_k) : 1\le k\le K_n\}|Y_1^n \;\sim\; Di[\{K_{n,k} : 1\le k\le K_n\}] \tag{4.15}$$

So if $\{W_i : 1 \leq i \leq n\}$ is an i.i.d. sample of the exponential distribution with parameter 1, by setting

$$F_d(T_1) = (\sum_{1\leq i\leq n} W_i)^{-1}(\sum_{1\leq i\leq K_{n,1}} W_i) \tag{4.16}$$

and so on ($2 \leq k \leq K_n$), we will simulate (4.15).

To simulate F_c, we will generate an i.i.d. sample $\{Z_i : 1 \leq i \leq N\}$ according to μ and an i.i.d. sample $\{U_i : 1 \leq i \leq N\}$ of the $B(1,\alpha)$ distribution. (Note that, if V_i is uniformly distributed on the unit interval, $1 - V_i^{\frac{1}{\alpha}} \sim B(1,\alpha)$). Finally, if

$$G_k = U_k \prod_{1\leq \ell\leq k-1} (1 - U_\ell) \tag{4.17}$$

we will simulate F_c by

$$F_c^N = \sum_{1\leq k\leq N} G_k \varepsilon_{Z_k}. \tag{4.18}$$

The error term is uniformly bounded. Indeed,

$$\sup_{B\in\mathcal{S}} |F_c(B) - F_c^N(B)| = \sum_{N+1\leq k<\infty} G_k = L_N. \tag{4.19}$$

Since $L_N = 1 - H_N$ (see (3.49)), its density is known (corollary 3.16),

$$\frac{d}{dx} P[L_N \leq x] = \frac{\alpha^N}{(N-1)!} (\ell n \frac{1}{x})^{N-1} x^{\alpha-1} 1_{[0,1]}(x). \tag{4.20}$$

In particular,

$$E(L_N) = (\frac{\alpha}{\alpha+1})^N \tag{4.21}$$

and

$$V(L_N) = (\frac{\alpha}{\alpha+1})^{2N} \big[\{1 + \frac{1}{\alpha(\alpha+2)}\}^N - 1\big]. \tag{4.22}$$

References

Abramowitz, M. and I.A. Stegun (1964), Editors, *Handbook of Mathematical Functions.* National Bureau of Standards. (Applied Mathematics Series 55).

Antoniak, Ch.E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *Annals of Statistics*, 2, 1152-1174.

Blackwell, D. and J.B. MacQueen (1973), "Ferguson distributions via Polya urn schemes." *Annals of Statistics*, 1, 353-355.

Cifarelli, D.M. and E. Regazzini (1990), "Distribution functions of means of a Dirichlet process." *Annals of Statistics*, 18, 429-442.

Dellacheric, C. and P.A. Meyer (1980), *Probabilité et Potentiel (B): Théorie des Martingales*, Paris: Hermann.

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

Ferguson, T.S. (1974), "Prior distributions on spaces of probability measures." *Annals of Statistics*, 2, 615-629.

Florens, J.-P., M. Mouchart and J.-M. Rolin (1990), *Elements of Bayesian Statistics.* New-York: Marcel Dekker.

Neveu, J. (1972), *Martingales en temps discret.* Paris: Masson.

Rolin, J.-M. (1992), "Some useful properties of the Dirichlet process." Institut de Statistique, Université Catholique de Louvain, Discussion Paper No 9202.

Wilks, S.S. (1962), *Mathematical Statistics.* New York: Wiley.

Yamato, H. (1984). "Characteristic functions of means of distributions chosen from a Dirichlet process." *Annals of Probability*, 12, 262-267.

Yamato, H. (1984). "Properties of samples from distributions chosen from a Dirichlet process." *Bulletin of Informatics and Cybernetics*, 21, 77-83.