An exact predictive recursion for Bayesian nonparametric analysis of incomplete data[*]

Ubaldo Garibaldi

IMEM-CNR, via Dodecaneso 33, 16146, Genova (Italy)
e-mail: garibaldi@fisica.unige.it

and

Paolo Viarengo

University of Naples "Federico II", Naples (Italy)

Abstract: This paper presents a new derivation of nonparametric distribution estimation with right-censored data. It is based on an extension of the predictive inferences to compound evidence. The estimate is recursive and exact, and no stochastic approximation is needed: it simply requires that the censored data are processed in decreasing order. Only in this case the recursion provides exact posterior predictive distributions for subsequent samples under a Dirichlet process prior. The resulting estimate is equivalent to the Susarla-VanRyzin estimator and to the beta-Stacy process.

Contents

1. Preliminaries on Dirichlet processes with incomplete data

An important extension of Bayesian nonparametric theory is the treatment of incomplete data as they involve indirect observations or latent variables. The first completely Bayesian approach to the problem of dealing with observations censored on the right was made by Susarla and Van Ryzin [15] who used a Dirichlet process [5, 2] as a prior for the random distribution F and obtained in closed form the mean of the posterior distribution of F given the data. Blum and Susarla [3] complemented this result by showing that the conditional survival function turns out to be a Dirichlet processes mixture, as considered by Antoniak [1] with specified transition and mixing measures.

In the class of models under consideration, instead of the "precise, microscopic" values of the random sample $t_1, \ldots, t_n$ i.i.d $\sim F$, the unknown distribution, the evidence is given by less precise "fuzzy, macroscopic" observations. That is the t's are not directly observed; instead we only know that $t_1 \in A_1, \ldots, t_n \in A_n$, where the A's are subsets of $\mathbb{R}$. In the case of right-censoring, A_i is a singleton if t_i is uncensored, while $A_i = (c_i, \infty)$ is an interval containing $t_i > c_i$, where c_i is the censoring time.

In the case of "precise" evidence, standard Bayesian calculations are induced by placing a Dirichlet process prior distribution with parameter α on F, denoted D_α. This prior is indexed by a positive finite measure $\alpha(\cdot)$ defined on the range Ω of t_i. The base distribution $F_0(\cdot) = \frac{\alpha(\cdot)}{\alpha(\Omega)}$ is obtained by normalizing the measure $\alpha(\cdot)$ having total mass $\alpha(\Omega)$, which in turn measures the strength of the prior belief. The meaning of F_0 is that of initial (predictive) probability on the first observation, that is $\mathbb{P}(t_1 \in B) = F_0(B)$ for a measurable set $B \subset \Omega$. Moreover F_0 can be seen as a prior guess about the unknown distribution F, as $F_0(\cdot) = \mathbb{E}[F(\cdot)]$ is the expectation of F with regards to the prior D_α. Conditional on the precise data $D_p = \{t_i\}$, the posterior mean $F_n(\cdot) = \mathbb{E}[F(\cdot)|D_p]$ of the random probability measure F, is the expectation with regards to the posterior $D_{\alpha+\sum \delta_{t_i}}$, and its meaning is the posterior predictive distribution $\mathbb{P}(t_{n+1} \in \cdot|D_p) = F_n(\cdot) = (\alpha(\cdot) + \sum \delta_{t_i}(\cdot))/(\alpha(\Omega) + n)$.

In the case of "compound evidence", that is conditional on the data $D_c = \{A_i\}$, the posterior results to be a mixture of Dirichlet processes, which is cumbersome [14]. (This type of mixture comes from the incomplete data, and it is different in principle from those mixtures with a Dirichlet process mixing distribution, which have been studied only with the advent of Monte Carlo Markov chain (among others [8, 9]) and related solutions, such as sequential importance sampling [13].

In order to escape the complication of mixtures induced by incomplete data, one can use the alternative and powerful approach of Blackwell and MacQueen [2], which have shown that the extension of the Polya updating rule to a continuous space, based only on the predictive distribution, is equivalent to Ferguson's setting. In this framework the predictive distributions $F_n(\cdot) = F_0(\cdot|D_p)$ are easy to calculate, and F_n is a simple recursive function of F_{n-1}. Extending this approach to compound evidence, Newton et al. [10, 11, 12] proposed a

class of computationally efficient methods to approximate mixtures of Dirichlet processes, involving a recursive algorithm. The advantage is that the predictive uncertainty can be encoded in a single measure approximating Bayes estimator in general, and avoiding the high computational cost required for both an exact solution and the Markov chain Monte Carlo calculations. Unfortunately the typical problem of Newton's estimate is that $\widehat{F}_n(\cdot)$ depends on the order in which data are processed, i.e. the recursive algorithm is order dependent. Hence one can approximate the final result averaging on (a sample of) all permutations. A recent paper by Martin and Ghosh [7] explores the above-said fast recursive algorithm as a special case of stochastic approximation.

In the present note, in the right censoring case $A_1 \subset \cdots \subset A_n$, we prove that no stochastic approximation is necessary, as there is a sole rational pattern which recursively produces the exact posterior predictive distributions for subsequent samples under a Dirichlet process prior.

By "exact" we mean that our result coincides with the Susarla- Van Ryzin estimator under squared error loss.

1.1. On $n = 1$

On the basis of a single observation, the Dirichlet process implies that the (exact) updating rule for the measure is linear:

$$\alpha(\cdot|t_1) = \alpha(\cdot) + \delta_{t_1}(\cdot) \tag{1}$$

and the associated predictive probability is

$$\mathbb{P}(t_2 \in B|t_1) = \frac{\alpha(B|t_1)}{\alpha(\Omega|t_1)} = \frac{\alpha}{1+\alpha}F_0(B) + \frac{1}{1+\alpha}1_B(t_1). \tag{2}$$

The extension to missing data, posing $\alpha_1(B) := \alpha(B|t_1 \in A_1)$, is given by:

$$\begin{aligned} \alpha_1(B) &= \alpha(B) + \sum_{t_1 \in A_1} 1_B(t_1)\mathbb{P}(t_1|A_1) \tag{3} \\ &= \alpha(B) + \mathbb{P}(t_1 \in B|A_1) \\ &= \alpha(B) + \frac{\alpha(B \cap A_1)}{\alpha(A_1)} \tag{4} \\ \alpha_1(\Omega) &= \alpha + 1 \end{aligned}$$

where $\alpha_1(B)$ is determined as the expectation value of $\alpha(B|t_1)$ weighted by $\mathbb{P}(t_1|A_1)$, see Appendix A. (We reserve the subscript n to measures and probabilities conditioned to observables $A_1, \ldots, A_n$).

The associated probability is

$$F_1(B) := \mathbb{P}(t_2 \in B|A_1) = \frac{\alpha_1(B)}{\alpha + 1} = \frac{\alpha}{1+\alpha}F_0(B) + \frac{1}{1+\alpha}\mathbb{P}(t_1 \in B|A_1). \tag{5}$$

This result was first derived in [1] where missing data are analyzed in the field of mixtures of Dirichlet processes. It is interesting that Bayes' rule enters in the

second term of (4) through probability calculus. The corresponding posterior calculations are given by the initial F_0 on t_1.

The usual Polya scheme (2) is contained in (5) if A_1 reduces to t_1. If the information is complete, it implies the addition of a unit mass concentrated at the observed value, which coincides with the discreteness property of F. If the information is not complete, in general one needs to add an entire probability distribution to the urn, whose support is the observed A_1. In this case F is modified proportionally to its value.

1.2. Newton's approximate recursive algorithm

After precise evidence $t_1, \ldots, t_n$ the updating rule for the measure is

$$\alpha(t_{n+1} \in B | t_1, \ldots, t_n) = \alpha(B) + \delta_{t_1}(B) + \cdots + \delta_{t_n}(B) \tag{6}$$

which can be easily transformed into a recursive algorithm

$$\alpha(t_{n+1} \in B | t_1, \ldots, t_n) = \alpha(B | t_1, \ldots, t_{n-1}) + \delta_{t_n}(B). \tag{7}$$

Does the idea that the predictive uncertainty can be encoded in a single measure, and that learning occurs by adding measure, hold also for missing data? Thus would lead us to the following recursive generalization of (4)

$$\begin{aligned}
\alpha_n(t_{n+1} \in B) &= \alpha_{n-1}(B) + \mathbb{P}_{n-1}(t_n \in B | A_n) \tag{8} \\
F_n(t_{n+1} \in B) &= \frac{\alpha + n - 1}{n + \alpha} F_{n-1}(B) + \frac{1}{n + \alpha} \mathbb{P}_{n-1}(t_n \in B | A_n)
\end{aligned}$$

approximating uncertainty in F with a single Dirichlet process centered at F_{n-1}, on the basis of the well-known Polya sequence characterization. Again Bayes' rule enters in the second term taking F_{n-1} as the updated prior distribution for t_n. As constructed $F_n(B)$ is not the exact posterior predictive distribution for t_{n+1} in general, except the case $n = 1$. It will be only the case of no information loss, that is when $A_i = t_i$, and (8) and (6) coincide. In facts the value depends on the order in which $A_1, \ldots, A_n$ are processed.

The main hindrance of this recursion method is the dependence of the result upon the data order. The suggested algorithm in [10, 11, 12] implies to process data in some order through (8), to re-calculate it over a permutation of the data and average the results. Does a privileged order, providing an exact result, exist?

2. An exact recursive algorithm

If no information is lost the updating of $F_0(\cdot)$ and $\alpha(\cdot)$ after n data is given as follows, since the added measure is a point mass at t_i for each datum, unconditioned to any other quantity:

$$\alpha(t_{n+1} \in B | t_1, \ldots, t_n) = \alpha(B) + \delta_{t_1}(B) + \cdots + \delta_{t_n}(B) \tag{9}$$

In fact $\delta_{t_i}(\cdot)$ is a point mass distribution concentrated in t_i.

Otherwise if there is information loss, that is $D_c = (t_1 \in A_1) \cap \cdots \cap (t_n \in A_n)$, $A_i \subset \Omega$, and the exact generalization of (4) is, with $\alpha_n(B) := \alpha(B|D_c) = \alpha(B|A_1, \ldots, A_n)$:

$$\alpha_n(B) = \alpha(B) + \mathbb{E}[\sum_{i=1}^{n} 1_B(t_i)|D] \tag{10}$$

$$= \alpha(B) + \sum_{t_1, \ldots, t_n \in D_c} \left(\sum_{i=1}^{n} 1_B(t_i)\right) \mathbb{P}(t_1, \ldots, t_n|A_1, \ldots, A_n) \tag{11}$$

where $\mathbb{P}$ is based on the initial probability $\alpha(\cdot)/\alpha(\Omega)$.

Equation (11) holds for all cases of predictive inferences based on uncertain or indirect evidence. If Ω is finite the predictive probability can be exactly computed, but in general an exact recursive method does not exist [10]. The computation is easy to perform, although its length is rapidly increasing as a function of $\#\Omega$. One must compute the probability of all unobserved patterns $(t_1, \ldots, t_n)$, that are the elementary events compatible with the evidence, by means of the initial measure $\alpha(\cdot)$, and then normalize to the evidence, obtaining thus the joint conditional distribution $\mathbb{P}(t_1, \ldots, t_n|A_1, \ldots, A_n) = \frac{\mathbb{P}(t_1, \ldots, t_n)}{\mathbb{P}(A_1, \ldots, A_n)}$ that rules (11). Hence the empirical weights result as an average on the probability of any pattern. If Ω is continuous this procedure is impossible, as in (11) the sum is substituted by integrals.

Fortunately the case of right-censoring is peculiar, because the compound events $\{A_i\}$ are all of the same kind, that is they are events in the tail. We can show that (11) can be represented in the form (8) in an exact way, if the data are reordered and processed so that $A_1 \subset \cdots \subset A_n$, that is $A_i = (c_i, \infty)$, and $c_1 \geq \cdots \geq c_n$. Recall that we are speaking about the order in which data are *processed*, not the order in which they are *collected*, which is inessential due to the exchangeability property of the process.

To this end we re-write (10) in the following form, interchanging $\mathbb{E}$ and $\sum$;

$$\alpha_n(B) = \alpha(B) + \sum_{i=1}^{n} \mathbb{P}(t_i \in B|D_c) \tag{12}$$

where $\sum_{i=1}^{n} \mathbb{P}(t_i \in B|D_c)$ is the expected number of unobserved events that happen in B conditioned to D_c.

2.1. On $n = 2$

Given two compound events $D_c = A_1, A_2$, and $A_1 \subset A_2$, that is $c_1 \geq c_2$, (12) implies:

$$\alpha_2(B) =: \alpha(B|A_1, A_2) = \alpha(B) + \mathbb{P}(t_1 \in B|D_c) + \mathbb{P}(t_2 \in B|D_c). \tag{13}$$

Let us denote $t_i \in B$ with B_i. From Appendix B, for $(B_1 \cap A_1) \subset A_1 \subset A_2$, we have that $\mathbb{P}(B_1|A_1, A_2) = \mathbb{P}(B_1|A_1)$; while $\mathbb{P}(B_2|A_1, A_2) = \mathbb{P}_1(B_2|A_2)$.

Hence substituting in (13) we have

$$\alpha_2(B) = \alpha(B) + \mathbb{P}(B_1|A_1) + \mathbb{P}_1(B_2|A_2) = \alpha_1(B) + \mathbb{P}_1(B_2|A_2), \qquad (14)$$

that is exactly the (8) for $n = 2$.

2.2. On $n = n$

From (12), posing $D_n = A_1, \ldots, A_n$, we pass from $\alpha_n(B) = \alpha(B) + \sum_{i=1}^{n} \mathbb{P}(B_i|D_n)$ to $\alpha_{n+1}(B) = \alpha(B) + \sum_{i=1}^{n+1} \mathbb{P}(B_i|D_{n+1})$ adding A_{n+1} in all the evidence, and adding the term $\mathbb{P}(t_{n+1} \in B|D_{n+1})$.

If A_{n+1} contains all previous evidence (that is c_{n+1} precedes all other censoring times), then $\mathbb{P}(B_i|D_n, A_{n+1}) = \mathbb{P}(B_i|D_n)$. The first n terms are just $\alpha_n(B)$. The last term is $\mathbb{P}(B_{n+1}|D_n, A_{n+1}) = \frac{\mathbb{P}(B_{n+1} \cap A_{n+1}|D_n)}{\mathbb{P}(A_{n+1}|D_n)} = \mathbb{P}_n(B_{n+1}|A_{n+1})$.

Hence $\alpha_{n+1}(B) = \alpha_n(B) + \mathbb{P}_n(B_{n+1}|A_{n+1})$. See also Appendix C.

The solution of the recursion, that is the final predictive probability, is equivalent to the Susarla-Van Ryzin estimator [15]. Moreover it appears as a predictive generalization of the alternative method to Efron's algorithm proposed by Dinse [4].

3. Example

We briefly illustrate the algorithm on the dataset used by Susarla and Van Ryzin [15], which is nothing else the "classical" dataset used in the original paper by Kaplan and Meier [6] and re-used in many successive works (see f.i. Walker and Muliere [16, 17], whose approach, based on the beta-Stacy process and the right-neutrality, is predictive, but not recursive). The data are $0.8, 1.0+, 2.7+, 3.1, 5.4, 7.0+, 9.2, 12.1+$, where a "+"denotes a censored observation. Our illustration is intended to make a numerical comparison of our algorithm with the exact results obtained by Susarla and Van Ryzin. They took F_0 to be the exponential distribution with mean $1/0.12$, and for $\alpha(\mathbb{R}^+)$ they considered three cases, $\alpha(\mathbb{R}^+) = 4$, 8 and 16. They obtained the mean of the posterior distribution of F, under a squared error loss, as t ranges over the eight censored and uncensored points. For the sake of maximizing discrepancies we make our comparison only for the case $\alpha(\mathbb{R}^+) = 4$.

The recursion proceeds as follows: first of all the four precise data are processed, the order being inessential. Introducing the cumulative weight function $G(t) = \alpha(0, t)$, from (7) one gets for $n = 1, \ldots, 4$

$$G_n(t) = G_{n-1}(t) + \Theta(t - t_n), \qquad (15)$$

where Θ is Heaviside's function. The Survival function after the precise data is given by $S_4(t) = 1 - G_4(t)/G_4(\infty)$, see Fig. 1. Restarting from $G_4(t)$, the recursive equation for $G_n(t)$ is:

$$G_n(t) = G_{n-1}(t) + \Theta(t - t_n)\frac{G_{n-1}(t) - G_{n-1}(t_n)}{G_{n-1}(\infty) - G_{n-1}(t_n)}, \qquad (16)$$

for $n = 5, \ldots, 8$, and the final Survival function is $S_8(t) = 1 - G_8(t)/G_8(\infty)$, see Fig. 1 again.
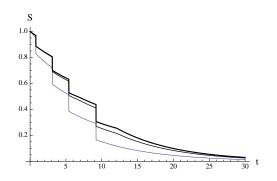
FIG 1. *The thin gray graph is $S_4(t)$, conditioned on the dead only; the two in black are $S_8(t)$, conditioned to the losses too. The thicker curve, obtained processing censored data in decreasing order, is the correct one; the other curve is obtained processing them in increasing order.*

TABLE 1

Maximum differences between the correct Survival function and the permuted ones. The table shows the value and the abscissa of the maximum difference. All permuted survival functions are dominated by the correct one, so that any average on them produces a systematic error on the final estimate. This deviation tends to increase with the lag of the last observed value from the first place, and (conditioned to it) with the the lag of the second-last one, and so on

MaxDiff	Abscissa	Order
0.00067	2.7	{12.1,7.0,1.0,2.7}
0.00067	2.7	{7.0,12.1,1.0,2.7}
0.00536	7.0	{12.1,2.7,7.0,1.0}
0.00584	7.0	{12.1,1.0,7.0,2.7}
0.00973	7.0	{12.1,2.7,1.0,7.0}
0.01019	7.0	{12.1,1.0,2.7,7.0}
0.01518	12.1	{1.0,12.1,7.0,2.7}
0.01581	12.1	{2.7,12.1,7.0,1.0}
0.01740	12.1	{1.0,12.1,2.7,7.0}
0.01803	12.1	{2.7,12.1,1.0,7.0}
0.01827	12.1	{7.0,12.1,2.7,1.0}
0.02746	12.1	{7.0,1.0,12.1,2.7}
0.02757	12.1	{2.7,1.0,12.1,7.0}
0.02772	12.1	{1.0,2.7,12.1,7.0}
0.02777	12.1	{7.0,2.7,12.1,1.0}
0.02988	12.1	{1.0,7.0,12.1,2.7}
0.03024	12.1	{2.7,7.0,12.1,1.0}
0.03500	12.1	{7,2.7.0,1.0,12.1}
0.03516	12.1	{7.0,1.0,2.7,12.1}
0.03750	12.1	{2.7,7.0,1.0,12.1}
0.03763	12.1	{1.0,7.0,2.7,12.1}
0.03943	12.1	{2.7,1.0,7.0,12.1}
0.03960	12.1	{1.0,2.7,7.0,12.1}

The result of our recursive algorithm is exactly the same if censored data are processed in decreasing order. Any permutation of this privileged order produces appreciable deviations, and the maximum deviation is obtained by the reversed order (Fig. 1). These deviations increase in the limit $\alpha(\mathbb{R}^+) \to 0$, where one obtains the Kaplan-Meier estimate in the empirical range. All "permuted" Survival functions are dominated by the correct one, so that any average on them produces a systematic error on the final estimate. Table 1 shows the value and
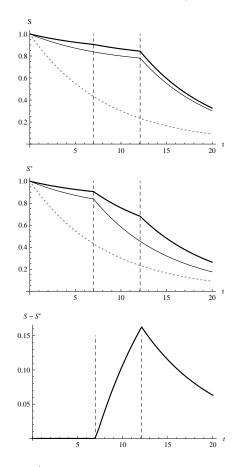
FIG 2. *The same prior, $\alpha(\mathbb{R}^+) = .4$, and only two censored data, 7.0+ and 12.1+. In the first two graphs the dotted line denotes the prior, the thin black line denotes the first updated Survival function, and the thick one the second (final) one. At each new-processed datum a cusp appears. The first graph describes S obtained in the correct order, that is $\{12.1+, 7.0+\}$, the second S^* in the reversed one. The third plot magnifies the difference between the two final estimates. The difference is null in the domain $[0, 7.0]$, it is maximum at the end of the data domain.*

the abscissa of the maximum difference between the correct Survival function and the "permuted" one is reported. This deviation roughly increases with the lag of the last observed value from the first place, and (conditioned to it) with the the lag of the second-last one, and so on.

To magnify the effect of permutation on the recursion, we consider the simplified case of Fig. 2, where one considers the same prior, $\alpha(\mathbb{R}^+) = .4$, and only two censored data, 7.0+ and 12.1+. In all the first two graphs the dotted line denotes the prior, the thin black line denotes the first updated Survival function, and the thick one the second (final) one. At each new-processed datum a cusp appears. The first updated one contains one cusp, the final contains two cusps. The first graph describes S obtained in the correct order, that is $\{12.1+, 7.0+\}$,

the second S^* in the reversed one. The third graph magnifies the difference between the two final Survival functions. The difference is null in the domain $[0, 7.0]$, it is maximum at the end of the data domain.

Acknowledgements

Appendixes

These appendixes show the demonstrations of the three statements on $n = 1, 2$ and n.

Appendix A. Considering that in the Dirichlet process $F_1(\cdot) = \mathbb{E}[F(\cdot)|t_1]$, i.e. it is the expectation of F taken over the Dirichlet conditioned to t_1, whose weight is $\alpha(\cdot|t_1) = \alpha(\cdot) + \delta_{t_1}(\cdot)$, its generalization to compound evidence is straightforward, due to probability calculus: $\mathbb{E}[F(\cdot)|A_1] = \int_{\mathbf{t_1} \in \mathbf{A_1}} \mathbb{E}[F(\cdot)|t_1]\mathbb{P}(t_1|A_1)dt_1 = \mathbb{E}[\mathbb{E}[F(\cdot)|t_1]|A_1]$. The generalization to n compound events reads

$$F_n(\cdot) = \mathbb{E}[\mathbb{E}[F(\cdot)|t_1, \ldots, t_n]|D_c],$$

where $D_c = (t_1 \in A_1) \cap \cdots \cap (t_n \in A_n)$

Appendix B. Suppose that $A_1 \subset A_2$. Now

$$\mathbb{P}(t_1 \in B|t_1 \in A_1, t_2 \in A_2) = \frac{\mathbb{P}(t_1 \in (B \cap A_1)|t_2 \in A_2)}{\mathbb{P}(t_1 \in A_1|t_2 \in A_2)}$$
$$= \frac{\alpha(t_1 \in (B \cap A_1)|t_2 \in A_2)}{\alpha(t_1 \in A_1|t_2 \in A_2)}.$$

As in the domain $t \in A_2$ it is true that $\alpha(t|t_2 \in A_2) = \alpha(t) + \mathbb{P}(t|A_2) = \alpha(t)\{1 + \frac{1}{\alpha(A_2)}\}$, and both $B_1 \cap A_1$ and A_1 lay in this region, it follows that they both increase their probability in the same ratio, and

$$\frac{\alpha(t_1 \in (B \cap A_1)|t_2 \in A_2)}{\alpha(t_1 \in A_1|t_2 \in A_2)} = \frac{\alpha(t_1 \in B \cap A_1)}{\alpha(t_1 \in A_1)}.$$

Then

$$\mathbb{P}(t_1 \in B|t_1 \in A_1, t_2 \in A_2) = \mathbb{P}(t_1 \in B|t_1 \in A_1).$$

Denoting $t_i \in B$ with B_i, in short we have $\mathbb{P}(B_1|A_1, A_2) = \mathbb{P}(B_1|A_1)$. It means that $t_2 \in A_2$ is "irrelevant" to $t_1 \in B|t_1 \in A_1$. In words A_1 "screens off" $t_1 \in B$ from any further evidence less precise than A_1. On the contrary a further more

precise evidence $t_1 \in A_1$ is relevant for $\mathbb{P}(t_2 \in B | t_2 \in A_2)$, so that A_1 in $\mathbb{P}(B_2 | A_1, A_2)$ cannot be neglected. Given that $\mathbb{P}_1(\cdot) := \mathbb{P}(\cdot | A_1)$, we have:

$$\mathbb{P}(B_2 | A_1, A_2) = \frac{\mathbb{P}(B_2 \cap A_2 | A_1)}{\mathbb{P}(A_2 | A_1)} = \frac{\mathbb{P}_1(B_2 \cap A_2)}{\mathbb{P}_1(A_2)} = \mathbb{P}_1(B_2 | A_2).$$

Appendix C. Using the screening of property induced by $A_1 \subset \cdots \subset A_n$ we have $\alpha_n(B) = \alpha(B) + \sum_{i=1}^{n} \mathbb{P}(B_i | D_n) = \alpha(B) + \mathbb{P}(B_1 | A_1) + \mathbb{P}(B_2 | A_2, A_1) + \cdots + \mathbb{P}(B_i | A_i, A_{i-1}, \ldots A_1) + \cdots + \mathbb{P}(B_n | A_n, \ldots, A_1)$.

Now $\alpha(B) + \mathbb{P}(B_1 | A_1) = \alpha_1(B)$, and one can omit A_1 in all remaining evidence, substituting $\mathbb{P}_1$ to $\mathbb{P}$. Hence $\alpha_n(B) = \alpha_1(B) + \mathbb{P}_1(B_2 | A_2) + \cdots + \mathbb{P}_1(B_i | A_i, A_{i-1}, \ldots A_2) + \cdots + \mathbb{P}_1(B_n | A_n, \ldots, A_2)$.

Going on with the same algorithm, if $A_1 \subset \cdots \subset A_n$ then $\alpha_n(B) = \alpha_{n-1}(B) + \mathbb{P}_{n-1}(B_n | A_n)$. QED.

References

[1] ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1552–1174. MR0365969

[2] BLACKWELL D., MACQUEEN J.B. (1973). Ferguson distributions via polya urn schemes, *Ann. Statist.* **1**, 353–355. MR0362614

[3] BLUM, K., SUSARLA, V. (1977). On the posterior distribution of a Dirichlet process given randomly right censored data. *Stoch. Proc. Appl.* **5**, 207–211. MR0474604

[4] DINSE, G. E. (1985). An alternative to Efron's redistribution-of-mass construction of the Kaplan–Meier estimator. *Amer. Statist.* **39**, 299–300.

[5] FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230. MR0350949

[6] KAPLAN, E. L, MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–481. MR0093867

[7] MARTIN, R., GHOSH, K. (2008). Stochastic approximation and Newton's estimate of a mixing distribution. *Stat. Sci.* **23**, 365–382. MR2483909

[8] MACEACHERN, S. N., MULLER, P. (1998). Estimating mixture of Dirichlet process models. *J. Comp. Graph. Statist.* **7**, 223–238.

[9] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, 249–265. MR1823804

[10] NEWTON, M. A., QUINTANA, F. A., ZHANG Y. Nonparametric Bayes methods using predictive updating. In DAY, D., MULLER, P., SINHA, D. (eds.) (1998). *Practical nonparametric and semiparametric Bayesian statistics*, pp. 45–61. Springer, New York. MR1630075

[11] NEWTON, M. A., ZHANG, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86**, 15–26. MR1688068

[12] NEWTON, M. A. (2002). A nonparametric recursive estimator of the mixing distribution. *Sankhya* Ser. A **64**, 306–322. MR1981761

[13] QUINTANA, F. A. (1998). Nonparametric Bayesian analysis for assessing homogeneity in k × 1 contingency tables with fixed right margin totals. *J. Amer. Statist. Assoc.* **93**, 1140–1149. MR1649208

[14] MELILLI, E., PETRIS, G. (1995) Bayesian inference for contingency tables with given marginals. J. Ital. Stat. Soc. **4**, 215–233.

[15] SUSARLA, V., VAN RYZIN, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897–902. MR0436445

[16] WALKER, S., MULIERE, P. (1997). Beta-Stacy processes and a generalization of the Polya-urn scheme. *Ann. Statist.* **25**, 1762–1780. MR1463574

[17] WALKER, S., MULIERE, P. (2000). Neutral to the right processes from a predictive perspective: a review and new developments. *Metron* **58** (3/4), 13–30. MR1835012