Statistical Science
2016, Vol. 31, No. 1, 30–33
DOI: 10.1214/15-STS538
Main article DOI: 10.1214/15-STS529

Relatives of the Ewens Sampling Formula in Bayesian Nonparametrics

Stefano Favaro and Lancelot F. James

Abstract. We commend Harry Crane on his review paper which serves to not only point out the ubiquity of the Ewens sampling formula (ESF) but also highlights some connections to more recent developments. As pointed out by Harry Crane, it is impossible to cover all aspects of the ESF and its relatives in the pages generously provided by this journal. Our task is to present additional commentary in regards to some, perhaps not so well-known, related developments in Bayesian noparametrics.

Key words and phrases: Age ordered ESF, Bayesian nonparametrics, posterior ESF, spatial neutral to the right process, species sampling problem.

1. INTRODUCTION

Let $X_1, \ldots, X_n | P$ be independent and identically distributed as P, where P is an unknown probability measure $P(A) = \Pr(X_1 \in A)$ with A being a subset of a measurable space $\mathcal{X}$. Despite the complexity of $\mathcal{X}$, the classic nonparametric estimator of P is given by $P_n(x) = n^{-1} \sum_{1 \le i \le n} \delta_{X_i}(x)$. Bayesian nonparametric statistics has its origins in David Blackwell's desire in the late 1960s to find appropriate Bayesian solutions to this problem, that is, to find priors $\mathscr{P}(\mathrm{d}P)$ over the space of distributions that lead to posterior distributions $\mathscr{P}(\mathrm{d}P | X_1, \ldots, X_n)$ that are tractable and ideally mimic nice properties of their frequentist counterparts. The Dirichlet process (DP) by Ferguson [6] was the first and still most prominent solution to this problem, producing a posterior distribution that is again a DP. Doksum [3] soon followed with neutral to the right (NTR) priors defined as random distribution functions over the real line $\mathcal{R}$, and showed that the posterior was also a NTR process. Doksum also showed that when $\mathcal{X} = \mathcal{R}$ the DP arises as a special case of NTR priors. These early works focused on the statistical problem where it was assumed that $\mathbf{X} = (X_1, \ldots, X_n)$ comes from a true distribution $\tilde{P}$. Furthermore, it was often assumed that $\tilde{P}$ was a nonatomic distribution, meaning there are no ties among the X_i's. So while it was of interest to calculate the Bayes estimate, which is the prediction rule $Q_n(\mathrm{d}x) = \mathbb{E}[P(\mathrm{d}x)|\mathbf{X}]$, its behavior is evaluated relative to $\tilde{P}$. That is to say, from the point of view of such statistical problems, the combinatorial aspects of the joint exchangeable distribution $M_n(\mathrm{d}\mathbf{x}) = \mathbb{E}[\prod_{1 \le i \le n} P(\mathrm{d}x_i)]$ was not of primary interest.

In the case of the DP, with prior law $\mathscr{P}_{0,\theta}(\mathrm{d}P|H)$, $M_n(\mathrm{d}\mathbf{x}) = M_n(\mathrm{d}\mathbf{x}|\theta)$ is the Blackwell–MacQueen Pólya urn distribution in Blackwell and MacQueen [2] which produces the Chinese restaurant process (CRP) and the Ewens' sampling formula (ESF). The result by Blackwell and MacQueen [2] shows that if the empirical measure P_n is based on $\mathbf{X}$ sampled from $M_n(\mathrm{d}\mathbf{x}|\theta)$, then as $n \to \infty$ it converges to a DP with law $\mathscr{P}_{0,\theta}(\mathrm{d}P|H)$. The problem considered in Antoniak [1] involves placing a prior distribution $\pi(\mathrm{d}\theta)$ on θ, which necessitates the involvement of $M_n(\mathrm{d}\mathbf{x}|\theta)$ and gives impetus to Antoniak's independent derivation of the ESF. The reader is referred to Proposition 3 in Antoniak [1] for a detailed account. The model in Lo [12] represents the prototype for the modern usage of Bayesian hierarchical mixture models where $\mathbf{X}$ are viewed as latent variables drawn from $M_n(\mathrm{d}\mathbf{x}|\theta)$. The approach proposed by Lo [12] incorporates Fubini's theorem which expresses the joint distribution of $(P, \mathbf{X})$ in terms of $P|\mathbf{X}$, and $\mathbf{X} \sim M_n(\mathrm{d}\mathbf{x}|\theta)$. Furthermore, Lemma 2 in Lo [12] establishes characteri-

Stefano Favaro is Associate Professor, Department of Economics and Statistics, Corso Unione Sovietica 218/bis, 10134 Torino, Italy (e-mail: stefano.favaro@unito.it). Lancelot F. James is Professor, Department of Information Systems, Business Statistics and Operations Management, Clear Water Bay, Kowloon, Hong Kong (e-mail: lancelot@ust.hk).

zations via the partition of $[n] = \{1, \ldots, n\}$ induced by $\mathbf{X}$ and hence the CRP. Nonetheless, the combinatorial complexity involved in such characterizations limited their practical usage at that time.

The purpose of our exposition so far was to note that in the 1970s and 1980s the primary focus in Bayesian nonparametrics was on priors over spaces of measures and the somewhat limited employment of combinatorial structures such as the CRP/ESF. The field has changed in a dramatic fashion since Ferguson [6]. Due in large part to advances in computer technology, Bayesian nonparametric ideas are now being applied to tackle a wide range of statistical problems where parametric assumptions are often infeasible. In particular, many applications exploit the flexible modeling features exhibited by the ESF/CRP, and generalized M_n, as applied to intricate missing data problems. The availability of explicit probability distributions for these mechanisms provides formal generative processes that allow one to learn model structure in the presence of incoming data via the standard Bayesian updating mechanism. For instance, generalized notions of the CRP provide natural priors on the space of partitions/clusters/groups that grow with the sample size. Of note are some recent developments related to problems arising out of Bayesian machine learning; see, for example, Teh and Jordan [14] and references therein. Rather than recount points that are generally familiar to a more specialized Bayesian nonparametric audience, in the next two sections we focus on less well-known, but highly pertinent, connections between the ESF and its relatives and models arising in Bayesian nonparametrics.

2. AGE-ORDERED ESF AND SPATIAL NTR PROCESSES

It is now well recognized (see, for instance, Pitman [13]) that, via Kingman's correspondence, there are bijective relations between the DP, the Blackwell–McQueen Pólya urn and the ESF. Here we point out a correspondence that is not so well known. Let $T_1, \ldots, T_n | F$ be independent and identically distributed survival times with unknown survival distribution $S(t) = 1 - F(t)$ and cumulative hazard $\Lambda(t) = \int_0^t F(\mathrm{d}s)/S(s-)$. If there are $K_n = k \leq n$ distinct ordered values $T_{(1)} > T_{(2)} > \cdots > T_{(k)}$, then, using the language of survival analysis, one can form death sets $D_j = \{i : T_i = T_{(j)}\}$ and risk sets $R_j = \{i : T_i \geq T_{(j)}\}$, with sizes $d_j = |D_j|$ and $r_j = |R_j|$. Hence, $r_j = r_{j-1} + d_j = \sum_{j \leq \ell \leq k} d_\ell$ with $r_k = n$ and $r_0 = 0$. It is evident that $\mathcal{D}_n = \{D_1, \ldots, D_k\}$ constitutes one of $k!$ orderings of a partition $\pi = (B_1, \ldots, B_k)$ of $[n]$. Let $\mathbf{d} = (d_1, \ldots, d_k)$, and consider the probabilities

$$p_{0,\theta}(\mathbf{d}) = \frac{\theta^k}{(\theta)^{\uparrow n}} \prod_{j=1}^{k} \frac{d_j!}{r_j} \tag{2.1}$$

and

$$\tilde{p}_{0,\theta}(\mathbf{d}) = \frac{n!}{\prod_{j=1}^{k} d_j!} p_{0,\theta}(\mathbf{d}).$$

These probabilities agree with variants of the age-ordered ESF, as derived in Donelley and Tavaré [4], that is, formula for $k \leq n$ distinct allelic types ordered by their ages. As the name suggests, there should be a connection between (2.1) and the DP, however, it is not an immediately obvious one. Formula (2.1) arises as a special case of a general formula in James [10] and Gnedin and Pitman [7]. Specifically, (2.1) is derived by choosing F to be a NTR process, as defined in Doksum [3], which involves a representation of the marginal distribution $M_n(\mathrm{d}\mathbf{t}) = \mathbb{E}[\prod_{1 \leq i \leq n} F(\mathrm{d}t_i)]$ in terms of a joint distribution of $(D_1, \ldots, D_k)$ and $(T_{(1)}, \ldots, T_{(k)})$. James [10] showed that if one integrates $M_n(\mathrm{d}\mathbf{t})$, with respect to $(T_{(1)}, \ldots, T_{(k)})$, then this yields generalized distributions on $\mathcal{D}_n$ where $p_{0,\theta}$ is a special case. With the exception of the DP, $M_n(\mathrm{d}\mathbf{t})$ does not produce independent and identically distributed unique values, when F is selected to be a NTR process.

For specifics, $\hat{S}(t) = \prod_{\{j : T_{(j)} \leq t\}} (1 - d_j/r_j)$ and $\hat{\Lambda}(t) = \sum_{\{j : T_{(j)} \leq t\}} d_j/r_j$ are the Kaplan–Meier and Nelson–Aalen estimators for S and Λ, respectively. Recall that a NTR survival process can always be represented as $S(t) = \prod_{\{j : \tau_j \leq t\}} (1 - \Delta_j)$, where (Δ_j, τ_j) are the points of a Poisson random measure N, with mean $\mathbb{E}[N(\mathrm{d}u, \mathrm{d}s)] = \rho(u|s)\Lambda_0(\mathrm{d}s)\,\mathrm{d}u$, on $([0,1], (0, \infty))$ where $\Lambda_0(\mathrm{d}s) = F_0(\mathrm{d}s)/S_0(s-)$ is a hazard rate. In particular, it follows that $F(\mathrm{d}s) = S(s-)\Lambda(\mathrm{d}s)$, where $\Lambda(t) = \sum_{\{j : \tau_j \leq t\}} \Delta_j$, is a random cumulative hazard corresponding to processes described in Hjort [9]. It is known that F has the DP law $\mathscr{P}_{0,\theta}(\mathrm{d}F | F_0)$ if $\rho(u|s) = \theta S_0(s-) u^{-1} (1-u)^{\theta S_0(s-)-1}$. The choice of F that produces (2.1) is $\rho(u) = \theta(\theta+1)(1-u)^{\theta-1}$, which is the term $(\theta + 1)$ multiplied by a Beta density function with parameter $(1, \theta)$. This can be seen as a special case of equation (42) in Gnedin and Pitman [7]. The prediction rule for this case can be deduced from James [10],

$$\mathbb{E}[F(\mathrm{d}t) | \mathbf{T}] = \sum_{\ell=1}^{k+1} q_\ell^* \tilde{F}_{\ell:n}(\mathrm{d}t) + \sum_{j=1}^{k} p_j^* \delta_{T_{(j)}}(\mathrm{d}t).$$

Here, setting $\Lambda_0(s) = s$, $\tilde{F}_{\ell:n}$ denotes a truncated exponential distribution with parameter $(\theta(\theta+1)/(\theta + r_{\ell-1})(\theta + r_{\ell-1} + 1))$ and support $T_{(\ell)} < t < T_{(\ell-1)}$. Furthermore, the transition probabilities for generating $\mathcal{D}_n$ according to $p_{0,\theta}$ in (2.1) are given by

$$p_j = \mathbb{E}[p_j^*|\mathcal{D}_n] = \frac{n}{(\theta+n)}\frac{r_j}{n}\frac{(d_j+1)}{r_j+1}\prod_{\ell=j+1}^{k}\frac{r_\ell}{r_\ell+1}$$

and

$$q_j = \mathbb{E}[q_j^*|\mathcal{D}_n] = \frac{\theta}{(\theta+n)}\frac{1}{r_{j-1}+1}\prod_{\ell=j}^{k}\frac{r_\ell}{r_\ell+1},$$

which can be used to generate a special case of a generalized ordered CRP. See Section 5.2.1 in James [10]. $F(t)$ is clearly not a DP. However, as in James [10], if one adds a spatial component (Δ_j, τ_j, x_j) with mean $\theta(\theta+1)(1-u)^{\theta-1}\Lambda_0(\mathrm{d}s)H(\mathrm{d}x)$, then this creates a spatial NTR process $F(\mathrm{d}t, \mathrm{d}x)$, where for $(t,x) \in \mathcal{R}_+ \times \mathcal{X}$, $P(\mathrm{d}x) = F(\infty, \mathrm{d}x)$ has the Dirchlet law $\mathscr{P}_{0,\theta}(\mathrm{d}P|H)$. Equation (42) in Gnedin and Pitman [7] yields the ordered partition distribution of the two-parameter family for the range $0 \le \alpha < 1$ and $\theta \ge 0$. Proposition 6.1 in James [10] used that result to show that the Pitman–Yor process with law $\mathscr{P}_{\alpha,\theta}(\mathrm{d}P|H)$ can be represented as a spatial NTR process $F(\infty, \mathrm{d}x)$. The analogous p_j and q_j can also be worked out explicitly. A remarkable aspect of this is that while Doksum's NTR specification of $F(t)$ does not contain the Pitman–Yor processes, there is a NTR process $F(t) = F(t, \mathcal{X})$ which produces a random partition of $[n]$, $(B_1, \ldots, B_k)$ whose law for $\theta \ge 0$ follows the two parameter (α, θ) CRP scheme.

3. POSTERIOR ESF AND SPECIES SAMPLING PROBLEMS

Species sampling problems refer to a broad class of statistical problems where samples are drawn from a population of individuals belonging to an (ideally) infinite number of species with unknown proportions. Given a sample of size n featuring $K_n = k$ species with frequency counts $(M_{1,n}, \ldots, M_{n,n}) = (m_{1,n}, \ldots, m_{n,n})$, interest lies in estimating global and local measures of species variety induced by considering an additional unobservable sample of size m: the former refer to the species variety of the whole additional sample, whereas the latter refer to the discovery probability at the $(n+m+1)$th step of the sampling process. These problems have originally appeared in ecology, and their importance has grown considerably in recent years, driven by challenging applications arising from bioinformatics, genetics, linguistics, networking and data confidentiality, design of experiments, machine learning, etc. The ESF, together with Pitman's two-parameter generalization, represents the cornerstone of the Bayesian nonparametric approach proposed by Lijoi et al. [11] for making inference measures of species variety. Indeed, it takes on the natural interpretation of a prior "sampling" model induced by assuming a DP prior on the unknown species composition of the population. Given that, the estimation of global and local measures of species variety relies on the study of distributional properties of a posterior ESF, namely, distributional properties of $(M_{1,n}, \ldots, M_{n+m,n+m})$ given $(M_{1,n}, \ldots, M_{n,n})$.

Under the assumption that P has the two-parameter Poisson–Dirchlet law (i.e., Pitman–Yor process) $\mathscr{P}_{\alpha,\theta}(\mathrm{d}P|H)$, the distribution of $M_{l,n+m}$ given $(M_{1,n}, \ldots, M_{n,n})$ arises by a direct application of Theorem 3 in Favaro et al. [5]. Along similar arguments, one may obtain the joint conditional distribution of $(M_{1,n+m}, \ldots, M_{n+m,n+m})$ given $(M_{1,n}, \ldots, M_{n,n})$. Of particular interest in Bayesian nonparametric inference for species sampling problems is the expectation and the large m asymptotic behavior of the distribution of $M_{l,n+m}$ given $(M_{1,n}, \ldots, M_{n,n})$. In particular, one has

$$\begin{aligned}
&\mathbb{E}[M_{l,n+m}|\mathbf{M}_n = (m_1, \ldots, m_n)] \\
&= \sum_{i=1}^{l}\binom{m}{l-i} m_i (i-\alpha)^{\uparrow(l-i)} \\
&\quad \cdot \frac{(\theta+n-i+\alpha)^{\uparrow(m-l+i)}}{(\theta+n)^{\uparrow m}} \\
&\quad + \binom{m}{l}(1-\alpha)^{\uparrow(l-1)}(\theta+k\alpha) \\
&\quad \cdot \frac{(\theta+n+\alpha)^{\uparrow(m-l)}}{(\theta+n)^{\uparrow m}}
\end{aligned} \tag{3.1}$$

with $k = \sum_{1\le i\le n} m_i$. Furthermore, let $B_{a,b}$ be a Beta random variable with parameter (a,b), and let S_q be a nonnegative random variable with density function proportional to $s^{q-1-1/\alpha} f_\alpha(s^{-1/\alpha})$, where f_α is the positive α-stable density function. As $m \to +\infty$,

$$\begin{aligned}
&\frac{M_{l,n+m}}{m^\alpha}\Big|(\mathbf{M}_n = (m_1, \ldots, m_n)) \\
&\quad \to \frac{\alpha(1-\alpha)^{\uparrow(l-1)}}{l!} B_{k+\theta/\alpha, n/\alpha-k} S_{(\theta+n)/\alpha}
\end{aligned} \tag{3.2}$$

almost surely, where $B_{k+\theta/\alpha,n/\alpha-k}$ and $S_{(\theta+n)/\alpha}$ are independent random variables. The limiting scale mixture $B_{k+\theta/\alpha,n/\alpha-k}S_{(\theta+n)/\alpha}$ provides the posterior counterpart of the well-known α-diversity of the two-parameter Ewens–Pitman sampling formula, which is recovered by setting $n = k = 0$. In particular, if $\alpha = 0$, then $M_{l,n+m} \to P_l$ weakly $m \to +\infty$, where P_l is a Poisson random variable with parameter θ/l.

Let $\mathcal{M}_{l,m} = \mathbb{E}[M_{l,n+m}|\mathbf{M}_n = (m_1, \ldots, m_n)]$ be the posterior expectation displayed in (3.1). Intuitively, $\mathcal{M}_{l,m}$ takes on the interpretation of the Bayesian nonparametric estimator, with respect to a squared loss function, of the number of species with frequency l in the enlarged sample of size $(n + m)$. Accordingly, (3.2) provides a tool for deriving corresponding large m asymptotic credible intervals. The estimator $\mathcal{M}_{l,m}$ is the prototypical example of a measure of global species variety, and other measures of global species variety may be introduced as suitable functions of it. For instance, for any integer $1 \le \tau \le n+m$, $\mathcal{M}_m(\tau) = \sum_{1\le l\le\tau} \mathcal{M}_{l,m}$ is the estimator of the so-called rare species variety, namely, the number of species with frequency less or the equal of a threshold τ. Then an estimator of the overall species variety is obtained by setting $\tau = n + m$. Besides these measures of global species variety, $\mathcal{M}_{l,m}$ also leads to measures of local species variety. Indeed, it is easy to show that $\mathcal{D}_{l,m} = (l-\alpha)\mathcal{M}_{l,m}/(\theta+n+m)$ is the estimator of the probability that the observation at the $(n + m + 1)$th draw coincides with a species with frequency l. In particular, $1 - \sum_{1\le l\le n+m} \mathcal{D}_{l,m}$ is the estimator of the probability of discovering a new species at the $(n+m+1)$th draw. For any $m \ge 1$, these estimators provide the natural Bayesian nonparametric counterparts of the celebrated Good–Toulmin estimator proposed by Good and Toulmin [8].

ACKNOWLEDGMENTS

Stefano Favaro is also affiliated with Collegio Carlo Alberto, Moncalieri, Italy and is supported in part by the European Research Council through StG N-BNP 306406. Lancelot F. James is supported in part by the Grant RGC-HKUST 601712 of the HKSAR.

REFERENCES

[1] ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

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

[3] DOKSUM, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2** 183–201. MR0373081

[4] DONNELLY, P. and TAVARÉ, S. (1986). The ages of alleles and a coalescent. *Adv. in Appl. Probab.* **18** 1–19. MR0827330

[5] FAVARO, S., LIJOI, A. and PRÜNSTER, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.* **23** 1721–1754. MR3114915

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

[7] GNEDIN, A. and PITMAN, J. (2005). Regenerative composition structures. *Ann. Probab.* **33** 445–479. MR2122798

[8] GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039

[9] HJORT, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18** 1259–1294. MR1062708

[10] JAMES, L. F. (2006). Poisson calculus for spatial neutral to the right processes. *Ann. Statist.* **34** 416–440. MR2275248

[11] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. MR2416792

[12] LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519

[13] PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory* (T. S. Ferguson, L. S. Shapley and J. B. MacQueen, eds.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward, CA. MR1481784

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