

On species sampling sequences induced by residual allocation models

Modelli di campionamento di specie indotta da assegnazione residua

Abel Rodríguez and Fernando A. Quintana

Abstract We discuss fully Bayesian inference in a class of species sampling models that are induced by residual allocation (sometimes called stick-breaking) priors on almost surely discrete random measures. This class provides a generalization of the well-known Ewens sampling formula that allows for additional flexibility while retaining computational tractability. In particular, the procedure is used to derive the exchangeable predictive probability functions associated with the generalized Dirichlet process of [3]. The procedure is illustrated with an application to genetics.

Key words: Exchangeable partition probability function; Generalized Dirichlet process; Probit-stick breaking process; Size-biased permutation.

1 Introduction

An exchangeable sequence of random variables $X_1, X_2, \dots$ defined on a probability space $(\mathcal{X}, \mathcal{B})$ follows a species sampling model (SSM) if its joint distribution can be characterized by a sequence of predictive rules where $X_1 \sim G_0$ and

$$p(X_{n+1} \in B | X_n, \dots, X_1) = \sum_{k=1}^{K^n} q_k^n(\mathbf{m}^n) \delta_{\tilde{X}_k}(B) + q_{K^n+1}^n(\mathbf{m}^n) G_0(B),$$

for some non-atomic measure G_0 on $(\mathcal{X}, \mathcal{B})$ and for all $B \in \mathcal{B}$. In the previous expression, δ_a denotes the degenerate probability measure placing probability 1 on a , $\mathbf{m}^n = (m_1^n, \dots, m_{K^n}^n)$ with $m_k^n = \sum_{j=1}^n I(X_j = \tilde{X}_k)$ being the number of values among

Abel Rodríguez

University of California, Santa Cruz, USA, e-mail: abel@ams.ucsc.edu

Fernando A. Quintana

Pontificia Universidad Católica de Chile, Santiago, Chile e-mail: quintana@mat.puc.cl

$X_1, \dots, X_n$ that are equal to $\tilde{X}_k$, $K^n = \max\{k : m_k^n > 0\}$ being the number of unique values among $X_1, \dots, X_i$, and q_k^n for $k \leq n$ and $n = 1, 2, \dots$ being a collection of functions of $\mathbf{m}^n$ (usually called the predictive probability functions, or PPFs) which for all n and $\mathbf{m}^n$ satisfy $q_k^n(\mathbf{m}^n) \geq 0$ and $\sum_{k=1}^{K^n+1} q_k^n(\mathbf{m}^n) = 1$.

A well-defined set of PPFs implies a symmetric joint probability distribution $p(m_1^n, \dots, m_{K^n}^n)$ for the number of species and the sample sizes associated with each one of them, which can be obtained through a recursion where $p(1) = 1$, $p(m_1^n, \dots, m_k^n + 1, \dots, m_{K^n}^n) = q_k^n(m_1^n, \dots, m_{K^n}^n) p(m_1^n, \dots, m_k^n, \dots, m_{K^n}^n)$ for $k \leq K^n$, and $p(m_1^n, \dots, m_{K^n}^n, 1) = q_{K^n+1}^n(m_1^n, \dots, m_{K^n}^n) p(m_1^n, \dots, m_{K^n}^n)$ [10, 5]. The function p is often called the exchangeable partition probability function (EPPF).

Constructing EPPFs is a difficult task, as ensuring that $X_1, X_2, \dots$ is an exchangeable sequence implies quite strict conditions on the PPFs [11, 5]. As a consequence, the number of species sampling models in the literature is rather small, with the most popular ones being those associated with the Dirichlet process (DP) [2, 1], the two-parameter Poisson-Dirichlet process (PDP) [10], and the normalized inverse Gaussian measures (NIGM) [6]. This list suggests that there is a close link between the class of nonparametric priors on almost-surely discrete distributions and the class of SSMs. Indeed, a well known result due to [11] establishes that the De Finetti measure of any SSM can be written as $G(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{X_k^*}(\cdot) + RG_0$ for some sequence of positive random variables $\omega_1, \omega_2, \dots$ and R such that $1 - R = \sum_{k=1}^{\infty} \omega_k \leq 1$ almost surely, $X_1^*, X_2^*, \dots$ is a random sample from a non-atomic G_0 , and the sequences $\omega_1, \omega_2, \dots$ and $X_1^*, X_2^*, \dots$ are independent.

2 PPFs and EPPFs from residual allocation models

The random probability measure G is said to follow an independent residual allocation prior if it can be represented as

$$G(\cdot) = \sum_{k=1}^N \omega_k \delta_{X_k^*}, \quad (1)$$

where $N \in \mathbb{N}$, $X_1^*, X_2^*, \dots$ is a sequence of independent and identically distributed realizations from some distribution G_0 and $\omega_k = z_k \prod_{\ell < k} \{1 - z_\ell\}$ where $z_k \sim H_k^\theta$ independently for all $k = 1, 2, \dots$ (with the convention $z_N = 1$ if $N < \infty$), and H_k^θ is a probability distribution on $[0, 1]$ indexed by θ . Examples include the Dirichlet process (where $N = \infty$, $\theta = b$, and $z_k \sim \text{Beta}(1, b)$ for some $b > 0$) and the Poisson-Dirichlet process (for which $\theta = (a, b)$, $z_k \sim \text{Beta}(1 - a, b + ka)$ and either $N = \infty$, $0 \leq a < 1$ and $b > -a$ or $N < \infty$, $a < 0$ and $b = -aN$).

[4] and [11] show that the EPPF associated with (1) is given by

$$p(m_1^n, \dots, m_{K^n}^n \mid \theta) = \sum_{(j_1, \dots, j_{K^n}) \in \mathcal{J}_{K^n}} \mathbb{E}_\theta \left\{ \prod_{k=1}^{K^n} \omega_{j_k}^{m_k^n} \right\}, \quad (2)$$

where $\mathcal{J}$ represents the set of all possible sequences of distinct (not necessarily consecutive) positive integers of length K^n .

The residual allocation construction provides a very general mechanism to define priors on random measures and, implicitly, on the associated SSMs. In the sequel, we will focus on the special case where all the stick-breaking ratios z_k are not only independent but also identically distributed, i.e., $z_k \sim H$ for all $k = 1, 2, \dots$

Lemma 1. *The EPPF induced by the stick breaking prior in (1) is given by*

$$p(m_1^n, \dots, m_{K^n}^n \mid \theta) = \sum_{\sigma \in \mathcal{P}_{K^n}} \left\{ \prod_{k=1}^{K^n} \frac{\gamma_\theta(m_{\sigma_k}^n, \sum_{j=k+1}^{K^n} m_{\sigma_j}^n)}{1 - \gamma_\theta(0, \sum_{j=k}^{K^n} m_{\sigma_j}^n)} \right\}, \quad (3)$$

where $\mathcal{P}_{K^n}$ denotes the set of all permutations of $\{1, \dots, K^n\}$, $\sigma = (\sigma_1, \dots, \sigma_{K^n})$ and $\gamma_\theta(x, y) = \mathbb{E}_\theta \{z^x(1-z)^y\}$ with $z \sim H$. $\diamond$

Note that the function $g_\theta(m_{\sigma_1}^n, \dots, m_{\sigma_{K^n}}^n) = \prod_{k=1}^{K^n} \frac{\gamma_\theta(m_{\sigma_k}^n, \sum_{j=k+1}^{K^n} m_{\sigma_j}^n)}{1 - \gamma_\theta(0, \sum_{j=k}^{K^n} m_{\sigma_j}^n)}$ satisfies $g_\theta(1) = 1$, but is in general not symmetric on its arguments and depends on the permutation σ , which encodes the order of the indices that identify the components that generate the unique labels associated with the observed species. Indeed, in the future it will be convenient to interpret $g_\theta(m_{\sigma_1}^n, \dots, m_{\sigma_{K^n}}^n)$ as describing the joint probability distribution of the exchangeable vector $(m_1^n, \dots, m_{K^n}^n)$ and the permutation σ , $p(m_1^n, \dots, m_{K^n}^n, \sigma \mid \theta)$.

3 A new species sampling models

[3] defined a new class of residual allocation priors by letting $N = \infty$ and $u_k \sim \text{Beta}(a, b)$. Since $a = 1$ implies that G follows a Dirichlet process, he called this construction a generalized Dirichlet processes (GDP). The EPPF associated with the GDP is

$$p(m_1^n, \dots, m_{K^n}^n \mid a, b) = \left\{ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \right\}^{K^n} \frac{\Gamma(b)}{\Gamma(b+n)} \sum_{\sigma \in \mathcal{P}_{K^n}} \prod_{k=1}^{K^n} \frac{\Gamma(a+m_k^n)}{\left\{ \frac{\Gamma(a+b+\sum_{j=k}^{K^n} m_{\sigma_j}^n)}{\Gamma(b+\sum_{j=k}^{K^n} m_{\sigma_j}^n)} - \frac{\Gamma(a+b)}{\Gamma(b)} \right\}}.$$

The GDP is particularly appealing for creating species sampling models because it allows for a multitude of asymptotic behaviors for the expected number of species.

Lemma 2. *The expected number of clusters $e_{a,b}(K^n)$ for the species sampling model induced by a generalized Dirichlet process where $z_k \sim \text{Beta}(a, b)$ is given by*

$$e_{a,b}(K^n) = \sum_{i=1}^n \frac{ia\Gamma(a+b)\Gamma(b+i-1)}{\Gamma(b)\Gamma(a+b+i) - \Gamma(a+b)\Gamma(b+i)}. \quad (4)$$

Note that for $a = 1$, equation (4) simplifies to $e_{a,b}(K^n) = \sum_{i=1}^n \frac{b}{b+i-1} \sim o(\log n)$, a well known result for the Dirichlet process. For more general values of a , Stirling's approximation can be used to show that

$$\frac{na\Gamma(a+b)\Gamma(b+n-1)}{\Gamma(b)\Gamma(a+b+n) - \Gamma(a+b)\Gamma(b+n)} \approx C(a,b)n^{-a}.$$

where $C(a,b) = \{a\Gamma(a+b)/\Gamma(b)\} \exp\{-2(a+1)\}$. Hence, for $a \leq 1$ the expected number of distinct species will grow slowly but without bound as n increases, with $e_{a,b}(K^n) \sim o(n^{1-a})$. This behavior is similar to the one associated with the Poisson Dirichlet process [12]. However, for $a > 1$ the expected number of species instead converges to a finite constant.

4 Bayesian inference for species sampling models

Given a prior $p(\theta)$, the posterior distribution of the parameters of the SSM is trivially given by

$$p(\theta \mid m_1^n, \dots, m_{K^n}^n) \propto p_\theta(m_1^n, \dots, m_{K^n}^n)p(\theta),$$

where $p_\theta(m_1^n, \dots, m_{K^n}^n)$ was given in (3). Since θ is typically a low dimensional vector, exploring this posterior distribution using a Markov chain Monte Carlo algorithm should be in principle straightforward. However, for realistic applications, evaluating the sum over all possible permutations of the set $\{1, 2, \dots, K^n\}$ is infeasible.

We can get around this issue by reinterpreting $g_\theta(m_{\sigma_1}^n, \dots, m_{\sigma_{K^n}}^n)$ as describing a joint distribution over both the permutation σ and the vector of observations $m_1^n, \dots, m_{K^n}^n$, and treating σ as a latent variable to be imputed as part of an augmented sampler. The joint posterior distribution of this augmented model can be written as

$$p(\theta, \sigma \mid m_1^n, \dots, m_{K^n}^n) \propto g_\theta(m_{\sigma_1}^n, \dots, m_{\sigma_{K^n}}^n)p(\theta),$$

which can be explored by alternatively sampling from the full conditional associated with the parameters θ that control the prior size of the weights, and the full conditional distribution of the permutation σ controlling the assignment of species to atoms. For sampling θ we favor a Metropolis-Hasting step with a multivariate Gaussian random walk proposal on a suitable transformation of θ . For sampling σ we use a combination of symmetric proposals that update blocks of elements of the permutation. Furthermore, this MCMC algorithm can be easily extended to estimate the predictive distribution of $\tilde{K}^{\tilde{n}}$, the number of new species to be observed if

the current sample size is extended from n to $n + \tilde{n}$ using “order-dependent” EPPFs derived from g_θ .

5 Illustration

We consider a real dataset discussed in [9], [8], and [7]. The data consists of $n = 2574$ randomly selected expressed sequence tags taken from a large cDNA library made from the 0 mm to 3 mm buds of tomato flowers. The number distinct tags observed in this sample is $K = 1814$; Table 1 shows the number of times each species size was observed in the sample (e.g., 1423 tags were observed only once).

Table 1 Data on the number of expressed sequence tags from a cDNA library made from the 0 mm to 3 mm buds of tomato flowers.

Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	23	27
Frequency	1423	253	71	33	11	6	2	3	1	2	2	1	1	1	2	1	1

We compare the species sampling model associated with the GDP against a the species sample models induced by the DP and the PDP. For this illustration, the algorithm described in Section 4 was used to generate of 20000 samples from the posterior distribution, obtained after a burn-in of 100000 iterations and thinning every 80 observations. The parameters a and b of the GDP are given exponential prior distributions with means 0.1 and 1000 respectively; the corresponding posterior means are $e(a | \mathbf{m}) = 0.053$ and $e(b | \mathbf{m}) = 1345$, while the respective 95% symmetric credible intervals are (0.001, 0.186) and (1172, 1584). The equivalent sample sizes for these parameters are 3385 for a and 5431 for b . In the case of the PDP, the parameters are given a uniform distribution on the $(0, 1)$ interval and an exponential prior with mean 800; the corresponding posterior means are 0.61 and 735.9, with 95% credible intervals of (0.53, 0.67) and (515.6, 1017.4). Finally, for the DP an exponential prior with mean 1000 is employed for the concentration parameters, which leads to a posterior mean of 2724.9 and 95% posterior credible interval of (2491.3, 2972.1).

Figure 1 presents the distribution of the number of unique species in a new sample of size 500 taken from this population under these three models as well as the posterior expected number of species for a new sample of size $\tilde{n}$. Not surprisingly given the large number of singleton species, the PDP favors a larger number of new species than the DP. On the other hand, the GDP seems to offer a compromise between both predictions, with a forecast of the number of species that seems to be in between those generated by the DP and the PDP.

References

1. Blackwell, D., MacQueen, J.B. Ferguson distribution via Pólya urn schemes. *Ann. Stat.* **1**, 353–355 (1973).
2. Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230 (1973).
3. Hjort, N. Bayesian analysis for a generalized Dirichlet process prior. Technical Report. University of Oslo (2000).
4. Kingman, J.F. The representation of partition structures. *J. Lond. Math. Soc.* **18**, 374–380 (1978).
5. Lee, J., Quintana, F.A., Müller, P., Trippa, L. Defining predictive probability functions for species sampling models. *Stat. Sc.* **28**, 209–222 (2013).
6. Lijoi, A., Mena, R.H., Prünster, I. Bayesian nonparametric analysis for a generalized Dirichlet process prior. *Stat. Infer. Stoch. Process* **8**, 283–309 (2005).
7. Lijoi, A., Mena, R.H., Prünster, I. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786 (2007).
8. Mao, C.X. Prediction of the conditional probability of discovering a new class. *J. Am. Stat. Assoc.* **99**, 1108–1118 (2004).
9. Mao, C.X., Lindsay, B.G. A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669–682 (2002).
10. Pitman, J. Exchangeable and partially exchangeable random partitions. *Probab. Theor. Relat. Field.* **102**, 145–158 (1995).
11. Pitman, J. Some developments of the Blackwell-MacQueen urn scheme, in: Ferguson, T.S., Shapeley, L.S., MacQueen, J.B. (Eds.), *Statistics, Probability and Game Theory*. Papers in Honor of David Blackwell. Hayward, CA:IMS, pp. 245–268 (1996).
12. Pitman, J. Combinatorial stochastic processes. Volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lectures from the 32nd Summer School on Probability Theory (2002).

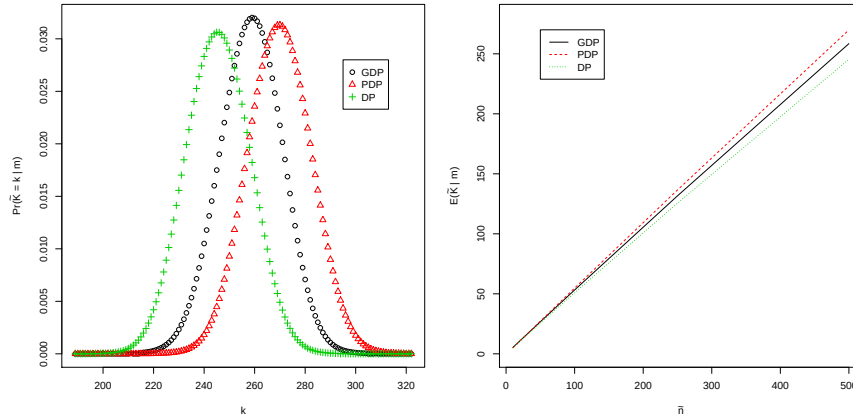


Fig. 1 Prediction for the number of new species for the expressed sequence tag data of [9]. The left panel shows the posterior distribution of the number of new species in a new sample of 500 individuals under each of three competing models: the generalized Dirichlet process (GDP), the Poisson-Dirichlet process (PDP) and the Dirichlet process (DP). The right panel shows the posterior expected number of species for a new sample of size $\bar{n}$ under each of the three competing models.