

Bayesian nonparametric analysis for a species sampling model with finitely many types

Annalisa Cerquetti

Abstract We derive a Bayesian nonparametric analysis for a species sampling model of Gibbs form with *finite* but *random* number of types, recently introduced in the literature. Our results rely on and complement existing analysis for models with infinite number of types. We adopt the classical notation of the theory of exchangeable partitions and the approach based on our groups sequential construction of Gibbs partitions. We obtain posterior predictive analysis for both the number of new species and the number of new observations belonging to new species.

Key words: Bayesian nonparametrics, Conditional Gibbs structures, Exchangeable partitions, Lah numbers, Species sampling models

1 Introduction

In the *species sampling problem* (Fisher, 1943; Engen, 1978; Pitman, 1996) a random sample is drawn from an hypothetically infinite population of individuals to make inference on the unknown total number of different species. In Lijoi et al. (2007, 2008) a Bayesian nonparametric approach to this problem has been recently devised under the assumption that the vector of multiplicities of different species observed is a random sample from an *exchangeable Gibbs* partition of type $\alpha \in [0, 1)$, and a posterior predictive inference has been derived on species richness for an additional sample. Actually exchangeable Gibbs partitions are defined for values of type parameter α in the larger subset $[-\infty, 1)$ (Gnedin and Pitman, 2006). These are models for random partitions of the positive integers which extend the Ewens-Pitman two-parameter family (Pitman, 1995) and are characterized by a probability

Annalisa Cerquetti

Dipartimento di Studi Geoeconomici, Linguistici, Statistici e Storici per l'Analisi Regionale, SAPIENZA University of Rome, Via del Castro Laurenziano 9, 00161 Rome, Italy e-mail: annalisa.cerquetti@uniroma1.it

function (EPPF) with the following structure

$$p(n_1, \dots, n_k) = V_{n,k} \prod_{j=1}^k (1 - \alpha)_{n_j-1}, \quad (1)$$

for $(x)_s = x(x+1) \cdots (x+s-1)$ rising factorials, and with weights $(V_{n,k})$ being solution to the *backward* recursion $V_{n,k} = (n - k\alpha)V_{n+1,k} + V_{n+1,k+1}$, for $V_{1,1} = 1$. In their paper Gnedin and Pitman even provide a constructive result for this class (cfr. Th. 12) showing that each element arises as a unique probability mixture of extreme partitions which are

- a) $PD(\alpha, \xi|\alpha|)$ partitions with $\xi = 1, \dots, \infty$ for $\alpha \in [-\infty, 0)$,
- b) $PD(0, \theta)$ partitions with $\theta \in [0, \infty)$ for $\alpha = 0$,
- c) $PK(\rho_\alpha|t)$ partitions with $t \in [0, \infty)$ for $\alpha \in (0, 1)$.

Here $PD(\cdot, \cdot)$ stands for the two-parameter Poisson-Dirichlet distribution (Pitman and Yor, 1997) and $PK(\rho_\alpha|t)$ for the conditional Poisson-Kingman distribution derived from the stable subordinator (cfr. Pitman, 2003). Lijoi et al. (2008) establish general distributional results and properties for EPPFs belonging to subclass c) to make conditional predictions according to a Bayesian nonparametric procedure. Focusing on class c) their analysis relies on the hypothesis that the total unknown number of different species is so large to be assumed infinite, an assumption that may be unrealistic in concrete applications.

Here we focus on a class of random partitions with *finite* but *random* number of different species, recently introduced in Gnedin (2010, Sect. 6), which belongs to subclass a) and contribute, deriving posterior predictive results in view of possible future Bayesian implementations.

First recall that for $\alpha < 0$, $\theta = |\alpha|\xi$ and $\xi = 1, 2, 3, \dots$, a $PD(\alpha, \xi|\alpha|)$ model has EPPF (see e.g. Pitman, 1996, 2006) of the form

$$p_{\xi, \alpha}(n_1, \dots, n_k) = \frac{(\xi|\alpha| - |\alpha|)_{k-1 \downarrow |\alpha|}}{(\xi|\alpha| + 1)_{n-1}} \prod_{j=1}^k (1 + |\alpha|)_{n_j-1}, \quad (2)$$

which arises by the following sequential procedure. Given the partition of $[n] = \{1, \dots, n\}$ in $K_n = k$ blocks with occupancy counts $\mathbf{n} = (n_1, \dots, n_k)$, the partition of $[n+1]$ is obtained through the following one-step prediction rules

$$p_j(\mathbf{n}) = \frac{n_j + |\alpha|}{n + |\alpha|\xi} \quad \text{and} \quad p_0(\mathbf{n}) = \frac{|\alpha|(\xi - k)}{\xi|\alpha| + n},$$

where $p_j(\mathbf{n})$ for $j = 1, \dots, k$ stands for the probability of randomly placing $n+1$ in an *old* block j , while $p_0(\mathbf{n})$ stands for the probability of $n+1$ to form a new block $k+1$. For $\alpha = -1$ the EPPF in (2) reduces to

$$p_{\xi, -1}(n_1, \dots, n_k) = \frac{(\xi - 1)_{k-1 \uparrow -1}}{(\xi + 1)_{n-1}} \prod_{j=1}^k n_j,$$

with one-step prediction rules

$$p_j(\mathbf{n}) = \frac{n_j + 1}{n + \xi} \quad \text{and} \quad p_0(\mathbf{n}) = \frac{(\xi - k)}{n + \xi}.$$

The corresponding limit frequencies $(\tilde{P}_{\xi,1}, \dots, \tilde{P}_{\xi,\xi})$ are well-known to have *stick-breaking* representation $\tilde{P}_{\xi,j} = W_j \prod_{i=1}^{j-1} (1 - W_i)$ for independent $W_i \sim \text{Beta}(2, \xi - i)$, $i = 1, \dots, \xi$ and $\text{Beta}(2, 0)$ a Dirac mass at 1.

As from Gneden and Pitman representation result, $PD(\alpha, \xi | \alpha)$ models are extreme points of a convex set of Gibbs partitions of type $\alpha < 0$, whose elements are in one to one correspondence with a set of mixing distributions over the set of the positive integers. Gneden (2010, Sect. 6) studies the particular model arising by mixing a $PD(\alpha, |\alpha| \xi)$ model for $\alpha = -1$ over ξ , with density

$$\mathbb{P}(K = \xi) = \frac{\gamma(1 - \gamma)_{\xi-1}}{\xi!},$$

for $\xi = 1, 2, \dots$ and $\gamma \in (0, 1)$, and shows it has Gibbs weights

$$V_{n,k} = \frac{(k-1)! (1 - \gamma)_{k-1} (\gamma)_{n-k}}{(n-1)! (1 + \gamma)_{n-1}} = \frac{(k-1)! \gamma(1 - \gamma)_{k-1}}{(n-1)! (\gamma + n - k)_k}. \quad (3)$$

The corresponding EPPF is given by

$$p_{\gamma,-1}(n_1, \dots, n_k) = \frac{(k-1)! (1 - \gamma)_{k-1} (\gamma)_{n-k}}{(n-1)! (1 + \gamma)_{n-1}} \prod_{j=1}^k n_j! \quad (4)$$

and is obtained by sequential construction with one-step prediction rules

$$p_j(\mathbf{n}) = \frac{(n - k + \gamma)(n_j + 1)}{n(n + \gamma)} \quad \text{for } j = 1, \dots, k \quad \text{and} \quad p_0(\mathbf{n}) = \frac{k(k - \gamma)}{n(n + \gamma)}.$$

2 Bayesian nonparametric analysis for Gneden's species sampling model with finitely many types

To realize a Bayesian posterior predictive analysis of Gneden's model we first rephrase the one-step prediction rules as m -steps prediction rules as in Cerquetti (2009, Sect. 2.1), which here we formulate as a group sequential random allocation of balls labelled $1, 2, \dots$ in a series of boxes. First notice that for $k^* \geq 1$ and $m \geq 1$, from (3) we obtain

$$V_{n+m,k+k^*} = \frac{(k+k^*-1)! (1 - \gamma)_{k+k^*-1} (\gamma)_{n+m-k+k^*}}{(n+m-1)! (1 + \gamma)_{n+m-1}},$$

then, by basic properties of rising factorials, specialization of the general formulas for Gibbs partitions easily follow. Start with box $B_{1,1}$ with a single ball. Given the placement of the first group of n balls in a $(n_1, \dots, n_k)$ configuration in k boxes, the new group of m balls labelled $\{n+1, \dots, n+m\}$ is:

a) allocated in the *old* k boxes in configuration $(m_1, \dots, m_k)$, for $m_j \geq 0$, $\sum_{j=1}^k m_j = m$, with probability

$$p_{\mathbf{m}}(\mathbf{n}) = \frac{1}{(n)_m} \frac{(\gamma+n-k)_m}{(\gamma+n)_m} \prod_{j=1}^k (n_j+1)_{m_j}; \quad (5)$$

b) allocated in k^* new boxes in configuration $(s_1, \dots, s_{k^*})$, for $\sum_{j=1}^{k^*} s_j = m$, $1 \leq k^* \leq m$, $s_j \geq 1$, with probability

$$p_{\mathbf{s}}(\mathbf{n}) = \frac{(k)_{k^*}}{(n)_m} \frac{(k-\gamma)_{k^*} (\gamma+n-k)_{m-k^*}}{(\gamma+n)_m} \prod_{j=1}^{k^*} s_j!; \quad (6)$$

c) $s < m$ balls are allocated at k^* new boxes in configuration $(s_1, \dots, s_{k^*})$ and the remaining $m-s$ balls in the *old* boxes in configuration $(m_1, \dots, m_k)$ for $\sum_{j=1}^m m_j = m-s$, $1 \leq s \leq m$, $\sum_{j=1}^{k^*} s_j = s$, $m_j \geq 0$, $s_j \geq 1$ with probability

$$p_{\mathbf{s}, \mathbf{m}}(\mathbf{n}) = \frac{(k)_{k^*}}{(n)_m} \frac{(k-\gamma)_{k^*} (\gamma+n-k)_{m-k^*}}{(\gamma+n)_m} \prod_{j=1}^k (n_j+1)_{m_j} \prod_{j=1}^{k^*} s_j!. \quad (7)$$

Now, by exploiting the definition of central and non-central *Lah numbers* as particular case of central and non-central *generalized Stirling numbers* of the first kind, these m -steps prediction rules allow to readily obtain Bayesian posterior predictive distributional results for an additional m -sample specializing results in Lijoi et al. (2008) by means of expressions derived in Cerquetti (2009). It is worth to notice that the results in Lijoi et al (2008) are expressed in terms of *generalized factorial coefficients* (see Charalambides, 2005). For the relation with generalized Stirling numbers see Cerquetti (2009, Sect. 1).

First notice that generalized Stirling numbers of the first kind $S_{n,k}^{-1,-\alpha}$ for $\alpha = -1$ admit an explicit expression known as *Lah numbers*

$$S_{n,k}^{-1,1} = \binom{n-1}{k-1} \frac{n!}{k!},$$

which are connection coefficients defined by

$$(x)_n = \sum_{k=0}^n \frac{n!}{k!} \binom{n-1}{n-k} (x)_{k\uparrow-1}.$$

Hence, by an application of (10) in Gnedin and Pitman (2006), (see Gnedin, 2010, eq. (4)) the distribution of the number K_n of occupied boxes for Gnedin's model is readily calculated as

$$\mathbb{P}(K_n = k) = V_{n,k} S_{n,k}^{-1,1} = \binom{n}{k} \frac{(1-\gamma)_{k-1} (\gamma)_{n-k}}{(1+\gamma)_{n-1}}.$$

Then it is enough to recall the definition of *non-central* Lah numbers with parameter of non-centrality r (see e.g. Charalambides, 2005) which correspond to non central generalized Stirling numbers of the first kind for $\alpha = -1$

$$S_{n,k}^{-1,1,r} = \frac{n!}{k!} \binom{n-r-1}{n-k},$$

to derive a Bayesian nonparametric analysis of the Gibbs class (4). Notice that for the sake of generality here we maintain the treatment in terms of random allocation of balls in boxes. The obvious traslation in terms of random partition of individuals among species follows easily.

2.1 Posterior predictive analysis for an additional m -sample

Now going back to the m -steps sequential construction, notice that in a Bayesian perspective it has a natural interpretation as follows. Given that the first n -observations ($X_1 = x_1, \dots, X_n = x_n$) from model (4) have produced an allocation in k different blocks with multiplicities $(n_1, \dots, n_k)$, we are interested in doing inference with respect to the placement in the already observed different k blocks (species) and/or in an unknown number K_m of new blocks (species) for an additional sample of $m \geq 1$ observations, say $X_{n+1}, \dots, X_{n+m}$. Here is what we are able to obtain.

Given the sufficiency of the number K_n of boxes induced by the basic sample, the joint distribution of the number of balls allocated in new k^* boxes in a specific configuration $(s_1, \dots, s_{k^*})$ given $(n_1, \dots, n_k)$ is obtained marginalizing (7) with respect to $(m_1, \dots, m_k)$ and by an application of the multinomial theorem for rising factorials as in Lijoi et al. (2008) (cfr. eq. (17) in Cerquetti, 2009), hence

$$\mathbb{P}(s_1, \dots, s_{k^*} | K_n = k) = \frac{(k)_{k^*}}{(\gamma+n)_m} \frac{(k-\gamma)_{k^*} (\gamma+n-k)_{m-k^*}}{(n)_m} \binom{m}{s} (n+k)_{m-s} \prod_{j=1}^{k^*} s_j!. \quad (8)$$

The joint distribution of the number of *new* boxes K_m and of the total number of balls falling in new boxes S_m , given K_n , (cfr. eq. (18) in Cerquetti, 2009) is given by

$$\begin{aligned} \mathbb{P}(K_m = k^*, S = s | K_n = k) &= \\ &= \frac{(k)_{k^*}}{(n)_m} \frac{(k-\gamma)_{k^*} (\gamma+n-k)_{m-k^*}}{(\gamma+n)_m} \binom{m}{s} (n+k)_{m-s} \binom{s}{k^*} \frac{(s-1)!}{(k^*-1)!} \end{aligned} \quad (9)$$

and arises by (8) summing over the space of all partitions of s elements in k^* blocks and exploiting the definition of Lah numbers.

Marginalizing (9) with respect to K_m , a probability distribution for the total number S_m of balls in new boxes is obtained in terms of Lah numbers according to eq. (19) in Cerquetti (2009) and eq. (11) in Lijoi et al. (2008)

$$\begin{aligned} \mathbb{P}(S_m = s | K_n = k) &= \\ &= \binom{m}{s} \frac{(n+k)_{m-s\uparrow}}{(n)_m(\gamma+n)_m} \sum_{k^*=0}^s \binom{s}{k^*} \frac{(s-1)!}{(k^*-1)!} (k)_{k^*} (k-\gamma)_{k^*} (\gamma+n-k)_{m-k^*}, \end{aligned}$$

and by an application of Proposition 2. in Lijoi et al. (2008) the corresponding expected value is given by

$$\mathbb{E}(S_m | K_n = k) = m \frac{V_{n+1,k+1}}{V_{n,k}} = m \frac{k(k-\gamma)}{n(n+\gamma)}.$$

The probability distribution of the number K_m of new blocks induced by the additional sample given the basic sample follows marginalizing (9) with respect to S_m and exploiting the definition of non-central Lah numbers with parameter of non centrality $r = -(n+k)$. An application of eq. (4) in Lijoi et al. (2007) yields

$$\mathbb{P}(K_m = k^* | K_n = k) = \frac{(k-\gamma)_{k^*} (\gamma+n-k)_{m-k^*}}{(n+\gamma)_m} \binom{m}{k^*} \frac{(n+k+k^*)_{m-k^*} (k)_{k^*}}{(n)_m}, \quad (10)$$

with the corresponding expected value providing the *Bayes estimator* for K_m under quadratic loss function,

$$\mathbb{E}(K_m | K_n = k) = \frac{(k)_{n+m}}{(n+\gamma)_m} \sum_{k^*=0}^m \binom{m}{k^*} \frac{k^*}{(k+k^*)_n} \frac{(k-\gamma)_{k^*} (\gamma+n-k)_{m-k^*}}{(n)_m}.$$

Notice that as $m \rightarrow \infty$, (10) agrees with the posterior result for the total number of boxes obtained in Gnedin (2010). In fact, by standard asymptotic $\Gamma(n+a)/\Gamma(n+b) \sim n^{a-b}$ and recalling the definition of rising and falling factorials in terms of Gamma function $(a)_{b\uparrow} = \Gamma(a+b)/\Gamma(a)$ and $(a)_{b\downarrow} = \Gamma(a+1)/\Gamma(a-b+1)$, equation (10) reduces to

$$\mathbb{P}(K = k^* | K_n = k) = \frac{(n-1)!}{(k-1)!} (\gamma+n-k)_k \frac{(k-\gamma)_{k^*} \Gamma(k+k^*)}{\Gamma(k^*+1) \Gamma(n+k+k^*)}. \quad (11)$$

The posterior distribution obtained in Gnedin (2010, cfr. eq. 10), in terms of $\varkappa = k+k^*$, is given by

$$\mathbb{P}(\varkappa = \varkappa | K_n = k) = \frac{(n-1)!}{(k-1)!(\varkappa+n-1)!} \prod_{i=1}^{k-1} (\varkappa-i) \prod_{j=1}^k (\gamma+n-j) \prod_{l=k}^{\varkappa-1} (l-\gamma)$$

for $1 \leq k \leq n$, $\varkappa \geq k$ and may be re-written as

$$= \frac{(n-1)!}{(k-1)!(\varkappa+n-1)!} (\varkappa-1)_{k-1\downarrow} (k-\gamma)_{\varkappa-k} (\gamma+n-k)_k,$$

hence, by substitution $\varkappa = k + k^*$, the result easily follows. The corresponding expected value which may play the role of the *asymptotic Bayesian estimator* for the numbers of block, results

$$\mathbb{E}(K|K_n = k) = \frac{(n-1)!}{(k-1)!} (n+\gamma-1)_{k\downarrow} \sum_{k^*=0}^{\infty} \frac{1}{(k^*-1)!} \frac{(k-\gamma)_{k^*}}{(k+k^*)_n}$$

which expressed in terms of $\varkappa = k + k^*$ yields

$$\mathbb{E}(\Xi|K_n = k) = \frac{(n-1)!}{(k-1)!} \frac{(n+\gamma-1)_{k\downarrow}}{(k-\gamma-1)!} \sum_{\varkappa=k}^{\infty} \frac{(\varkappa-\gamma-1)!}{(\varkappa-k-1)!} \frac{1}{(\varkappa)_n}.$$

Going back to the analysis for finite m , the distribution of the number S_m of balls in the new m -sample which belong to new boxes, given the number of boxes in the basic sample K_n and the number of new boxes K_m , follows by an application of Eq. (12) in Lijoi et al. (2008)

$$\mathbb{P}(S_m = s | K_m = k^*, K_n = k) = \binom{s-1}{k^*-1} \binom{n+k+m-s-1}{m-s} / \binom{n+k+m-1}{m-k^*}.$$

By Proposition 4 in Lijoi et al. (2008) (cfr. also Corollary 10, in Cerquetti, 2008) the probability that the m new balls don't belong to a subset of $(k-r)$ old boxes arises from (7) by summing over the ways to choose s balls from the m of the new group, by summing over the ways to partition s balls in a subset of k^* boxes, and over the ways to allocate $m-s$ balls in at most r old boxes and is equal to

$$\sum_{k^*=1}^m \frac{(k)_k^*}{(n)_m} \frac{(k-\gamma)_k^* (\gamma+n-k)_{m-k^*}}{(\gamma+n)_m} \binom{m}{k^*} \frac{1}{(r + \sum_j n_j + m)_{k^*-m}}.$$

As for the *conditional Gibbs structure* characterizing Gneden's model, (cfr. Prop. 3. in Lijoi et al. 2008), it may be obtained by the operation of *deletion of the first k classes* (Pitman, 2003, see also Cerquetti 2008, Prop. 12) and results

$$p(s_1, \dots, s_{k^*} | \mathbf{m}, \mathbf{n}) = \frac{\Gamma(k+k^*)\Gamma(k+k^*-\gamma)\Gamma(\gamma+n+m-k-k^*)}{\sum_{k^*=0}^m \binom{s}{k^*} \Gamma(k^*)_k \Gamma(s) \Gamma(k+k^*-\gamma) \Gamma(\gamma+n+m-k-k^*)} \prod_{j=1}^{k^*} s_j!.$$

Finally a Bayesian nonparametric estimator for the probability of ball $n+m+1$ th to fall in a new box given $K_n = k$, which in species sampling corresponds to the probability to discover a new species without observing the intermediate sample of dimension m , (known as the *discovery probability*) is readily derived by an application of Eq. (6) in Lijoi et al. (2007) as

$$\hat{D}_m^{n,k} = \sum_{k^*=0}^m \frac{(k)_{k^*+1}}{(n)_{m+1}} \frac{(k-\gamma)_{k^*+1}(\gamma+n-k)_{m-k^*}}{(\gamma+n)_{m+1}} \binom{m}{k^*} \frac{(m+n+k-1)!}{(n+k+k^*-1)!}.$$

3 Final remarks

We have presented an extension of the Bayesian nonparametric analysis of exchangeable Gibbs partitions to a new species sampling model characterized by a finite but random number of different species. Our results allow to fully implement a posterior predictive analysis when assuming a large (hypothetically infinite) number of species is not realistic for some specific observational settings.

References

1. Cerquetti, A.: Generalized Chinese restaurant construction of Gibbs partitions and related results. Available via arXiv <http://arxiv.org/abs/0805.3853v1> (2008)
2. Cerquetti, A.: A generalized sequential construction of exchangeable Gibbs partitions with application. Proceedings of S.Co. 2009, 14-16 September, Milano, Italy, pag-115–120, Maggioli Editore (2009)
3. Charalambides, C. A.: Combinatorial Methods in Discrete Distributions. Wiley, Hoboken NJ (2005)
4. Engen, S.: Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversity. Chapman and Hall Ltd (1978)
5. Fisher, R.A. Corbet, A.S., & Williams C.B.: The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal. Ecol.* **12**, 42–58 (1943)
6. Gnedin, A.: A species sampling model with finitely many types. *Elect. Comm. in Probab.* **15**, 79–88 (2010)
7. Gnedin, A. and Pitman, J.: Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, **138**, 3, 5674–5685 (2006)
8. Lijoi, A., Mena, R. and Prünster, I.: Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786 (2007)
9. Lijoi, A., Prünster, I., Walker, S.G.: Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.*, **18**, 1519–1547 (2008)
10. Pitman, J.: Exchangeable and partially exchangeable random partitions. *Prob. Th. Rel. Fields*, **102**, 145–158 (1995)
11. Pitman, J.: Some developments of the Blackwell-MacQueen urn scheme. In T.S. Ferguson, Shapley L.S., and MacQueen J.B., editors, *Statistics, Probability and Game Theory*, volume 30 of *IMS Lecture Notes-Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward, CA. (1996)
12. Pitman, J.: Poisson-Kingman partitions. In D.R. Goldstein, editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 40 of *Lecture Notes-Monograph Series*, pages 1–34. Institute of Mathematical Statistics, Hayward, California (2003)
13. Pitman, J.: Combinatorial Stochastic Processes. Ecole d’Eté de Probabilité de Saint-Flour XXXII - 2002. *Lecture Notes in Mathematics N. 1875*, Springer (2006)
14. Pitman, J., Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900 (1997)