Some further results for the two-parameter Poisson-Dirichlet partition model

Annalisa Cerquetti

Abstract We obtain some additional explicit results for the posterior partition generated by sampling from the random atoms of a two-parameter Poisson-Dirichlet model, conditional to a basic observed sample. Those results complement the large amount of conditional and unconditional results already obtained for this model, and have application in Bayesian nonparametric estimation in species sampling problems.

Key words: Exchangeability, Poisson-Dirichlet model, Random partitions, Species sampling problem

1 Introduction

The two parameter Poisson-Dirichlet model (Pitman and Yor, 1997) represents the most tractable and studied extension of the Ferguson-Dirichlet partition model in the large class of exchangeable Gibbs partitions devised by Gnedin and Pitman (2006). It is characterized by an exchangeable partition probabiliy function (EPPF) in the form

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

for $\alpha \in (-\infty, 1)$, $\theta > -\alpha$ and $(x)_{a\uparrow b} = x(x+b)\cdots(x+(a-1)b)$ the usual notation for generalized rising factorials. A bulk of results has been obtained for this model in the random partitions literature (see Pitman, 2006 for a comprehensive reference). Recently its mathematical tractability has been exploited to obtain explicit results in a Bayesian nonparametric approach to posterior estimation in species sampling

Annalisa Cerquetti
MEMOTEF, Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161 Roma. e-mail: annalisa.cerquetti@uniroma1.it

problems (see Lijoi et al. 2007, 2008, Cerquetti, 2011). In this setting the idea is to use the random discrete probability distribution corresponding to each exchangeable Gibbs partition, as a *prior* model on the unknown species relative abundances and to obtain a posterior predictive analysis conditionally on a basic n-sample for an additional m-sample of observations. Here we obtain a characterization of the two-parameter model with respect to a specific posterior distribution, and some additional explicit distributional results for the posterior partition probability function.

2 Main results

Let $(n_1,\dots,n_k)$ be the random partition induced by a sample $(X_1,\dots,X_n)$ of observations from a random discrete probability distribution $P(\cdot) = \sum_{i=1}^{\infty} P_i \delta_{X_i(\cdot)}$ corresponding to an EPPF in general Gibbs form of type α $p(n_1,\dots,n_k) = V_{n,k}\prod_{j=1}^{k}(1-\alpha)_{n_j-1}$. Let $X_1,\dots,X_m$ be an additional random m-sample from P. Then for k^* the number of new blocks, $s = \sum_{j=1}^{k^*} s_j$ the number of new observations in new blocks, and $(m_1,\dots,m_k)$ the allocation of the remaining $(m-s)$ new osbervations in old blocks, the following decomposition holds for the posterior distribution of the random partition $(s_1,\dots,s_{k^*},m_1,\dots,m_k)$ of the first m natural integers, given $(n_1,\dots,n_k)$

$$p_{\alpha,V_{n,k}}(s_1,\dots,s_{k^*},m_1,\dots,m_k|n_1,\dots,n_k) = \tag{2}$$

$$= p_{\alpha,V_{n,k}}(s_1,\dots,s_k^*|m_1,\dots,m_k,m-s,n_1,\dots,n_k)\,p_{\alpha,V_{n,k}}(m_1,\dots,m_k|m-s,n_1,\dots,n_k)$$

$$p_{\alpha,V_{n,k}}(m-s|n_1,\dots,n_k)$$

To obtain the explicit distribution for (2) under the two parameter EPPF (1) we need the explicit results for the three components.

First we obtain the following characterization of the two-parameter $PD(\alpha,\theta)$ model, with respect to the conditional distribution of S_m, the number of new observations in the new blocks, whose general form for α-Gibbs models is in Lijoi et al. (2008, cfr. Eq. (11)). Notice that to be consistent with Pitman's notation we make use of generalized Stirling numbers $S_{s,k}^{-1,-\alpha}$, (see Cerquetti, 2009).

Proposition 1. The extended two-parameter Poisson-Dirichlet model, for $\alpha \in (0,1)$ and $\theta > -\alpha$ and $\alpha < 0$ and $\theta = |\alpha|\xi$ for $\xi = 1,2,3,\dots$, is the unique Gibbs partition model of type $\alpha \in (-\infty,1)$ such that

$$\mathbb{P}(S_m = s|K_n = k) = \frac{1}{V_{n,k}}\binom{m}{s}(n-k\alpha)_{m-s\uparrow}\sum_{k^*=0}^{s} V_{n+m,k+k^*}S_{s,k^*}^{-1,-\alpha} =$$

$$= \binom{m}{s}(n-k\alpha)_{m-s}\left[\left(\frac{V_{n,k}}{V_{n+1,k}}\right)_m\right]^{-1}\left(\frac{V_{n+1,k+1}}{V_{n+1,k}}\right)_s.$$

Proof: The detailed proof is long. Here we just sketch it relies on the backward recursive relation characterizing the Gibbs weights $V_{n,k} = (n-k\alpha)V_{n+1,k} + V_{n+1,k+1}$, on the definition of generalized Stirling numbers as connection coefficients, $(x)_s = \sum_{y=0}^{s} S_{s,y}^{-1,-\alpha}(x)_{y\uparrow\alpha}$, and a known characterization of the $PD(\alpha,\theta)$ weights as the unique weights in the Gibbs class such that $V_{n,k} = C_k/V_n$. □

From Proposition 1. it follows that

$$\mathbb{P}(S_m = s | K_n = k) = \binom{m}{s} \frac{(\theta+k\alpha)_s (n-k\alpha)_{m-s}}{(\theta+n)_m}$$

which is a Beta-Binomial distribution of parameters $(m, (\theta+k\alpha), (n-k\alpha))$ hence

$$\mathbb{E}(S_m|K_n) = m\frac{(\theta+k\alpha)}{\theta+n}$$

and

$$Var(S_m|K_n) = \frac{m(\theta+k\alpha)(n-k\alpha)}{(\theta+n)^2} \frac{\theta+n+m}{\theta+n+1}.$$

Proposition 2. Given the observed sample partition $(n_1,\ldots,n_k)$ and the number $m-s$ of new observations in old blocks, the random allocation $(M_1,\ldots,M_k)$ in the k old blocks follows a Dirichlet Multinomial (or Multivariate Polya urn) distribution of parameters $(m-s, n_1-\alpha, \ldots, n_k-\alpha)$

$$Prob_{\alpha,\theta}(m_1,\ldots,m_k|m-s,n_1,\ldots,n_k) = \frac{m-s!}{\prod_{j=1}^{k} m_j!} \frac{\prod_{j=1}^{k}(n_j-\alpha)_{m_j}}{(n-k\alpha)_{m-s}}. \tag{3}$$

Proof. By the theory of the two-parameter Poisson-Dirichlet, the random vector arises by a Polya *urn model* construction for a k colors urn with initial composition $(n_1-\alpha,\ldots,n_k-\alpha)$. The thesis follows by a known result in probability theory about the number of successes in the different classes for a multicolor Polya urn. □

Now as already proved in Cerquetti (2011), conditionally given the basic sample $(n_1,\ldots,n_k)$, the number $m-s$ of new observations in old blocks, and the allocation $(m_1,\ldots,m_k)$ of the $m-s$ observation in old blocks, the partition of the s additional observations in new blocks follows a $PD(\alpha, \theta+k\alpha)$ partition model, hence

$$p_{\alpha,\theta}(s_1,\ldots,s_{k^*}|m_1,\ldots,m_k,m-s,n_1,\ldots,n_k) = \frac{(\theta+k\alpha+\alpha)_{k^*-1\uparrow\alpha}}{(\theta+k\alpha+1)_{s-1}} \prod_{j=1}^{k^*}(1-\alpha)_{s_j-1}. \tag{4}$$

In fact it corresponds to the random partition obtained by the operation of *deletion of classes* (cfr. Pitman, 2003, Sect. 4.3) and (4) follows by a known characterization of the two parameter model as the unique EPPF such that the *deleted* partition is still a *PD* partition with updated parameters $(\alpha, \theta+k\alpha)$.

By the previous considerations it follows that:

Proposition 3. Under the two parameter Poisson-Dirichlet (α,θ) model the posterior partition probability function of the new observations in old and new blocks, for $s=\sum_j s_j$ the random number of new observations in new block, is given by

$$p_{\alpha,\theta}(s_1,\dots,s_k^*,m_1,\dots,m_k|n_1,\dots,n_k)=$$

$$=\binom{m}{s}\frac{(m-s)!}{\prod_{j=1}^k m_j!}\frac{\prod_{j=1}^k (n_j-\alpha)_{m_j}(\theta+k\alpha)_{k^*\uparrow\alpha}}{(\theta+n)_m}\prod_{j=1}^k (1-\alpha)_{s_j-1}. \qquad (5)$$

Proof. The result easily arises by combining the conditional distributions obtained for the components of the decomposition in (2) and by elementary combinatorial calculus for generalized rising factorials. □

Remark 4. Notice that the random partition in (5) is not exchangeable, while by marginalizing $p_{\alpha,\theta}(s_1,\dots,s_{k^*},m_1,\dots,m_k|n_1,\dots,n_k)$ with respect to $(m_1,\dots,m_k)$, by means of a known extension of the multinomial theorem to rising factorials, one recovers the restricted EPPF to the first s positive integers $p_{\alpha,\theta}(s_1,\dots,s_k^*|n_1,\dots,n_k)$ as in Lijoi et al. (2008) (cfr. Cerquetti, 2009) given by

$$p_{\alpha,\theta}(s_1,\dots,s_{k^*}|n_1,\dots,n_k)=\binom{m}{s}\frac{(n-k\alpha)_{m-s}(\theta+k\alpha)_{k^*\uparrow\alpha}}{(\theta+n)_m}\prod_{j=1}^k (1-\alpha)_{s_j-1}.$$

References

1. Cerquetti, A.: A generalized sequential construction of exchangeable Gibbs partitions with application. Proceedings of S.Co. 2009, 14-16 September, Milano, Italy. (2009)
2. Cerquetti, A.: A decomposition approach to Bayesian nonparametric estimation for species richness under two-parameter Poisson-Dirichlet priors. Proceedings of ASMDA, Rome, June 2011. (2011)
3. Gnedin, A. and Pitman, J.: Exchangeable Gibbs partitions and Stirling triangles. J. Math. Sci., 138, 3, 5674–5685. (2006)
4. Lijoi, A., Mena, R. and Prünster, I.: Bayesian nonparametric estimation of the probability of discovering new species. Biometrika **9**4, 769–786 (2007)
5. Lijoi, A., Prünster, I., Walker, S.G.: Bayesian nonparametric estimator derived from conditional Gibbs structures. Ann. Appl. Probab., **1**8, 1519–1547 (2008)
6. Pitman, J.: Combinatorial Stochastic Processes. Ecole d'Eté de Probabilité de Saint-Flour XXXII - 2002. Lecture Notes in Mathematics N. 1875, Springer. (2006)
7. Pitman, J., Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**, 855–900 (1997)