Electronic Journal of Statistics
Vol. 7 (2013) 697–716
ISSN: 1935-7524
DOI: 10.1214/13-EJS784

Marginals of multivariate Gibbs distributions with applications in Bayesian species sampling

Annalisa Cerquetti*

Department of Methods and Models for Economics, Territory and Finance
University of Rome "La Sapienza"
Via del Castro Laurenziano, 9 00161 Rome, Italy
e-mail: annalisa.cerquetti@gmail.com

Abstract: Gibbs partition models are the largest class of infinite exchangeable partitions of the positive integers generalizing the product form of the probability function of the two-parameter Poisson-Dirichlet family. Here we call into question the current approach to Bayesian nonparametric estimation in species sampling problems under Gibbs *priors*, which incorrectly relies on treating exchangeable partition probability functions (EPPFs) as multivariate distributions on compositions of the positive integers. We show that once those multivariate distributions are correctly derived, results for corresponding sampling formulas can be obtained, generalized and sometimes fixed, working with marginals and a known result on falling factorial moments of a sum of non independent indicators. We provide an application of our findings to a recently proposed Bayesian nonparametric estimation under Gibbs priors of the predictive probability to observe a species already observed a certain number of times.

AMS 2000 subject classifications: Primary 60G57, 62G05; secondary 62F15.
Keywords and phrases: Exchangeable Gibbs partitions, falling factorial moments, multivariate Gibbs distributions, sampling formulas, species sampling problems, two parameter Poisson-Dirichlet model.

Received December 2012.

1. Introduction

Exchangeable random partitions of the positive integers are consistent sequences $\Pi = (\Pi_n)$ of random partitions of finite sets $[n] := \{1, \dots, n\}$, such that, for each *particular* partition $\{A_1, \dots, A_k\}$ of $[n]$, where the blocks are assumed to be listed in order of appearance, for $|A_i| = n_i \ \forall i$,

$$\mathbb{P}(\Pi_n = \{A_1, \dots, A_k\}) = p(|A_1|, \dots, |A_k|),$$

for some non-negative *symmetric* function p of *compositions* $(n_1, \dots, n_k)$ of n, satisfying $p(1) = 1$ and $p(n_1, \dots, n_k) = \sum_j p(\dots, n_j + 1, \dots) + p(n_1, \dots, n_k, 1)$, called the *exchangeable partition probability function* (EPPF) determined by Π,

*Partially supported by grant PRIN MIUR:2008CEFF37.

(see [28] for a comprehensive reference). By Kingman's correspondence, ([21]), every exchangeable partition Π has the same distribution as one generated by an infinite exchangeable sequence (X_n) driven by some random discrete probability measure P representable as $P(\cdot) := \sum_{i=1}^{\infty} P_i \delta_{\hat{X}_i}(\cdot)$, for $\hat{X}_i$ i.i.d. with non-atomic distribution $H(\cdot)$ independent of the (P_i). By the exchangeable equivalence relation $i \sim j$ iff $X_i = X_j$, then

$$p(n_1, \ldots, n_k) = \sum_{(i_1, \ldots, i_k)} \mathbb{E}\left[\prod_{j=1}^{k} P_{i_j}^{n_j}\right], \tag{1}$$

where $(i_1, \ldots, i_k)$ ranges over all ordered k-tuples of distinct positive integers and (P_i) is any rearrangement of the ranked atoms $(P_i^{\downarrow})$ of P. Exchangeable *Gibbs* partitions ([14]) are the largest class of infinite exchangeable partitions with EPPF in the *Gibbs* product form

$$p_{\alpha,V}(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1-\alpha)_{n_j - 1}, \tag{2}$$

for $\alpha \in (-\infty, 1)$, $V = (V_{n,k})$ weights satisfying the backward recursive relation $V_{n,k} = (n - k\alpha)V_{n+1,k} + V_{n+1,k+1}$, for $V_{1,1} = 1$, and $(x)_y = (x)(x+1)\cdots(x+y-1)$. By Theorem 12 in [14] each exchangeable partition with EPPF in (2) arises as a probability mixture of extreme partitions, namely: Fisher's (1943) partitions ([12]) for $\alpha < 0$, Ewens (θ) partitions ([6, 21]) for $\alpha = 0$, and Poisson-Kingman conditional partitions driven by the stable subordinator ([27]) for $\alpha \in (0, 1)$. A particularly tractable example of (2) is the *two parameter* (α, θ) *Poisson-Dirichlet* model ([25, 29]), which is well-known to arise for $V_{n,k} = (\theta + \alpha)_{k-1\uparrow\alpha}/(\theta+1)_{n-1}$, for $\alpha \in (0,1)$, $\theta > -\alpha$ or $\alpha < 0$, $\theta = |\alpha|\xi$ for $\xi = 1, 2, \ldots$ and $(x)_{y\uparrow\alpha} = x(x+\alpha)\cdots(x+(y-1)\alpha)$.

When the order of the blocks is irrelevant, and interest is just in the sizes of the blocks, by an application of Eq. (2.7) in [28], given an infinite EPPF in the form (2), for each $n \geq 1$ the corresponding joint distribution of the random vector $(N_{1,n} \ldots, N_{K_n,n}, K_n)$ of the sizes and number of the blocks in *exchangeable random order* that, from now on, we term *multivariate Gibbs distribution* of parameters (α, V), is given by

$$\mathbb{P}_{\alpha,V}(N_1^{ex} = n_1, \ldots, N_{K_n}^{ex} = n_k, K_n = k) = \frac{n!}{\prod_{j=1}^{k} n_j!} \frac{1}{k!} V_{n,k} \prod_{j=1}^{k} (1-\alpha)_{n_j-1}. \tag{3}$$

The corresponding *Gibbs sampling formula*, encoding the *partition of* n by the vector of the numbers of blocks of different sizes, is obtained by the obvious change of variable in (3) and multiplying by the number $k!/\Pi_{i=1}^{n} c_i!$ of compositions of n providing the same partition of n, i.e. the same rearrangement in decreasing order, and corresponds to

$$\mathbb{P}_{\alpha,V}(C_{1,n} = c_1, \ldots, C_{n,n} = c_n) = n! V_{n,k} \prod_{i=1}^{n} \frac{[(1-\alpha)_{i-1}]^{c_i}}{(i!)^{c_i} c_i!}, \tag{4}$$

where $c_i = \sum_{j=1}^{k} 1\{n_j = i\}$, for $i = 1, \dots, n$, $\sum_{i=1}^{n} ic_i = n$ and $\sum_{i=1}^{n} c_i = k$. Note that this is the general *Gibbs* analog of the *Ewens sampling formula* ([6])

$$\mathbb{P}_\theta(C_{1,n} = c_1, \dots, C_{n,n} = c_n) = \frac{n!\theta^k}{(\theta)_n} \prod_{i=1}^{n} \frac{1}{(i)^{c_i} c_i!}, \tag{5}$$

which gives the distribution of the number of blocks of different sizes of the Dirichlet (θ) partition model, ([11, 20]), whose EPPF is well-known to arise for $\alpha = 0$ in the (α, θ) model. A comprehensive reference for the study of (5), also called *component frequency spectrum*, for general combinatorial random structures is [1].

In this paper we study marginals of (3), both conditional and unconditional, to 1) drastically simplify the Bayesian approach to nonparametric estimation in species sampling problems under Gibbs *priors* as introduced in [22] and further developed in [23], 2) generalize, and sometimes fix, results obtained in the same spirit, with a view toward estimation of rare species richness, in [9] and [10]. Species sampling problems arise in many contexts, like population genetics, ecology or biology, when interest lies in studying richness and diversity of large populations of different species whose relative abundances are unknown. In a Bayesian nonparametric approach one assumes that the species *labels* of the first n individuals observed are a sample from a *species sampling model* (Pitman, 1996) that is an infinite exchangeable sequence (X_n) driven by an almost surely discrete *de Finetti* measure P with prediction rule

$$\mathbb{P}(X_{n+1} \in \cdot | X_1, \dots, X_n) = \sum_{j=1}^{K_n} p_{j,n} \delta_{\hat{X}_j}(\cdot) + q_n H(\cdot),$$

for $p_{j,n} = p(\dots, n_j + 1, \dots)/p(n_1, \dots, n_k)$, $q_n = p(n_1, \dots, n_k, 1)/p(n_1, \dots, n_k)$, $p(\cdot)$ the EPPF determined by (X_n) and $(\hat{X}_1, \dots, \hat{X}_{K_n})$ the distinct values among $(X_1, \dots, X_n)$. Given the vector of the multiplicities $(n_1, \dots, n_k)$ of the first k species observed in a initial sample of size n, posterior predictive estimation on a further sample of m observations is then obtained under the *prior* assumption that the unknown relative abundances of the different species in the population (P_i) – the random atoms in the infinite series representation of P – belong to the Gibbs family, or which is the same, that p belongs to the model (2). Unfortunately, in both [22] and [23], where the posterior predictive analysis of Gibbs partition models has been firstly devised, like in [9] and [10], providing further estimation results based on conditional falling factorial moments of *components* of (4), *conditional* and *unconditional multivariate distributions* of the vector of sizes and number of the blocks are erroneously identified with *conditional* and *unconditional EPPFs* (cf. e.g. Eq. (2) in Lijoi *et al.* 2007, Eq. (3) and Proposition 1 and 3 in Lijoi *et al.* 2008), whereas the former are probability distributions on the space of random compositions of the positive integers, and the latter are probability functions of a particular partition of the sets $[n]$ for $n \geq 1$. Such a mistake heavily affects the complexity of the proofs and sometimes induces wrong results.

Here, after correctly identifying conditional and unconditional *multivariate Gibbs distributions*, we derive corresponding marginals and obtain in a direct

way *joint* falling factorial moments of any order of corresponding Gibbs sampling formulas, both conditional and unconditional, and explicit formulas for distributions of interest generalizing the particular cases obtained in [9]. Our analysis, besides providing a more effective technique for Bayesian nonparametric applications, establishes the first systematic study of joint multivariate distributions on spaces of compositions of the positive integers arising from Gnedin-Pitman's Gibbs partitions theory. The paper is organized as follows: in Section 2 we obtain marginals of (3) and, resorting to a result in [18] for a sum of non independent indicators, derive general formulas for joint falling factorial moments of (4), together with explicit marginal distributions and their expected values. In Section 3, first we fix the results in [23], then we derive *conditional multivariate Gibbs distributions* and their marginals, for sizes and number of *new* blocks induced by an additional m-sample. Exploiting the same technique adopted in Section 2, a complete analysis is performed for *conditional Gibbs sampling formulas*, which generalizes the results in [9], deriving *joint* falling factorial moments and general marginals for the entire Gibbs class. In Section 4 we introduce *multivariate Pólya-Gibbs distributions* arising by the conditional allocation of the additional sample in *old* blocks and derive joint falling factorial moments and marginals of the corresponding vector of counts. Finally, in Section 5, we apply the proposed technique to fix, by a very short proof, a wrong result on a Bayesian nonparametric estimator of the m-step ahead probability to detect at observation $n+m+1$ a species already observed a certain number of times, which recently appeared in Favaro *et al.* (2012b).

2. Marginals of multivariate Gibbs distributions

To obtain the marginal distributions for general Multivariate Gibbs distributions (3) it is enough to resort to the definition of generalized *central* Stirling numbers of parameters $(-1,-\alpha)$

$$S_{n,k}^{-1,-\alpha} = \frac{n!}{k!} \sum_{(n_1,\ldots,n_k)} \prod_{j=1}^{k} \frac{(1-\alpha)_{n_j-1}}{n_j!}, \tag{6}$$

where the sum ranges over all $(n_1,\ldots,n_k)$ compositions of n, (see Sect. A.3 for further details). From now on we refer to (3) omitting the *ex* power in the notation.

Proposition 1. *Under a general Gibbs partition model (2) of parameters (α,V), for each $n\geq 1$ the r-dimensional marginal of (3), for $0\leq k-r\leq n-\sum_{j=1}^{r} n_j$, is given by*

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(N_1=n_1,\ldots,N_r=n_r,K_n=k)\\&\quad=\frac{n!}{\prod_{j=1}^{r} n_j!(n-\sum_{j=1}^{r} n_j)!}\prod_{j=1}^{r}(1-\alpha)_{n_j-1}\frac{V_{n,k}}{k_{[r]}}S_{n-\sum_{j=1}^{r} n_j,k-r}^{-1,-\alpha},\end{aligned} \tag{7}$$

for $(x)_{[n]}=(x)(x-1)\cdots(x-n+1)$.

Proof. From (3)

$$\mathbb{P}_{\alpha,V}(N_1 = n_1, \dots, N_r = n_r, K_n = k)$$
$$= \frac{n!}{\prod_{j=1}^r n_j!} \prod_{j=1}^r (1-\alpha)_{n_j-1} \frac{V_{n,k}}{k!} \sum_{(b_1,\dots,b_{k-r})} \frac{1}{\prod_i b_i!} \prod_{i=1}^{k-r} (1-\alpha)_{b_i-1},$$

for $(b_1, \dots, b_{k-r})$ such that $b_i > 0\ \forall i$ and $\sum_i b_i = n - \sum_{j=1}^r n_j$. Multiplying and dividing by $(n - \sum_{j=1}^r n_j)!$ and $(k-r)!$ an application of (6) yields (7). □

By (6), for each Gibbs model (2), $\mathbb{P}_{\alpha,V}(K_n = k) = V_{n,k} S^{-1,-\alpha}_{n,k}$, and a close inspection of (7) shows that, for every i, the conditional law of N_i given K_n does not involve the choice of the specific Gibbs weights $(V_{n,k})$. As an example the law of N_1 given $K_n = k$, for $0 \le k-1 \le n-n_1$ and $n_1 = 1, \dots, n-k+1$, will correspond to

$$\mathbb{P}_\alpha(N_1 = n_1 | K_n = k) = \binom{n}{n_1} \frac{(1-\alpha)_{n_1-1}}{k} \frac{S^{-1,-\alpha,}_{n-n_1,k-1}}{S^{-1,-\alpha}_{n,k}}.$$

A formula for the joint law of $(N_1, \dots, N_r)$ on the event $\{K_n \ge r\}$ can be obtained by appropriate marginalization of (7).

2.1. Joint falling factorial moments of Gibbs sampling formulas

Joint falling factorial moments for the *Ewens' sampling formula* (5) of order $(r_1, \dots, r_n)$, for r_l non negative integers and $n - \sum_l l r_l \ge 0$, are in [7] (cf. Eq. (41.9)) and correspond to

$$\mathbb{E}_\theta \left[\prod_{l=1}^n (C_{l,n})_{[r_l]} \right] = \frac{n!}{(n - \sum_{l=1}^n l r_l)!} \frac{(\theta)_{n-\sum_{l=1}^n l r_l}}{(\theta)_n} \prod_{l=1}^n \left(\frac{\theta}{l} \right)^{r_l}.$$

Under the same conditions, the generalization to the (α, θ) *Poisson-Dirichlet* sampling formula has been obtained in [30] and is given by

$$\mathbb{E}_{\alpha,\theta} \left[\prod_{l=1}^n (C_{l,n})_{[r_l]} \right] = \frac{n!}{(n - \sum_{l=1}^n l r_l)!} \frac{(\theta+\alpha)_{\sum_l r_l - 1 \uparrow \alpha}}{(\theta+1)_{n-1}}$$
$$\times \prod_{l=1}^n \left(\frac{(1-\alpha)_{l-1}}{l!} \right)^{r_l} (\theta + \alpha \sum_l r_l)_{n - \sum l r_l}.$$

In the following Proposition we generalize those results to the *general Gibbs sampling formula* (4) by resorting to a result in [18], first established in [5] then studied in [19]. See also [16, 17].

Proposition 2. *Under a general (α, V) Gibbs partition model, joint falling factorial moments of the vector of counts $(C_{1,n}, \dots, C_{n,n})$ of order $(r_1, \dots, r_n)$*

for $\sum_l lr_l \le n$ *are given by*

$$\mathbb{E}_{\alpha,V}\left[\prod_{l=1}^{n}(C_{l,n})_{[r_l]}\right] = \frac{n!}{\prod_{l=1}^{n}(l!)^{r_l}}\frac{\prod_{l=1}^{n}[(1-\alpha)_{l-1}]^{r_l}}{(n-\sum_l lr_l)!} \times \sum_{k-\sum_l r_l=0}^{n-\sum_l lr_l} V_{n,k}S^{-1,-\alpha}_{n-\sum_l lr_l,k-\sum_l r_l}, \tag{8}$$

for $0 \le k-\sum_{l=1}^{n} r_l \le n-\sum_{l=1}^{n} lr_l$. *For* $r_l = r \le \lceil \frac{n}{l} \rceil$ *and* $r_j = 0$ *for every* $j \ne l$, *the* r*-th falling factorial moment of* $C_{l,n}$ *results*

$$\mathbb{E}_{\alpha,V}\left[(C_{l,n})_{[r]}\right] = \frac{n![(1-\alpha)_{l-1}]^r}{(l!)^r(n-lr)!}\sum_{k-r=0}^{n-rl} V_{n,k}S^{-1,-\alpha}_{n-rl,k-r}, \tag{9}$$

and corrects the summation limits in Eq. (11) in Favaro et al. (2012a).

Proof. For $K_n = k$, let $C_{l,n} = \sum_{j=1}^{k} 1\{N_j = l\}$. Then by a result for a sum of non independent indicators r.v.s in Johnson & Kotz (2005, Sect. 10.2), or Charalambides (2005, Example 1.12), for $r \le n$

$$\mathbb{E}_{\alpha,V}\left[(C_{l,n})_{[r]}\right] = \mathbb{E}_{\alpha,V}(\sum_{j=1}^{k} 1\{N_j = l\})_{[r]} = r!\sum_{(a_1,\ldots,a_r)} \mathbb{P}_{\alpha,V}(N_{a_1} = l,\ldots,N_{a_r} = l), \tag{10}$$

where the summation is extended over all r-combinations $(a_1,\ldots,a_r)$ of $\{1,\ldots,k\}$. Since in our case the number of blocks K_n is random, and the vector $(N_1,\ldots,N_r|K_n = k)$ is exchangeable then, for $l = 1,\ldots,n$,

$$\begin{aligned}\mathbb{E}_{\alpha,V}\left[(C_{l,n})_{[r]}\right] &= \sum_{k-r=0}^{n-rl}\mathbb{E}\left[(C_{l,n})_{[r]}|K_n = k\right]\mathbb{P}_{\alpha,V}(K_n = k)\\ &= \sum_{k-r=0}^{n-rl} r!\binom{k}{r}\mathbb{P}(N_1 = l,\ldots,N_r = l|K_n = k)\mathbb{P}_{\alpha,V}(K_n = k)\\ &= \sum_{k-r=0}^{n-rl} r!\binom{k}{r}\mathbb{P}_{\alpha,V}(N_1 = l,\ldots,N_r = l,K_n = k).\end{aligned}$$

By a similar argument

$$\begin{aligned}&\mathbb{E}_{\alpha,V}\left[\prod_{l=1}^{n}(C_{l,n})_{[r_l]}\right]\\ &= \sum_{k-\sum_l r_l=0}^{n-\sum_l lr_l}\left(\prod_{l=1}^{n} r_l!\right)\frac{k!}{\prod_l r_l!(k-\sum_l r_l)!}\\ &\quad\times \mathbb{P}_{\alpha,V}(N_1 = 1,\ldots,N_{r_1} = 1,\ldots,N_{\sum_l r_l - r_n+1} = n,\ldots,N_{\sum_l r_l} = n,K_n = k).\end{aligned} \tag{11}$$

Inserting (7) in (11) the result follows. □

Remark 1. Notice that (8) generalizes the result in [9] Eq. (11), stated in terms of *generalized factorial coefficients*, (see Sect. A.3 for the relationship with generalized Stirling numbers), which corresponds, after fixing the summation limits, to (9). Actually the summation limits in (8), (9) and (11) are written in an unconventional way. This is to highlight that they must correspond to the parameters of the generalized Stirling numbers, as it is compulsory for the summation to make sense, and allows to detect some inconsistent summations in the results in [9]. We will adopt this unconventional notation all over the paper.

The distribution of $C_{l,n}$ for a general (α, V) Gibbs model, generalizing Proposition 2 (Dirichlet case), Proposition 4 (two parameter Poisson-Dirichlet case) and Proposition 8 (Gnedin-Fisher case ([13])) in [9], arises by the known relationship between discrete probability distributions and falling factorial moments (see e.g. [18]), and corresponds to

$$\mathbb{P}_{\alpha,V}(C_{l,n} = x) = \frac{n![(1-\alpha)_{l-1}]^x}{x!(l!)^x} \sum_{r=0}^{\lceil \frac{n}{l} \rceil - x} \frac{(-1)^r [(1-\alpha)_{l-1}]^r}{r!(l!)^r (n-rl-lx)!} \times \sum_{k-r-x=0}^{n-rl-xl} V_{n,k} S^{-1,-\alpha}_{n-rl-xl,k-r-x}, \tag{12}$$

for $x = 0, \ldots, \lceil n/l \rceil$. Its with expected value follows from (9) for $r = 1$. The distribution and the expected value of $C_{1,n}$, the number of singletons, generalizing (41.10) and (41.11) in [7] to the entire Gibbs family, follow for $l = 1$.

3. Conditional multivariate Gibbs distributions

The study of *conditional exchangeable random partitions*, i.e. exchangeable partitions starting with an initial allocation of the first n natural integers in a certain number j of blocks, has been initiated in [22], in view of proposing a Bayesian conditional nonparametric estimation of the richness of a population of species under *priors* on the unknown relative abundances belonging to the Gibbs class. This is the first paper suffering from the problem we pointed out in the Introduction. In this setting the standard setup is as follows: given an initial sample of n observations, inducing a partition of the first n integers in j blocks with observed multiplicities $(n_1, \ldots, n_j)$, a further sample of size $m \geq 1$ is collected. Let $K_m^{(n)}$ be the number of *new* blocks generated by the additional $m \geq 1$ integers, $(S_1, \ldots, S_{K_m^{(n)}})$ the vector of the sizes of the *new* blocks in exchangeable random order, $\Sigma_m = \sum_{i=1}^{K_m^{(n)}} S_i$ the total number of *new* integers in *new* blocks and $(M_{1,m}, \ldots, M_{j,m})$ the vector of the allocations of the additional $m - \Sigma_m$ observations in the j old blocks. The next Proposition corrects the mistake in [23] fixing formulas (9) in Proposition 1 and (19) in Proposition 3, both missing the combinatorial coefficients.

Proposition 3. *Under a general (α, V) Gibbs partition model, given the initial allocation of n integers in j blocks, the joint conditional distribution of*

$(K_m^{(n)}, \Sigma_m, S_1, \ldots, S_{K_m^{(n)}})$, *for* $S_1, \ldots, S_{K_m^{(n)}}$ *in exchangeable random order, that we term conditional Multivariate Gibbs distribution of parameters* (α, m, n, j), *for* $s_i \geq 1\ \forall i$ *and* $s = \sum_{i=1}^k s_i$, *corresponds to*

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(K_m^{(n)} = k, \Sigma_m = s, S_1 = s_1, \ldots, S_{K_m^{(n)}} = s_k | n_1, \ldots, n_j)\\ &= \frac{m!}{s_1! \cdots s_k! k! m-s!} \frac{V_{n+m,j+k}}{V_{n,j}} (n - j\alpha)_{m-s} \prod_{i=1}^k (1-\alpha)_{s_i-1}. \end{aligned} \tag{13}$$

Conditioning on Σ_m *yields*

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(K_m^{(n)} = k, S_1 = s_1, \ldots, S_{K_m^{(n)}} = s_k | K_n = j, \Sigma_m = s)\\ &= \frac{s!}{s_1! \cdots s_k! k!} \frac{V_{n+m,j+k}}{\sum_{i=0}^s V_{n+m,j+i} S_{s,i}^{-1,-\alpha}} \prod_{i=1}^k (1-\alpha)_{s_i-1}. \end{aligned} \tag{14}$$

Proof. By easy combinatorics and telescoping product from the one-step prediction rule under (2) the conditional probability of any *particular* partition of the set $[n+m]-[n]$ in k new blocks of size $s_i \geq 1$, $\sum_{i=1}^k s_i = s$, $s \leq m$, with allocation in j old blocks of $m_i \geq 0$, $\sum_{i=1}^j m_i = m - s$ integers, corresponds to

$$p_{\mathbf{m}}^{\mathbf{s}}(\mathbf{n}) = \frac{V_{n+m,j+k}}{V_{n,j}} \prod_{i=1}^j (n_i - \alpha)_{m_i} \prod_{i=1}^k (1-\alpha)_{s_i-1}, \tag{15}$$

for $\mathbf{n} = (n_1, \ldots, n_j)$, $\mathbf{s} = (s_1, \ldots, s_k)$ and $\mathbf{m} = (m_1, \ldots, m_j)$. The joint full conditional of $(K_m^{(n)}, S_1, \ldots, S_{K_m^{(n)}}, \Sigma_m, M_{1,m}, \ldots, M_{j,m})$ easily arises from (3) and the standard multinomial coefficient and corresponds to

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(S_1 = s_1, \ldots, S_k = s_k, \Sigma_m = s, K_m^{(n)} = k, M_{1,m} = m_1, \ldots, M_{j,m} = m_j | \mathbf{n})\\ &= \frac{m!}{\prod_{i=1}^k s_i! k! \prod_{i=1}^j m_i!} \frac{V_{n+m,j+k}}{V_{n,j}} \prod_{i=1}^k (1-\alpha)_{s_i-1} \prod_{i=1}^j (n_i - \alpha)_{m_i}. \end{aligned} \tag{16}$$

By marginalizing with respect to all possible allocations $(m_1, \ldots, m_j)$ (13) follows. (14) is an easy consequence of Eq. (11) in [23]. Additionally, by (14) and Eq. (4) in [22], the conditional distribution given $K_m^{(n)}$, does not involve the choice of the Gibbs model as in the unconditional case, and fixes formula (34) in [23]. □

Remark 2. Further results for the conditional moments of any order of $K_m^{(n)}$ and for the conditional asymptotic distribution of a proper normalization of $K_m^{(n)}$ under (α, θ) Poisson-Dirichlet partition models are in [8]. A simplified approach to the posterior analysis of the two-parameter model, exploiting the *deletion of classes property* and the Beta-Binomial distribution of $\Sigma_m | K_n = j$, is in [2]. A general result for *conditional* α *diversity* for Poisson-Kingman partition models driven by the stable subordinator ([27]) has been obtained in [3].

Given the corrected formulas in Proposition 3, the derivation of estimators for quantities of interest in Bayesian nonparametric species sampling modeling becomes an easy task. In the next Proposition, mimicking the technique adopted in the previous section for the unconditional case, we derive marginals of (13) as the tools to obtain *joint* conditional falling factorial moments of the *conditional Gibbs sampling formula*, which accounts for the conditional probability distribution of the number of new blocks of different sizes. For $W_{l,m} = \sum_{i=1}^{K_m^{(n)}} 1\{S_i = l\}$, $1 \leq l \leq m$, $\sum_{l=1}^{m} W_{l,m} = K_m^{(n)}$ and $\sum_l l W_{l,m} = \Sigma_m$, by the obvious change of variable in (13), and multiplying by $k!/\prod_{l=1}^{m} w_l!$, it corresponds to

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(W_{1,m} = w_1, \ldots, W_{m,m} = w_m | n_1, \ldots, n_j)\\ &\quad = \frac{m! V_{n+m,j+k}}{V_{n,j}} \frac{(n - j\alpha)_{m-s}}{(m-s)!} \prod_{l=1}^{m} \frac{[(1-\alpha)_{l-1}]^{w_l}}{(l!)^{w_l} w_l!},\end{aligned} \tag{17}$$

for $\sum_l w_l = k$ and $\sum_l l w_l = s$. In what follows we will resort to the convolution relation which defines *non-central* generalized Stirling numbers $S_{n,k}^{-1,-\alpha,\gamma}$ in terms of *central* generalized Stirling numbers (see Eq. (35) in Section A.3).

Proposition 4. *Under a general* (α, V) *Gibbs partition model the r-dimensional marginal of (13), for* $(s_1, \ldots, s_r) : \sum_i s_i \leq s \leq m$ *and* $0 \leq k - r \leq m - \sum_{i=1}^{r} s_i$, *is given by*

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(S_1 = s_1, \ldots, S_r = s_r, K_m^{(n)} = k | n_1, \ldots, n_j)\\ &\quad = \frac{m![\prod_{i=1}^{r}(1-\alpha)_{s_i-1}]}{\prod_{i=1}^{r} s_i!(m - \sum_{i=1}^{r} s_i)!} \frac{(k-r)!}{k!} \frac{V_{n+m,j+k}}{V_{n,j}} S_{m-\sum_{i=1}^{r} s_i, k-r}^{-1,-\alpha,-(n-j\alpha)}.\end{aligned} \tag{18}$$

Proof. Multiplying and dividing (13) by $(s - \sum_{i=1}^{r} s_i)!$ and $(m - \sum_{i=1}^{r} s_i)!$ and marginalizing yields

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(S_1 = s_1, \ldots, S_r = s_r, K_m^{(n)} = k | n_1, \ldots, n_j)\\ &\quad = \frac{m![\prod_{i=1}^{r}(1-\alpha)_{s_i-1}]}{\prod_{i=1}^{r} s_i!(m - \sum_{i=1}^{r} s_i)!} \frac{1}{k!} \frac{V_{n+m,j+k}}{V_{n,j}}\\ &\qquad \times \sum_{s - \sum_{i=1}^{r} s_i = k-r}^{m - \sum_{i=1}^{r} s_i} \frac{(m - \sum_{i=1}^{r} s_i)!(n - j\alpha)_{m-s}}{(s - \sum_{i=1}^{r} s_i)!(m-s)!}\\ &\qquad \times \sum_{(b_1, \ldots, b_{k-r})} \frac{(s - \sum_{i=1}^{r} s_i)!}{\prod_i b_i!} \prod_{i=1}^{k-r} (1-\alpha)_{b_i - 1} =\end{aligned}$$

further multiplying and dividing by $(k-r)!$ we obtain

$$\begin{aligned}&= \frac{m![\prod_{i=1}^{r}(1-\alpha)_{s_i-1}]}{\prod_{i=1}^{r} s_i!(m - \sum_{i=1}^{r} s_i)!} \frac{(k-r)!}{k!} \frac{V_{n+m,j+k}}{V_{n,j}}\\ &\quad \times \sum_{s - \sum_{i=1}^{r} s_i = k-r}^{m - \sum_{i=1}^{r} s_i} \binom{m - \sum_{i=1}^{r} s_i}{s - \sum_{i=1}^{r} s_i} (n - j\alpha)_{m-s} S_{s - \sum_{i=1}^{r} s_i, k-r}^{-1,-\alpha},\end{aligned}$$

and the result follows by an applications of (35). □

The next Proposition, which generalizes Theorem 2 in [9], provides the joint falling factorials moments of any order of (17). Let $W_{l,m}^{(n)}$ stands for $W_{l,m}|n_1,\dots,n_j$.

Proposition 5. *Under a general (α, V) Gibbs partition model, joint falling factorial moments of order $(r_1,\dots,r_m)$ of the conditional sampling formula (17), for $m-\sum_l lr_l \geq 0$, are given by*

$$\mathbb{E}_{\alpha,V}\left[\prod_{l=1}^{m}(W_{l,m}^{(n)})_{[r_l]}\right] = \frac{m!\prod_{l=1}^{m}[(1-\alpha)_{l-1}]^{r_l}}{(m-\sum_l lr_l)!\prod_l (l!)^{r_l}}\frac{1}{V_{n,j}}\sum_{k-\sum_l r_l=0}^{m-\sum_l lr_l} V_{n+m,j+k}S^{-1,-\alpha,-(n-j\alpha)}_{m-\sum_l lr_l,k-\sum_l r_l}. \quad (19)$$

For $r_l = r \leq \lceil \frac{m}{l} \rceil$ and $r_j = 0$ for $j \neq l$ then

$$\mathbb{E}_{\alpha,V}[(W_{l,m}^{(n)})_{[r]}] = \frac{m!}{(m-rl)!}\frac{[(1-\alpha)_{l-1}]^r}{(l!)^r}\frac{1}{V_{n,j}}\sum_{k-r=0}^{m-rl} V_{n+m,j+k}S^{-1,-\alpha,-(n-j\alpha)}_{m-rl,k-r}, \quad (20)$$

which fixes the admissible values of r in Theorem 2. in [9] which is expressed in terms of non central generalized factorial coefficients (see Eq. (36) in A.3).

Proof. By the analogy between (18) and (7) the proof moves along the same lines as the proof of Proposition 2. □

In [9], (cf. Propositions 2, 6 and 9), explicit one dimensional marginals of (17) have been derived for the Dirichlet (θ), the (α,θ) Poisson-Dirichlet and the Gnedin-Fisher (γ) ([13]) partition models. By (20) the general result for the entire Gibbs family, providing the conditional analog of (12), corresponds to

$$\mathbb{P}_{\alpha,V}(W_{l,m}^{(n)} = x) = \frac{[(1-\alpha)_{l-1}]^x}{x!(l!)^x}\frac{m!}{V_{n,j}}\sum_{r=0}^{\lceil \frac{m}{l} \rceil - x}\frac{(-1)^r[(1-\alpha)_{l-1}]^r}{r!(l!)^r(m-rl-xl)!} \times \sum_{k-r-x=0}^{m-rl-xl} V_{n+m,j+k}S^{-1,-\alpha,-(n-j\alpha)}_{m-rl-xl,k-r-x}, \quad (21)$$

for $x = 0,\dots,\lceil m/l \rceil$. Its expected value, which provides the Bayesian nonparametric estimator under quadratic loss function, for the number of *new* species represented l times, arises from (20) for $r = 1$ and corresponds to Eq. (17) in [9] expressed in terms of generalized non central factorial coefficients. The conditional distribution of the number of new singleton species $W_{1,m}^{(n)}$ and its expected value follow easily for $l = 1$.

4. Multivariate Pólya-Gibbs distributions

In this Section we focus on the conditional random allocation of the additional m integers in the j *old* blocks. First we derive the conditional joint distribution

of the random vector $(M_{1,m},\dots,M_{j,m},\Sigma_m)$ of the numbers of *new* observations falling in the j *old* blocks and of the total number of *new* observations Σ_m falling in *new* blocks, for $\sum_i M_{i,n}+\Sigma_m=m$. Then, similarly to the previous sections, we move attention to the corresponding vector of counts and its joint falling factorial moments. From (16) an application of (6) yields

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(M_{1,m}=m_1,\dots,M_{j,m}=m_j,\Sigma_m=s|n_1,\dots,n_j)\\&=\frac{m!}{\prod_{i=1}^j m_i!s!}\prod_{i=1}^j(n_i-\alpha)_{m_i}\sum_{k=0}^s\frac{V_{n+m,j+k}}{V_{n,j}}S^{-1,-\alpha}_{s,k},\end{aligned}\tag{22}$$

for $m_i\geq 0$ for $i=1,\dots,j$ and $\sum_{i=1}^j m_i=m-s$.

Remark 3. Note here that, since the number of old blocks is fixed, (22) may be interpreted as a generalization of *multivariate Pólya distributions* (Dirichlet mixtures of multinomial distributions). If Q_V is the conditional law, given $(n_1,\dots,n_j)$, of the vector $(\tilde{P}_{1,n},\dots,\tilde{P}_{j,n},R_{j,n})$, for $\tilde{P}_{i,n}=\tilde{P}_i|n_1,\dots,n_j$ the conditional random relative abundance of the i-th species to appear, and $R_{j,n}=1-\sum_{i=1}^j\tilde{P}_{i,n}$, then (22) turns out to be a Q_V-multinomial mixture that we term *multivariate Pólya-Gibbs distribution* of parameters $(m,n_1-\alpha,\dots,n_j-\alpha,V)$. Moreover Q_V will be the limit law, for $m\to\infty$, of the random vector $\left(M^{(n)}_{1,m}/m,\dots,M^{(n)}_{j,m}/m,\Sigma_m/m\right)$, where $M^{(n)}_{i,m}$ stands for a component of (22). Notice that for the two-parameter Poisson-Dirichlet (α,θ) model, by a result in [26], (see Sect. 3.7, Corollary 20),

$$(\tilde{P}_{1,n},\dots,\tilde{P}_{j,n},R_{j,n})\sim Dir[n_1-\alpha,\dots,n_j-\alpha,\theta+j\alpha],$$

and substituting $V_{n,j}=(\theta+\alpha)_{j-1\uparrow\alpha}/(\theta+1)_{n-1}$ in (22) yields

$$\begin{aligned}&\mathbb{P}_{\alpha,\theta}(M_{1,m}=m_1,\dots,M_{j,m}=m_j,\Sigma_m=s|n_1,\dots,n_j)\\&=\frac{m!}{\prod_{i=1}^j m_i!s!}\frac{\prod_{i=1}^j(n_i-\alpha)_{m_i}(\theta+j\alpha)_s}{(\theta+n)_m},\end{aligned}\tag{23}$$

which is a proper *multivariate Pólya distribution* of parameters $(m,n_1-\alpha,\dots,n_j-\alpha,\theta+j\alpha)$.

The next Proposition provides the general marginal of (22) that we need to obtain joint falling factorial moments of the corresponding vector of counts.

Proposition 6. *Under a general (α,V) Gibbs model, given the initial allocation of n integers in j old blocks, the conditional joint marginal distribution of the vector of the sizes $(M_{1,m}\dots,M_{r,m})$ of the additional new observations falling in the first r old blocks, for $1\leq r\leq j$, corresponds to*

$$\begin{aligned}&\mathbb{P}_{\alpha,V}(M_{1,m}=m_1,\dots,M_{r,m}=m_r|n_1,\dots,n_j)\\&=\frac{m!\prod_{i=1}^r(n_i-\alpha)_{m_i}}{\prod_{i=1}^r m_i!(m-\sum_{i=1}^r m_i)!}\sum_{k=0}^{m-\sum_{i=1}^r m_i}\frac{V_{n+m,j+k}}{V_{n,j}}S^{-1,-\alpha,-(n-(j-r)\alpha-\sum_{i=1}^r n_i)}_{m-\sum_{i=1}^r m_i,k}.\end{aligned}\tag{24}$$

Proof. By (22), the joint marginal of the sizes of the first r blocks and Σ_m is obtained as

$$\mathbb{P}_{\alpha,V}(M_{1,m}=m_1,\dots,M_{r,m}=m_r,\Sigma_m=s|n_1,\dots,n_j)$$
$$=\frac{m!\prod_{i=1}^r(n_i-\alpha)_{m_i}(n-j\alpha-\sum_{i=1}^r n_i+r\alpha)_{m-s-\sum_{i=1}^r m_i}}{\prod_{i=1}^r m_i!(m-s-\sum_{i=1}^r m_i)!s!}\sum_{k=0}^{s}\frac{V_{n+m,j+k}}{V_{n,j}}S^{-1,-\alpha}_{s,k},$$

marginalizing with respect to Σ_m, multiplying and dividing by $(m-\sum_{i=1}^r m_i)!$ and then changing the order of marginalization yields

$$\mathbb{P}_{\alpha,V}(M_{1,m}=m_1,\dots,M_{r,m}=m_r|n_1,\dots,n_j)$$
$$=\frac{m!\prod_{i=1}^r(n_i-\alpha)_{m_i}}{\prod_{i=1}^r m_i!(m-\sum_{i=1}^r m_i)!}\sum_{k=0}^{m-\sum_{i=1}^r m_i}\frac{V_{n+m,j+k}}{V_{n,j}}\sum_{s=k}^{m-\sum_{i=1}^r m_i}\binom{m-\sum_{i=1}^r m_i}{s}$$
$$\times(n-j\alpha-\sum_{i=1}^r n_i+r\alpha)_{m-s-\sum_{i=1}^r m_i}S^{-1,-\alpha}_{s,k},$$

and the result follows by an application of (35). □

Now let $O^{(n)}_{l,m}=\sum_{i:n_i\le l}1\{n_i+M_{i,m}=l|n_1,\dots,n_j\}$, for $l=1,\dots,n+m$, be the number of *old* blocks of size l after the allocation of the additional m-sample, then, to obtain the joint falling factorial moments of any order of $(O^{(n)}_{1,m},\dots,O^{(n)}_{n+m,m})$ we exploit the result (10) recalled in the proof of Proposition 2, namely

$$\mathbb{E}\left[(O^{(n)}_{l,m})_{[r]}\right]=r!\sum_{(\xi_1,\dots,\xi_r)}\mathbb{P}(M_{\xi_1}=l-n_1,\dots,M_{\xi_r}=l-n_r|n_1,\dots,n_j). \quad (25)$$

Specializing (24) for $m_i=l-n_i$ the next result follows from (25) as the analog of Propositions 2 and 5.

Proposition 7. *Under a general (α,V) Gibbs model, the joint falling factorial moments of the vector of the number of old blocks of different size $(O^{(n)}_{1,m},\dots,O^{(n)}_{n+m,m})$, after the allocation of the additional m-sample, given the initial allocation $(n_1,\dots,n_j)$ are given by*

$$\mathbb{E}_{\alpha,V}\left[\left(\prod_{l=1}^{n+m}(O^{(n)}_{l,m})_{[r_l]}\right)\right]$$
$$=\prod_{l=1}^{n+m}r_l!\sum_{(\Xi_{r_1},\dots,\Xi_{r_{n+m}})}\frac{m!\prod_{l=1}^{n+m}\prod_{i=1}^{r_l}(n_{\xi_i}-\alpha)_{l-n_{\xi_i}}}{\prod_{l=1}^{n+m}\prod_{i=1}^{r_l}(l-n_{\xi_i})!(m-\sum_l lr_l+\sum_l\sum_{i=1}^{r_l}n_{\xi_i})!}$$
$$\times\sum_{k=0}^{m-\sum_l lr_l+\sum_l\sum_{i=1}^{r_l}n_{\xi_i}}\frac{V_{n+m,j+k}}{V_{n,j}}S^{-1,-\alpha,-(n-(j-\sum_l r_l)\alpha-\sum_l\sum_i n_{\xi_i})}_{m-\sum_l lr_l+\sum_l\sum_{i=1}^{r_l}n_{\xi_i},k},$$

for $\boldsymbol{\Xi}_{r_1} = (\xi_1, \ldots, \xi_{r_1}), \ldots, \boldsymbol{\Xi}_{r_{n+m}} = (\xi_{\sum_l r_l - r_{n+m}}, \ldots, \xi_{\sum_{l=1}^{n+m} r_l})$, $\xi_i : n_{\xi_i} \leq l$, *and each* $\boldsymbol{\Xi}_{r_l}$ *ranging over all the combinations of* r_l *elements of* j. *For* $r_l = r$ *and* $r_j = 0$ *for* $j \neq l$, *then*

$$\mathbb{E}_{\alpha,V}\left[(O_{l,m}^{(n)})_{[r]}\right] = r! \sum_{(\xi_1,\ldots,\xi_r)} \frac{m! \prod_{i=1}^{r} (n_{\xi_i} - \alpha)_{l-n_{\xi_i}}}{\prod_{i=1}^{r}(l - n_{\xi_i})!(m - rl + \sum_{i=1}^{r} n_{\xi_i})!} \times \sum_{k=0}^{m-lr+\sum_{i=1}^{r} n_{\xi_i}} \frac{V_{n+m,j+k}}{V_{n,j}} S^{-1,-\alpha,-(n-(j-r)\alpha-\sum_{i=1}^{r} n_{\xi_i})}_{m-lr+\sum_{i=1}^{r} n_{\xi_i},k} \tag{26}$$

for $\xi_i : n_{\xi_i} \leq l$, *which agrees with the result in Theorem 1. in [9] apart from some typos in the summation limits.*

Proof. By the analogy between (24) and (7) the proof moves along the same lines as the proof of Proposition 2. □

From (26) the the conditional marginal law of $O_{l,m}^{(n)}$, under (α, V) Gibbs models, generalizing Proposition 5 (two-parameter Poisson-Dirichlet case) and Proposition 9 (one parameter Gnedin-Fisher case) in [9] to the entire (α, V) Gibbs family, is given by

$$\begin{aligned}
&\mathbb{P}_{\alpha,V}(O_{l,m}^{(n)} = y) \\
&= \sum_{r=0}^{\lceil \frac{m-rl+\sum_{i=1}^{r+y} n_{\xi_i}}{l} \rceil - y} \frac{(-1)^r (r+y)!}{y! r!} \frac{1}{V_{n,j}} \\
&\quad \times \sum_{(\xi_1,\ldots,\xi_{r+y})} \frac{m!}{\prod_{i=1}^{r+y}(l - n_{\xi_i})!(m - rl - yl + \sum_{i=1}^{r+y} n_{\xi_i})!} \prod_{i=1}^{r+y} (n_{\xi_i} - \alpha)_{l-n_{\xi_i}} \\
&\quad \times \sum_{k=0}^{m-rl-ly+\sum_{i=1}^{r+y} n_{\xi_i}} V_{n+m,j+k} S^{-1,-\alpha,-(n-(j-r-y)\alpha-\sum_{i=1}^{r+y} n_{\xi_i})}_{m-lr-ly+\sum_{i=1}^{r+y} n_{\xi_i},k}.
\end{aligned}$$

Its expected value, which plays the role of the Bayesian nonparametric estimator, under quadratic loss function, of the number of *old* species represented l times, follows easily from (26) for $r = 1$, and agrees with Eq. (15) in [9] while fixing some typos in the summation limits.

Remark 4. Relying on the technique presented in this paper, conditional r-th falling factorial moments of $Z_{l,m}^{(n)} = O_{l,m}^{(n)} + W_{l,m}^{(n)}$, the total number of *old* and *new* blocks of size l after the allocation of the additional m-sample, as derived in Th. 3 in [9] by means of a very complex procedure, may be obtained in a straightforward way by the full conditional joint distribution (16). Multiplying by the way to choose t blocks among the *old* and $r - t$ among the *new* for every t, combining (20) and (26) we get

$$\mathbb{E}_{\alpha,V}\left[(Z^{(n)}_{l,m})_{[r]}\right]$$
$$= \sum_{t=0}^{r}\binom{r}{t}t!\sum_{(\xi_{i_1},\ldots,\xi_{i_t})}\frac{m![(1-\alpha)_{l-1}]^{r-t}\prod_{i=1}^{t}(n_{\xi_i}-\alpha)_{l-n_{\xi_i}}}{\prod_{i=1}^{t}(l-n_{\xi_i})!(l!)^{r-t}(m-tl+\sum_{i=1}^{t}n_{\xi_i}-(r-t)l)!}$$
$$\times \sum_{k-r+t=0}^{m-rl+\sum n_{\xi_i}}\frac{V_{n+m,k+j}}{V_{n,j}}S^{-1,-\alpha,-(n-j\alpha-\sum n_{\xi_i}+t\alpha)}_{m-rl+\sum n_{\xi_i},k-r+t}$$

which agrees with Theorem 3. in [9].

5. Application

In species sampling problems, given a basic n-sample $(n_1,\ldots,n_j)$, interest may be in estimating the *probability* to observe at step $n+m+1$ a species already represented l times both belonging to an *old* species or to a *new* species eventually arising in the m-additional sample which is still to be observed. This is the topic of a recent paper by Favaro *et al.* (2012b), ([10]), and can be seen as a generalization of the problem of estimating the *discovery probability*, i.e. the probability to discover at step $n+m+1$ a *new* species, not represented in the previous $n+m$ observations, already solved in [22]. In this Section we show how working with marginals of conditional multivariate Gibbs distributions greatly simplifies the derivation of the results obtained in [10] and additionally allows to fix a mistake. First recall that by sequential construction of exchangeable Gibbs partitions, the probability to observe an *old* species observed l times in the basic n-sample at observation $n+1$, easily follows by one-step prediction rules for general (α,V) Gibbs EPPFs (see e.g. [28]). For $c_{l,n}=\sum_{i=1}^{j}1\{n_i=l\}$, for $l=1,\ldots,n$ then

$$p_{l,n}(n_1,\ldots,n_j)=c_{l,n}\frac{p(n_1,\ldots,l+1,\ldots,n_j)}{p(n_1,\ldots,l,\ldots,n_j)}=c_{l,n}\frac{V_{n+1,j}}{V_{n,j}}(l-\alpha).$$

By a similar argument, given a basic sample $(n_1,\ldots,n_j)$, but assuming as in [10] an intermediate m-sample still to be observed, the probability to sample at observation $n+m+1$ a species represented l times among *new* species can be expressed as

$$P^{n+m+1}_{new,l}(\alpha,V)=\frac{V_{n+m+1,j+K^{(n)}_m}}{V_{n+m,j+K^{(n)}_m}}(l-\alpha)W^{(n)}_{l,m}, \tag{27}$$

for $K^{(n)}_m$ and $W^{(n)}_{l,m}$ as previously defined.

In the following Proposition we correct the Bayesian nonparametric estimator, under quadratic loss function, of (27), (see Theorem 2. in [10]), reducing a cumbersome proof to few straightforward steps.

Proposition 8. *Under a general* (α,V) *Gibbs partition model, for* $W^{(n)}_{l,m}=\sum_{i=1}^{K^{(n)}_m}1\{S_i=l|K_n=j\}$, *the Bayesian nonparametric estimator of* $P^{m+n+1}_{new,l}(\alpha,V)$

is given by

$$\mathbb{E}^{(\alpha,V)}_{(S_1,\dots,S_{K_m^{(n)}},K_m^{(n)}|K_n=j)}\left(\frac{V_{n+m+1,j+K_m^{(n)}}}{V_{n+m,j+K_m^{(n)}}}(l-\alpha)W_{l,m}^{(n)}\right)$$
$$=(l-\alpha)\sum_{k-1=0}^{m-l}\frac{V_{n+m+1,j+k}}{V_{n,j}}\binom{m}{l}(1-\alpha)_{l-1}S^{-1,-\alpha,-(n-j\alpha)}_{m-l,k-1}. \tag{28}$$

Proof. Let $f(K_m^{(n)})=\frac{V_{n+m+1,j+K_m^{(n)}}}{V_{n+m,j+K_m^{(n)}}}$ then, by definition of $W_{l,m}^{(n)}$,

$$\begin{aligned}&\mathbb{E}^{(\alpha,V)}_{(S_1,\dots,S_{K_m^{(n)}},K_m^{(n)}|K_n=j)}\left(\frac{V_{n+m+1,j+K_m^{(n)}}}{V_{n+m,j+K_m^{(n)}}}(l-\alpha)W_{l,m}^{(n)}\right)\\&\quad=(l-\alpha)\sum_{k=1}^{m-l+1}\mathbb{E}^{(\alpha,V)}_{(S_1,\dots,S_{K_m^{(n)}}|K_m^{(n)}=k,K_n=j)}\left(f(k)\sum_{i=1}^{k}1\{S_i=l|K_n=j\}\right)\\&\qquad\times\mathbb{P}_{\alpha,V}(K_m^{(n)}=k|K_n=j)\\&\quad=(l-\alpha)\sum_{k-1=0}^{m-l}f(k)k\mathbb{P}(S_1=l|K_m^{(n)}=k,K_n=j)\mathbb{P}_{\alpha,V}(K_m^{(n)}=k|K_n=j).\end{aligned} \tag{29}$$

Specializing (18) for $s_i=l$ for every i, and inserting the marginal for $r=1$ in (29), the result follows. □

Remark 5. The mistake in Theorem 2 in [10] is in the summation limits. The marginalization over the possible numbers of new species arising in the additional m-sample, giving rise to at least one species represented l times, actually ranges between 1, for $m=l$, and $m-l+1$, which stands for one species of size l and $m-l$ species of size 1. In [10] the summation ranges until $m-l$ thus producing the wrong result.

For completeness, by an analogous approach, we provide a simplified derivation for the Bayesian nonparametric estimator of the probability to observe at step $n+m+1$ a species represented l times among the *old* species, namely

$$P_{old,l}^{m+n+1}(\alpha,V)=\frac{V_{n+m+1,j+K_m^{(n)}}}{V_{n+m,j+K_m^{(n)}}}(l-\alpha)O_{l,m}^{(n)}.$$

Proposition 9. *Under a general (α,V) Gibbs partition model, for $O_{l,m}^{(n)}=\sum_{i=1}^{j}1\{n_i+M_{i,m}=l|n_1,\dots,n_j\}$, then a Bayesian nonparametric estimator under quadratic loss function of $P_{old,l}^{m+n+1}(\alpha,V)$, for $c_\xi=\sum_{i=1}^{j}1\{n_i=\xi\}$, is given by*

$$\mathbb{E}^{(\alpha,V)}_{(M_{1,m},\dots,M_{j,m},K_m^{(n)}|n_1,\dots,n_j)}\left(\frac{V_{n+m+1,j+K_m^{(n)}}}{V_{n+m,j+K_m^{(n)}}}(l-\alpha)O_{l,m}^{(n)}\right)$$

$$= (l-\alpha)\sum_{\xi=1}^{l} c_\xi \binom{m}{l-\xi} (\xi-\alpha)_{l-\xi} \sum_{k=0}^{m-l+\xi} \frac{V_{n+m+1,j+k}}{V_{n,j}} S^{-1,-\alpha,-(n-j\alpha+\xi-\alpha)}_{m-l+\xi,k}. \tag{30}$$

Proof. Let $f(K_m^{(n)}) = \frac{V_{n+m+1,j+K_m^{(n)}}}{V_{n+m,j+K_m^{(n)}}}$, then by definition of $O_{l,m}^{(n)}$

$$\begin{aligned}
&\mathbb{E}^{(\alpha,V)}_{(M_{1,m},\ldots,M_{j,m},K_m^{(n)}|n_1,\ldots,n_j)} \left(\frac{V_{n+m+1,j+K_m^{(n)}}}{V_{n,m,j+K_m^{(n)}}} (l-\alpha) O_{l,m}^{(n)} \right) \\
&\quad = (l-\alpha)\sum_{k=0}^{m} f(k) \\
&\qquad \times \mathbb{E}^{(\alpha,V)}_{(M_{1,m},\ldots,M_{j,m}|K_m^{(n)}=k,n_1,\ldots,n_j)} \left(\sum_{i=1}^{j} 1\left\{ n_i + M_{i,m} = l | K_m^{(n)} = k, n_1, \ldots, n_j \right\} \right) \\
&\qquad \times \mathbb{P}_{\alpha,V}(K_m^{(n)} = k | K_n = j) \\
&\quad = (l-\alpha)\sum_{k=0}^{m} f(k) \sum_{i:n_i \le l} \mathbb{P}_{\alpha,V}(M_{i,m} = l - n_i, K_m^{(n)} = k | n_1, \ldots, n_j)
\end{aligned}$$

and the result follows by an application of (24) by the change of variable $c_\xi = \sum_{i=1}^{j} 1\{n_i = \xi\}$. □

Appendix A

This Appendix contains some basic facts on rising and falling factorial numbers, partitions and compositions of the natural integers, together with known results and definitions of generalized *central* and *non central* Stirling numbers that are exploited in the proofs and derivations all over the paper. The main reference is [28]. Additionally, to facilitate the reading of the results contained in [22, 23, 9] and [10], the relationship between central and non central generalized *factorial* coefficients and generalized *Stirling* numbers is reported.

A.1. Generalized rising factorials

For $n = 0, 1, 2, \ldots$, and arbitrary real x and h, $(x)_{n\uparrow h}$ denotes the nth factorial power of x with increment h (also called generalized *rising* factorial) $(x)_{n\uparrow h} := x(x+h)\cdots(x+(n-1)h) = \prod_{i=0}^{n-1}(x+ih) = h^n(x/h)_n$, where $(x)_n$ stands for $(x)_{n\uparrow 1}$, and $(x)_{n\uparrow 0} = x^n$, for which the following multiplicative law holds $(x)_{n+r\uparrow h} = (x)_{n\uparrow h}(x+nh)_{r\uparrow h}$. From e.g. [24] (see Eq. 2.41 and 2.45) a generalized version of the multinomial theorem also holds,

$$(\sum_{j=1}^{p} z_j)_{n\uparrow h} = \sum_{n_j \ge 0, \sum n_j = n} \frac{n!}{n_1! \cdots n_p!} \prod_{j=1}^{p} (z_j)_{n_j \uparrow h}. \tag{31}$$

For $m_j > 0$, for every j, and $\sum_j m_j = m$, an application of the multiplicative law yields $(z_j)_{n_j+m_j-1} = (z_j)_{m_j-1}(z_j + m_j - 1)_{n_j}$ and by (31)

$$\sum_{n_j \geq 0, \sum n_j = n} \frac{n!}{n_1! \cdots n_p!} \prod_{j=1}^{p} (z_j)_{n_j+m_j-1} = \prod_{j=1}^{p} (z_j)_{m_j-1} (\sum_{j=1}^{p} (z_j + m_j - 1))_n$$
$$= \prod_{j=1}^{p} (z_j)_{m_j-1} (m + \sum_{j=1}^{p} z_j - p)_n,$$

which simplifies the proof of Lemma 1 in [23].

A.2. Partitions and compositions

A *partition* of the finite set $[n] := \{1, \ldots, n\}$ into k blocks is an *unordered* collection of non-empty disjoint sets $\{A_1, \ldots, A_k\}$ whose union is $[n]$, where the blocks A_i are assumed to be listed in order of appearance, i.e. in the order of their least elements. The sequence $(|A_1|, \ldots, |A_k|)$ of the sizes of blocks, $(n_1, \ldots, n_k)$, defines a *composition* of n, i.e. a sequence of positive integers with sum n. Two sequences that differ in the order of their terms define *different compositions* of n but the *same partition of* n. Let $\mathcal{P}^k_{[n]}$ denotes the space of all partitions of $[n]$ with k blocks. From [28] (see Eq. (1.9)) the number of ways to partition $[n]$ into k blocks and assign each block a W combinatorial structure such that the number of W-structures on a set of j elements is w_j, in terms of sum over *compositions* of n into k parts, is given by

$$B_{n,k}(w_\bullet) = \frac{n!}{k!} \sum_{(n_1,\ldots,n_k)} \prod_{i=1}^{k} \frac{w_{n_i}}{n_i!}, \tag{32}$$

where $B_{n,k}(w_\bullet)$ is a polynomial in variables $w_1, \ldots, w_{n-k+1}$ known as the (n,k)th *partial Bell polynomial*.

A.3. Generalized Stirling numbers and factorial coefficients

(For a comprehensive treatment see [15], see also [28] Ex. 1.2.7). For arbitrary distinct reals η and β, these are the connection coefficients $S^{\eta,\beta}_{n,k}$ defined by

$$(x)_{n\downarrow\eta} = \sum_{k=0}^{n} S^{\eta,\beta}_{n,k} (x)_{k\downarrow\beta} \tag{33}$$

and correspond to $S^{\eta,\beta}_{n,k} = B_{n,k}((\beta - \eta)_{\bullet-1\downarrow\eta})$, where $(x)_{n\downarrow h} = (x)_{n\uparrow-h}$ are generalized *falling* factorials and $(x)_{[n]} = (x)_{n\downarrow 1}$. Hence for $\eta = -1$, $\beta = -\alpha$, and $\alpha \in (-\infty, 1)$, $S^{-1,-\alpha}_{n,k}$ is defined by

$$(x)_n = \sum_{k=0}^{n} S^{-1,-\alpha}_{n,k} (x)_{k\uparrow\alpha}, \tag{34}$$

and for $w_{n_i} = (1-\alpha)_{n_i-1}$ and $\alpha \in [0,1)$, $B_{n,k}((1-\alpha)_{\bullet-1}) = S^{-1,-\alpha}_{n,k}$. In [22, 23, 9, 10] the treatment is in term of *generalized factorial coefficients*, which are the connection coefficients $\mathcal{C}^{\alpha}_{n,k}$ defined by $(\alpha y)_n = \sum_{k=0}^{n} \mathcal{C}^{\alpha}_{n,k}(y)_k$, (see e.g. [4]). From the definition of generalized rising factorials and (34), if $x = y\alpha$ then $(y\alpha)_n = \sum_{k=0}^{n} S^{-1,-\alpha}_{n,k}\alpha^k(y)_k$, hence $S^{-1,-\alpha}_{n,k} = \mathcal{C}^{\alpha}_{n,k}\alpha^{-k}$. Additionally, specializing formula (16) in [15], the following convolution relation holds, which defines *non-central* generalized Stirling numbers

$$S^{-1,-\alpha,\gamma}_{n,k} = \sum_{s=k}^{n} \binom{n}{s} S^{-1,-\alpha}_{s,k}(-\gamma)_{n-s}, \tag{35}$$

and consequently,

$$\mathcal{C}^{\alpha,\gamma}_{n,k} = \alpha^k S^{-1,-\alpha,\gamma}_{n,k} = \sum_{s=k}^{n} \binom{n}{s} \mathcal{C}^{\alpha}_{s,k}(-\gamma)_{n-s}. \tag{36}$$

An easy variation of equation (38) in [23] (see also (2.49) in [4]) provides a definition of *non-central* generalized Stirling numbers as connection coefficients.

Acknowledgements

I wish to thank an Associate Editor and a referee for their careful reading and their precious suggestions on how to improve the presentation of the paper, and Warren J. Ewens for his kind support and encouragement.

References

[1] Arratia, R., Barbour, A. D., Tavaré, S. (2003) *Logarithmic combinatorial structures: a probabilistic approach.* EMS Monographs in Mathematics. MR2032426

[2] Cerquetti, A. (2011a) A decomposition approach to Bayesian nonparametric estimation under two-parameter Poisson-Dirichlet priors. *Proceedings of ASMDA 2011 - Rome, Italy.* Available at: http://geostasto.eco.uniroma1.it/utenti/cerquetti/asmda2011last.pdf

[3] Cerquetti, A. (2011b) Conditional α-diversity for exchangeable Gibbs partition driven by the stable subordinator. *Proceedings of S.Co. Conferemce, 2011, Padova, Italy.* Available at: http://homes.stat.unipd.it/mgri/SCo2011/Papers/CS/CS-7/cerquetti.pdf.

[4] Charalambides, C. A. (2005) *Combinatorial Methods in Discrete Distributions.* Wiley, Hoboken NJ. MR2131068

[5] De Moivre, A. (1718) *The doctrine of chances: Or a method of calculating the probabilities of events in play.* London. Pearson.

[6] Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol., 3, 87–112.* MR0325177

[7] Ewens, W. and Tavaré, S. (1995) The Ewens sampling formula. In Multivariate discrete distributions (Johnson, N.S., Kotz, S. and Balakrishnan, N. eds.). Wiley, NY.
[8] Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. (2009) Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. B*, **71**, 993–1008. MR2750254
[9] Favaro, S., Lijoi, A. and Prünster, I. (2012a) Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.* (to appear)
[10] Favaro, S., Lijoi, A. and Prünster, I. (2012b) A new estimator of the discovery probability. *Biometrics*, **68**, 1188–1196.
[11] Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230. MR0350949
[12] Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943) 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.
[13] Gnedin, A. (2010) A species sampling model with finitely many types. *Electron. Commun. Prob.*, **15**, 79–88. MR2606505
[14] Gnedin, A. and Pitman, J. (2006) Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, **138**, 3, 5674–5685. MR2160320
[15] Hsu, L. C, and Shiue, P. J. (1998) A unified approach to generalized Stirling numbers. *Adv. Appl. Math.*, **20**, 366-384. MR1618435
[16] Iyer, P. V. K. (1949) Calculation of factorial moments of certain probability distributions. *Nature*. **164**, 282.
[17] Iyer, P. V. K. (1958) A theorem on factorial moments and its applications. *Ann. Math. Statist.*, **29**, 254–261. MR0093841
[18] Johnson, N. S. and Kotz, S. (2005) *Univariate discrete distributions* 3rd Ed. Wiley, NY. MR2163227
[19] Jordan, M. C. (1867) De quelques formules de probabilité. *Comptes Rendus. Académie des Sciences, Paris*, **65**, 993–994.
[20] Kingman, J. F. C. (1975) Random discrete distributions. *J. Roy. Statist. Soc. B*, **37**, 1–22. MR0368264
[21] Kingman, J. F. C (1978) The representation of partition structure. *J. London Math. Soc.* **2**, 374–380. MR0509954
[22] Lijoi, A., Mena, R. and Prünster, I. (2007) Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, **94**, 769–786. MR2416792
[23] Lijoi, A., Prünster, I. and Walker, S. G. (2008) Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.*, **18**, 1519–1547. MR2434179
[24] Normand, J. M. (2004) Calculation of some determinants using the *s*-shifted factorial. *J. Phys. A: Math. Gen.* **37**, 5737-5762. MR2066627
[25] Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, **102**, 145-158. MR1337249
[26] Pitman, J. (1996) Some developments of the Blackwell-MacQueen urn scheme. In T.S. Ferguson, Shapley L.S., and MacQueen J.B., editors, *Statis-*

tics, Probability and Game Theory, vol. 30 of *IMS Lecture Notes-Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward, CA. MR1481784

[27] Pitman, J. (2003) 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. IMS, Hayward, California. MR2004330

[28] Pitman, J. (2006) *Combinatorial Stochastic Processes*. Ecole d'Eté de Probabilité de Saint-Flour XXXII - 2002. Lecture Notes in Mathematics N. 1875, Springer. MR2245368

[29] Pitman, J. and Yor, M. (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, **25**, 855–900. MR1434129

[30] Yamato, H. and Sibuya, M. (2000) Moments of some statistics of Pitman sampling formula. *Bull. Inform. Cybernet.*, 32, 1. MR1792352