CONDITIONAL FORMULAE FOR GIBBS–TYPE EXCHANGEABLE RANDOM PARTITIONS

By Stefano Favaro[*], Antonio Lijoi[*] and Igor Prünster[*]

University of Torino, University of Pavia and University of Torino

Gibbs–type random probability measures and the exchangeable random partitions they induce represent an important framework both from a theoretical and applied point of view. In the present paper, motivated by species sampling problems, we investigate some properties concerning the conditional distribution of the number of blocks with a certain frequency generated by Gibbs–type random partitions. The general results are then specialized to three noteworthy examples yielding completely explicit expressions of their distributions, moments and asymptotic behaviours. Such expressions can be interpreted as Bayesian nonparametric estimators of the rare species variety and their performance is tested on some real genomic data.

1. Introduction. Let $\mathbb{X}$ be a complete and separable metric space equipped with the Borel σ–algebra $\mathscr{X}$ and denote by $\mathscr{P}$ the space of probability distributions defined on $(\mathbb{X}, \mathscr{X})$ with $\sigma(\mathscr{P})$ denoting the Borel σ–algebra of subsets of $\mathscr{P}$. By virtue of de Finetti's representation theorem, a sequence of $\mathbb{X}$–valued random elements $(X_n)_{n \geq 1}$, defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, is exchangeable if and only if there exists a probability measure Q on the space of probability distributions $(\mathscr{P}, \sigma(\mathscr{P}))$ such that

$$(1) \qquad \mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n] = \int_{\mathscr{P}} \prod_{i=1}^{n} P(A_i) \, Q(\mathrm{d}P)$$

for any $A_1, \ldots, A_n$ in $\mathscr{X}$ and $n \geq 1$. The probability measure Q directing the exchangeable sequence $(X_n)_{n \geq 1}$ is also termed *de Finetti measure* and takes on the interpretation of prior distribution in Bayesian applications. The representation theorem can be equivalently stated by saying that, given an exchangeable sequence $(X_n)_{n \geq 1}$, there exists a random probability measure

(r.p.m.) $\tilde{P}$, defined on $(\mathbb{X}, \mathscr{X})$ and taking values in $(\mathscr{P}, \sigma(\mathscr{P}))$, such that

$$(2) \qquad \mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n | \tilde{P}] = \prod_{i=1}^{n} \tilde{P}(A_i)$$

almost surely, for any $A_1, \ldots, A_n$ in $\mathscr{X}$ and $n \geq 1$. In this paper we will focus attention on almost surely discrete r.p.m.s, i.e., $\tilde{P}$ is such that $\mathbb{P}[\tilde{P} \in \mathscr{P}_d] = 1$ with $\mathscr{P}_d$ indicating the set of discrete probability measures on $(\mathbb{X}, \mathscr{X})$ or, equivalently, $(X_n)_{n \geq 1}$ is directed by a de Finetti measure Q that is concentrated on $\mathscr{P}_d$. An almost surely discrete r.p.m. (without fixed atoms) can always be written as

$$(3) \qquad \tilde{P} = \sum_{i \geq 1} \tilde{p}_i \delta_{\hat{X}_i}$$

for some sequences $(\hat{X}_i)_{i \geq 1}$ and $(\tilde{p}_i)_{i \geq 1}$ of, respectively, $\mathbb{X}$–valued random locations and non–negative random weights such that $\mathbb{P}[\sum_{i \geq 1} \tilde{p}_i = 1] = 1$ almost surely.

In the following we will assume that the two sequences in (3) are independent. These specifications imply that a sample $(X_1, \ldots, X_n)$ from the exchangeable sequence generates a random partition Π_n of the set of integers $\mathbb{N}_n := \{1, \ldots, n\}$, in the sense that any $i \neq j$ belongs to the same partition set if and only if $X_i = X_j$. The random number of partition sets in Π_n is denoted as K_n with respective frequencies $N_1, \ldots, N_{K_n}$. Accordingly, the sequence $(X_n)_{n \geq 1}$ associated to a r.p.m. $\tilde{P}$ as in (3) induces an exchangeable random partition $\Pi = (\Pi_n)_{n \geq 1}$ of the set of natural numbers $\mathbb{N}$. The distribution of Π is characterized by the sequence of distributions $\{p_k^{(n)} : 1 \leq k \leq n, \, n \geq 1\}$ such that

$$(4) \qquad p_k^{(n)}(\boldsymbol{n}) = \mathbb{P}[K_n = k, \, \boldsymbol{N} = \boldsymbol{n}],$$

with $\boldsymbol{N} = (N_1, \ldots, N_{K_n})$ and $\boldsymbol{n} = (n_1, \ldots, n_k)$. Hence, (4) identifies, for any $n \geq 1$, the probability distribution of the random partition Π_n of $\mathbb{N}_n$ and is known as *exchangeable partition probability function* (EPPF), a concept introduced by J. Pitman [21] as a major development of earlier results on exchangeable random partitions due to J.F.C. Kingman (see, e.g., [15, 16]). It is worth noting that EPPFs can be defined either by starting from an exchangeable sequence associated to a discrete r.p.m. and looking at the induced partitions or by defining directly the partition distribution. In the latter case, the distribution of the random partitions Π_n must satisfy certain consistency conditions and a symmetry property that guarantees exchangeability. A comprehensive account on exchangeable random partitions can be

found in [23] together with an overview of the numerous application areas and relevant references.

1.1. *Gibbs–type r.p.m.s and partitions.* We now recall the definition of a general class of r.p.m.s and of the exchangeable random partitions they induce together with some of distinguished special cases. This important class, introduced and thoroughly studied in [10], is characterized by the fact that its members induce exchangeable random partitions admitting EPPFs with product form, a feature which is crucial for guaranteeing mathematical tractability. Before introducing the definition, set $\mathcal{D}_{n,j} := \{(n_1, \ldots, n_j) \in \{1, \ldots, n\}^j : \sum_{i=1}^{j} n_i = n\}$ and denote by $(a)_q = \Gamma(a+q)/\Gamma(a)$ the q–th ascending factorial of a.

DEFINITION 1. *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to an almost surely discrete r.p.m.* (3) *for which locations $(\hat{X}_i)_{i \geq 1}$ and weights $(\tilde{p}_i)_{i \geq 1}$ are independent. Then the r.p.m. $\tilde{P}$ and the induced exchangeable random partition are said of* Gibbs–type *if, for any $n \geq 1$, $1 \leq j \leq n$ and $(n_1, \ldots, n_j) \in \mathcal{D}_{n,j}$ the corresponding EPPF can be represented as follows*

$$(5) \qquad p_j^{(n)}(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1-\sigma)_{n_i-1},$$

for $\sigma \in (-\infty, 1)$ and a set of non-negative weights $\{V_{n,j} : n \geq 1, 1 \leq j \leq n\}$ satisfying the recursion $V_{n,j} = V_{n+1,j+1} + (n - \sigma j)V_{n+1,j}$ with $V_{1,1} = 1$.

Hence, a Gibbs–type random partition is completely specified by the choice of $\sigma < 1$ and the weights $V_{n,j}$'s. The role of σ is crucial since it determines the clustering structure as well as the asymptotic behaviour of Gibbs–type models. As for the latter aspect, for any $n \geq 1$ define

$$c_n(\sigma) := \mathbb{1}_{(-\infty,0)}(\sigma) + \log(n)\,\mathbb{1}_{\{0\}}(\sigma) + n^\sigma\,\mathbb{1}_{(0,1)}(\sigma).$$

Then, for any Gibbs–type r.p.m. there exists a strictly positive and almost surely finite random variable S_σ, usually termed σ–*diversity*, such that

$$(6) \qquad \frac{K_n}{c_n(\sigma)} \xrightarrow{\text{a.s.}} S_\sigma,$$

for $n \to +\infty$. See [22, Section 6.1] for details. Finally, it is worth recalling that the solutions of the backward recursions defining the $V_{n,j}$'s form a convex set whose extreme points are determined in [10, Theorem 12] providing a complete characterization of Gibbs–type models according to the values of σ they assume. In the next subsection we concisely point out three important explicit special cases to be dealt with also in the sequel.

1.2. *Examples.* We will illustrate three noteworthy examples of Gibbs–type r.p.m.s that correspond to different choices of σ and the $V_{n,j}$'s in Definition 1. The first one is the *Dirichlet process* [9], which corresponds to a Gibbs–type r.p.m. characterized by $\sigma = 0$ and $V_{n,j} = \theta^j/(\theta)_n$ with $\theta > 0$. The implied EPPF coincides with

$$(7) \qquad p_j^{(n)}(n_1, \ldots, n_j) = \frac{\theta^j}{(\theta)_n} \prod_{i=1}^{j}(n_i - 1)!$$

and is well–known in Population Genetics as the *Ewens model*. See [6] and references therein.

The most interesting special case for our purposes is a generalization of (7) that has been provided by J. Pitman in [21]. It corresponds to the exchangeable random partition generated by the *two–parameter Poisson–Dirichlet process*, which coincides with a Gibbs–type r.p.m. with $\sigma \in (0,1)$ and, for any $\theta > -\sigma$, $V_{n,j} = \prod_{i=0}^{j-1}(\theta + i\sigma)/(\theta)_n$. The EPPF turns out to be

$$(8) \qquad p_j^{(n)}(n_1, \ldots, n_j) = \frac{\prod_{i=0}^{j-1}(\theta + i\sigma)}{(\theta)_n} \prod_{i=1}^{j}(1 - \sigma)_{n_i - 1}.$$

Clearly, the Ewens model (7) is recovered from (8) by letting $\sigma \to 0$. The r.p.m. and the partition distribution associated to (8) will be equivalently termed $\mathrm{PD}(\sigma, \theta)$ process or *Pitman model*.

Finally, another notable example of Gibbs–type r.p.m. has been recently provided in [11]. It is characterized by $\sigma = -1$ and weights of the form

$$(9) \qquad V_{n,j} = (\gamma)_{n-j} \frac{\prod_{i=1}^{j-1}(i^2 - \gamma i + \zeta)}{\prod_{i=1}^{n-1}(i^2 + \gamma i + \zeta)},$$

where ζ and γ are chosen such that $\gamma \geq 0$ and $i^2 - \gamma i + \zeta > 0$ for all $i \geq 1$. In the sequel we will term both the r.p.m. and the induced exchangeable random partition as *Gnedin model*.

1.3. *Aims and outline of the paper.* The main applied motivation of the present study is related to species sampling problems. Indeed, in many applications that arise, e.g., in population genetics, ecology and genomics, a population is a composition of individuals (e.g., animals, plants or genes) of different species: the $\hat{X}_i$'s and the $\tilde{p}_i$'s in (3) can then be seen as species labels and species proportions, respectively. In most cases one is interested in the $\tilde{p}_i$'s or in some functionals of them: this naturally leads to work with the random partitions induced by an exchangeable sequence. The number

of distinct partition blocks K_n takes on the interpretation of the number of different species detected in the observed sample $(X_1, \ldots, X_n)$ and the N_j's are the species frequencies. Given the relevance and intuitiveness of such an applied framework, throughout the paper we will often resort to the species metaphor even if the tools we will introduce and the results we will achieve are of interest beyond the species sampling framework.

Our first goal consists in analyzing certain distributional properties of Gibbs–type r.p.m.s. Specifically, we are interested in determining the probability distribution of the number of partition blocks having a certain size or frequency. In other words, given an exchangeable sequence $(X_n)_{n \geq 1}$ as in (1) associated to a Gibbs–type r.p.m., we investigate distributional properties of: (i) the number of species with frequency l in a sample of size n, namely, $M_{l,n} = \sum_{i=1}^{K_n} \mathbb{1}_{\{l\}}(N_i)$; (ii) the number of species $M_{l,n+m} = \sum_{i=1}^{K_{n+m}} \mathbb{1}_{\{l\}}(N_i)$ with frequency l in an enlarged sample of size $n+m$, for $m \geq 0$, conditionally on the species composition detected within a n–size sample $(X_1, \ldots, X_n)$. Note that the latter problem is considerably more challenging since it requires to account for the allocation of $(X_{n+1}, \ldots, X_{n+m})$ between "old" and "new" species together with the sequential modification of their frequencies, conditional on $(X_1, \ldots, X_n)$.

Solving problem (ii) is also the key for achieving our second goal, namely the derivation of estimators for *rare species* variety, where rare species are identified as those with a frequency not greater than a specific abundance threshold τ. This is of great importance in numerous applied settings. For example, in ecology conservation of biodiversity is a fundamental theme and it can be formalized in terms of the number of species whose frequency is greater than a specified threshold. Indeed, any form of management on a sustained basis requires a certain number of sufficiently abundant species (the so–called breeding stock). We shall address the issue be relying on a Bayesian nonparametric approach: the de Finetti measure associated to a Gibbs–type r.p.m. represents the nonparametric prior distribution and relying on the conditional (or posterior) distributions in (ii) one derives the desired estimators as conditional (or posterior) expected values. Bayesian estimators for overall species variety, namely the estimation of the distinct species (regardless of the respective frequencies), have been introduced and discussed in [17, 19, 20, 8]. Further contributions at the interface between Bayesian Nonparametrics and Gibbs–type random partitions can be found in [12, 13, 18]. None of the existing work provides estimators for the number of species with specific abundance. Here we fill in this important gap and, besides providing general results valid for the whole family of Gibbs–type r.p.m.s, we specialize them to the three examples outlined in Subsection 1.2.

This leads to explicit expressions that are of immediate use in applications.

The paper is structured as follows. Section 2 provides distributional results on the unconditional structure of $M_{l,n}$ and the conditional structure of $M_{l,n+m}$, given the species composition detected in a sample of size n, for general Gibbs–type r.p.m.s together with the corresponding estimators. Section 3 focuses on the three special cases of the Dirichlet process, and the models of Gnedin and of Pitman. In particular, for these special cases we also provide asymptotic results concerning the conditional distribution of $M_{l,n+m}$, given the species composition detected in a sample of size n, as the size of the additional sample m increases. The framework for genomic applications, including platforms under which such estimation problems arise, is presented in Section 4, where the methodology is also tested on real genomic data. In Section 5 the proofs of the results of Sections 2 and 3 and some useful techniques are described.

2. Distribution of cluster frequencies.

2.1. *Probability distribution of $M_{l,n}$*. We start our analysis of distributional properties of Gibbs–type random partitions by focusing on the unconditional distribution of the number of blocks with a certain size l, $M_{l,n}$. The blocks with relatively low frequency are typically referred to as *small blocks* (see e.g. [25]), which, in terms species sampling, will represent the rare species.

First note that a simple change of variable in the EPPF (5) yields the probability distribution of $\boldsymbol{M}_n := (M_{1,n}, \ldots, M_{n,n})$. Specifically, the so–called Gibbs–type sampling formula determines the probability distribution of $\boldsymbol{M}_n$ and it corresponds to

$$(10) \qquad \mathbb{P}[\boldsymbol{M}_n = (m_1, \ldots, m_n)] = V_{n,j}\, n! \prod_{i=1}^{n} \left(\frac{(1-\sigma)_{i-1}}{i!} \right)^{m_i} \frac{1}{m_i!},$$

for any $(m_1, \ldots, m_n) \in \{0, 1, \ldots\}^n$ such that $\sum_{i=1}^{n} i m_i = n$ and $\sum_{i=1}^{n} m_i = j$. The next proposition provides explicit expressions for the r–th factorial moments of $M_{l,n}$ in terms of generalized factorial coefficients $\mathscr{C}(n, k; \sigma)$. Recall that, for any $n \geq 1$ and $k \leq n$, $\mathscr{C}(n, k; \sigma)$ is defined as $(\sigma t)_n = \sum_{k=0}^{n} \mathscr{C}(n, k; \sigma)(t)_k$ for $\sigma \in \mathbb{R}$ and, moreover, is computable as $\mathscr{C}(n, k; \sigma) = (1/k!) \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-\sigma j)_n$ with the proviso $\mathscr{C}(0, 0; \sigma) = 1$, $\mathscr{C}(n, 0; \sigma) = 0$ for any $n > 0$ and $\mathscr{C}(n, k; \sigma) = 0$ for any $k > n$. For an exhaustive account on generalized factorial coefficients the reader is referred to [4].

PROPOSITION 1. *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a Gibbs–type r.p.m. Then, for any $l = 1, \ldots, n$ and $r \geq 1$,*

$$(11) \quad \mathbb{E}\left[(M_{l,n})_{[r]}\right] = \left(\frac{(1-\sigma)_{l-1}}{l!}\right)^r (n)_{[lr]} \sum_{j=1}^n V_{n,j} \frac{\mathscr{C}(n-rl, j-r; \sigma)}{\sigma^{j-r}},$$

where $(a)_{[q]} = a(a-1) \cdots (a-q+1)$ for any $q \geq 1$.

By using standard arguments involving probability generating functions, one can use the factorial moments (11) for determining the probability distribution of $M_{l,n}$. This will be illustrated for the three examples in Section 3. The asymptotic behaviour of $M_{l,n}$, as $n \to \infty$, is determined in [23, Lemma 3.11]: if $\tilde{P}$ is a Gibbs–type r.p.m. with $\sigma \in (0,1)$, then for any $l \geq 1$

$$(12) \quad \frac{M_{l,n}}{n^\sigma} \xrightarrow{\text{d}} \frac{\sigma(1-\sigma)_{l-1}}{l!} S_\sigma,$$

as $n \to +\infty$, where S_σ is the σ–diversity defined in (6). Some recent interesting developments on the asymptotic behaviour of the random variable $M_{l,n}$ associated to a generic exchangeable random partition are provided in [25].

2.2. *Conditional formulae.* Unlike the study of unconditional properties of Gibbs–type random partitions, that are the focus of a well–established literature with plenty of results, the investigation of conditional properties for this family of partitions has been only recently started in [20] and many issues are still to be addressed. We are going to focus on determining the distribution of $M_{l,n+m}$ conditional on the number of distinct species K_n, and on their respective frequencies $N_1, \ldots, N_{K_n}$, recorded in the sample $(X_1, \ldots, X_n)$. This will also serve as a tool for predicting the value of the number of distinct species that will appear l times in the enlarged sample $(X_1, \ldots, X_{n+m})$, given the observed sample $(X_1, \ldots, X_n)$.

Let $X_1^*, \ldots, X_{K_n}^*$ denote the labels identifying the K_n distinct species detected in the sample $(X_1, \ldots, X_n)$. One can, then, define

$$L_m^{(n)} := \sum_{i=1}^m \prod_{j=1}^{K_n} \mathbb{1}_{\{X_j^*\}^c}(X_{n+i}) = \text{card}(\{X_{n+1}, \ldots, X_{n+m}\} \cap \{X_1^*, \ldots, X_{K_n}^*\}^c)$$

as the number of observations from the additional sample of size m that do not coincide with any of the K_n distinct species in the basic sample. Correspondingly $X_{K_n+1}^*, \ldots, X_{K_n+K_m^{(n)}}^*$ are the labels identifying the additional

$K_m^{(n)} = K_{n+m} - K_n$ distinct species generated by these $L_m^{(n)}$ observations. Then we can define

$$S_{K_n+i} := \sum_{j=1}^{m} \mathbb{1}_{\{X_{K_n+i}^*\}}(X_{n+j}), \qquad S_q := \sum_{j=1}^{m} \mathbb{1}_{\{X_q^*\}}(X_{n+j})$$

for $i = 1, \ldots, K_m^{(n)}$ and $q = 1, \ldots, K_n$, where one obviously has $\sum_{i=1}^{K_m^{(n)}} S_{K_n+i} = L_m^{(n)}$. For our purposes, it is useful to resort to the decomposition $M_{l,n+m} = O_{l,m} + N_{l,m}$ where

$$(13) \qquad O_{l,m} := \sum_{q=1}^{K_n} \mathbb{1}_{\{l\}}(N_q + S_q) \qquad N_{l,m} := \sum_{i=1}^{K_m^{(n)}} \mathbb{1}_{\{l\}}(S_{K_n+i})$$

for any $l = 1, \ldots, n+m$. It is apparent that $O_{l,m} = 0$ for any $l > n+m$ and $N_{l,m} = 0$ for any $l > m$. Hence, $O_{l,m}$ is the number of distinct species, among the K_n detected in the basic sample $(X_1, \ldots, X_n)$, that have frequency l in the enlarged sample of size $n+m$. Analogously $N_{l,m}$ is the number of additional distinct species, generated by $L_m^{(n)}$ observations in $(X_{n+1}, \ldots, X_{n+m})$, with frequency l in the enlarged sample. For notational convenience we introduce random variables $O_{l,m}^{(n)}$ and $N_{l,m}^{(n)}$ that are defined in distribution as follows

$$\mathbb{P}[O_{l,m}^{(n)} = x] = \mathbb{P}[O_{l,m} = x \,|\, K_n = j, \, \boldsymbol{N} = \boldsymbol{n}]$$
$$\mathbb{P}[N_{l,m}^{(n)} = y] = \mathbb{P}[N_{l,m} = y \,|\, K_n = j, \, \boldsymbol{N} = \boldsymbol{n}]$$

for any $1 \leq j \leq n$, $\boldsymbol{n} \in \mathcal{D}_{n,j}$ and $n, m \geq 1$. Moreover, we set $\mathcal{C}_{j,r}$ as the space of all vectors $\mathbf{c}^{(r)} = (c_1, \ldots, c_r) \in \{1, \ldots, j\}^r$ such that $c_i \neq c_\ell$ for any $i \neq \ell$ and $\max_{1 \leq i \leq j} n_{c_i} \leq l$. Finally

$$I_\sigma \left(l, m, r, \boldsymbol{n}, \mathbf{c}^{(r)} \right)$$

$$:= r! \binom{m}{l - n_{c_1}, \ldots, l - n_{c_r}, m - l + |\boldsymbol{n}_{\mathbf{c}^{(r)}}|} \prod_{i=1}^{r} (n_{c_i} - \sigma)_{l - n_{c_i}},$$

where $|\boldsymbol{n}_{\mathbf{c}^{(r)}}| := \sum_{i=1}^{r} n_{c_i}$. The next result provides an explicit expression for the r–th factorial moments of $O_{l,m}^{(n)}$ in terms of noncentral generalized factorial coefficients defined by $\mathscr{C}(n, k; \gamma, \sigma) := (\sigma t - \gamma)_n = \sum_{k=0}^{n} \mathscr{C}(n, k; \sigma, \gamma)(t)_k$ with $\sigma, \gamma \in \mathbb{R}$. Recall also that $\mathscr{C}(n, k; \sigma, \gamma) = (1/k!) \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-\sigma j - \gamma)_n$ with the proviso $\mathscr{C}(0, 0; \sigma, \gamma) = 1$, $\mathscr{C}(n, 0; \sigma) = (-\gamma)_n$ for any $n > 0$ and $\mathscr{C}(n, k; \sigma, \gamma) = 0$ for any $k > n$.

THEOREM 1.   *Let $(X_n)_{n\geq 1}$ be an exchangeable sequence associated to a Gibbs–type r.p.m. Then, for any $l = 1, \ldots, n + m$, $r \geq 1$ and $\boldsymbol{n} \in \mathcal{D}_{n,j}$*

$$(14) \quad \mathbb{E}\left[(O_{l,m}^{(n)})_{[r]}\right] = \sum_{\boldsymbol{c}^{(r)} \in \mathcal{C}_{j,r}} I_\sigma\left(l, m, r, \boldsymbol{n}, \boldsymbol{c}^{(r)}\right)$$

$$\times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}\left(m - rl + |\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|, k; \sigma, -n + |\boldsymbol{n}_{\boldsymbol{c}^{(r)}}| + (j - r)\sigma\right)}{\sigma^k}.$$

It is worth observing that the moments in (14), for any $r \geq 1$, characterize the distribution of $O_{l,m}^{(n)}$. Such a distribution is interpretable as the posterior probability distribution, given the observations $(X_1, \ldots, X_n)$, of the number of distinct species that (i) appear with frequency l in a sample of size $n + m$; (ii) had been already detected within $(X_1, \ldots, X_n)$. Therefore we will refer to $O_{l,m}^{(n)}$ as the number of "old" species with frequency l. The Bayesian nonparametric estimator, under a quadratic loss function, coincides with the expected value of $O_{l,m}^{(n)}$ and is easily recovered from (14).

COROLLARY 1.   *Let $(X_n)_{n\geq 1}$ be an exchangeable sequence associated to a Gibbs–type r.p.m. Conditionally on a sample $(X_1, \ldots, X_n)$, the expected number of "old" distinct species that appear with frequency l, for any $l = 1, \ldots, n + m$, in a sample of size $n + m$ is given by*

$$(15) \quad \hat{O}_{l,m}^{(n)} := \mathbb{E}[O_{l,m}^{(n)}] = \sum_{t=1}^{l} \binom{m}{l-t} m_t (t - \sigma)_{l-t}$$

$$\times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}\left(m - (l - t), k; \sigma, -n + t + (j - r)\sigma\right)}{\sigma^k},$$

with $m_t \geq 0$ being the number of distinct species with frequency t observed in the basic sample, namely $m_t = \sum_{i=1}^{K_n} \mathbb{1}_{\{t\}}(N_i)$. Moreover, $(K_n, M_{1,n}, \ldots, M_{l,n})$ is sufficient for predicting $O_{l,m}^{(n)}$ over the whole sample of size $n + m$.

An analogous result of Theorem 1 can be established for $N_{l,m}^{(n)}$. Indeed, if we set

$$J_\sigma(l, m, r) := \binom{m}{l, \ldots, l, m - rl} [(1 - \sigma)_{l-1}]^r$$

one can show the following theorem.

THEOREM 2.   *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a Gibbs–type r.p.m. Then, for any $l = 1, \ldots, m$ and $r \geq 1$*

$$(16) \quad \mathbb{E}\left[(N_{l,m}^{(n)})_{[r]}\right] = J_\sigma(l,m,r) \sum_{k=0}^{m-rl} \frac{V_{n+m,j+k+r}}{V_{n,j}} \frac{\mathscr{C}(m-rl,k;\sigma,-n+j\sigma)}{\sigma^k}.$$

Hence, (16) characterizes the probability distribution of $N_{l,m}^{(n)}$. This can be seen as the posterior probability distribution, conditional on the observations $(X_1, \ldots, X_n)$, of the number of distinct species that (i) appear with frequency l in a sample of size $n+m$; (ii) do not coincide with any of the K_n distinct species already detected within $(X_1, \ldots, X_n)$. For this reason $N_{l,m}^{(n)}$ is referred to as the number of "new" species with frequency l. Thus, the Bayesian nonparametric estimator, under a quadratic loss function, is easily recovered from (16).

COROLLARY 2.   *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a Gibbs–type r.p.m. Conditionally on a sample $(X_1, \ldots, X_n)$, the expected number of "new" distinct species that appear with frequency l, for any $l = 1, \ldots, m$, in a sample of size $n + m$ is given by*
(17)

$$\hat{N}_{l,m}^{(n)} := \mathbb{E}[N_{l,m}^{(n)}] = \binom{m}{l}(1-\sigma)_{l-1} \sum_{k=0}^{l} \frac{V_{n+m,j+k+1}}{V_{n,j}} \frac{\mathscr{C}(m-l,k;\sigma,-n+j\sigma)}{\sigma^k}.$$

Hence, K_n is sufficient for predicting $N_{l,m}^{(n)}$.

REMARK 1.   According to the definition of the random variable $N_{l,m}^{(n)}$, one has

$$(18) \qquad \hat{E}_m^{(n)} := \mathbb{E}[K_m^{(n)} \mid K_n = j, \boldsymbol{N} = \boldsymbol{n}] = \sum_{l=1}^{m} \hat{N}_{l,m}^{(n)}$$

providing an alternative derivation of the Bayesian nonparametric estimator for the number of "new" distinct species derived in [20]. A detailed discussion of the estimator (18) and its relevance in genomics can be found in [19].

At this point we turn our attention to characterizing the following random variable

$$(19) \qquad\qquad\qquad M_{l,m}^{(n)} \stackrel{\mathrm{d}}{=} O_{l,m}^{(n)} + N_{l,m}^{(n)}$$

whose probability distribution coincides with the distribution of the number $M_{l,n+m}$ of clusters of size l featured by $(X_1, \ldots, X_{n+m})$ conditional on the basic sample $(X_1, \ldots, X_n)$. In particular, if we set

$$
H_\sigma(l, m, r, t, \boldsymbol{n}, \boldsymbol{c}^{(t)})
$$
$$
:= t! \binom{m}{l, \cdots, l, l - n_{c_1}, \ldots, l - n_{c_t}, m - rl + |\boldsymbol{n}_{\boldsymbol{c}^{(t)}}|}
$$
$$
\times [(1 - \sigma)_{l-1}]^{r-t} \prod_{i=1}^{t} (n_{c_i} - \sigma)_{l - n_{c_i}}
$$

an analogous result ot Theorem 1 and Theorem 2 can be established for $M_{l,m}^{(n)}$.

THEOREM 3.   *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a Gibbs–type r.p.m. Then, for any $l = 1, \ldots, m + n$ and $r \geq 1$*

(20)

$$
\mathbb{E}\left[\left(M_{l,m}^{(n)}\right)_{[r]}\right] = \sum_{t=0}^{r} \binom{r}{t} \sum_{\boldsymbol{c}^{(t)} \in \mathcal{C}_{j,t}} H_\sigma(l, m, r, t, \boldsymbol{n}, \boldsymbol{c}^{(t)})
$$
$$
\times \sum_{k=0}^{m - rl + |\boldsymbol{n}_{\boldsymbol{c}^{(t)}}|} \frac{V_{n+m, j+k+r-t}}{V_{n,j}} \frac{\mathscr{C}(m - rl + |\boldsymbol{n}_{\boldsymbol{c}^{(t)}}|, k; \sigma, -n + |\boldsymbol{n}_{\boldsymbol{c}^{(t)}}| + (j - t)\sigma)}{\sigma^k}.
$$

Hence (20) characterizes the probability distribution of $M_{l,m}^{(n)}$. Such a probability distribution is interpreted as the posterior probability distribution, given the observation $(X_1, \ldots, X_n)$, of the number of distinct species that appear with frequency l in a sample of size $n + m$. Thus, the Bayesian non-parametric estimator, under a quadratic loss function, is easily recovered from (20). Clearly, according to (19), this also corresponds to the sum of the estimators in (15) and (17).

3. Illustrations. We now apply the general results of Section 2 and specialize them to some noteworthy examples of Gibbs–type models. We will devote particular attention to the two–parameter Poisson–Dirichlet process since it is particularly suited for species sampling applications in general [17] and for genomic applications in particular, as will be seen in Section 4.

3.1. *The Dirichlet process.* Denote the signless Stirling number of the first kind by $|s(n, k)|$ and recall that $\lim_{\sigma \to 0} \sigma^{-k} \mathscr{C}(n, k, \sigma) = |s(n, k)|$ for

any $n \geq 1$ and $1 \leq k \leq n$. Now, let $\tilde{P}$ be a Dirichlet process with parameter θ and considering the form of the $V_{n,j}$ weights and Theorem 1, one readily obtains

$$\mathbb{E}\left[(M_{l,n})_{[r]}\right] = \frac{(n)_{[rl]}}{l^r \, (\theta)_n} \sum_{j=1}^{n-rl+r} \theta^j \, |s(n-rl, j-r)| = \frac{(n)_{[rl]}}{l^r \, (\theta)_n} \, (\theta)_{[n-rl]}.$$

Using the classical sieve formula one easily shows the following, which appears to be new even in the case of Ewens partitions with the exception of the case $l = 1$ obtained in [7].

PROPOSITION 2. *If $(X_n)_{n \geq 1}$ is an exchangeable sequence associated to a Dirichlet process with parameter $\theta > 0$, then, for any $n \geq 1$ and $l = 1, \ldots, n$, the distribution of $M_{l,n}$ is of the form*

$$(21) \qquad \mathbb{P}[M_{l,n} = m_l] = \frac{n!}{m_l!(\theta)_n} \frac{\theta^{m_l}}{l^{m_l}} \sum_{t=0}^{[n/l]-m_l} \frac{(-1)^t(\theta)_{[n-m_l-tl]}}{(n-m_l l - tl)!} \frac{\theta^t}{l^t}.$$

On the basis of the result stated in Proposition 2, one can derive the asymptotic behaviour of $M_{l,n}$, namely that, for any $l \geq 1$

$$(22) \qquad\qquad\qquad M_{l,n} \xrightarrow{\mathrm{d}} W_l,$$

as $n \to +\infty$, where W_l is a random variable distributed according to a Poisson distribution with parameter θ/l. The limit result (22) is known in the literature and has been originally obtained in [1, 3]. See also [2] and references therein.

Turning attention to the conditional case, one can easily derive the following results. Theorem 1 provides an expression for the probability distribution of $O_{l,m}^{(n)}$, i.e.

$$\mathbb{P}\left[O_{l,m}^{(n)} = m_l\right]$$

$$= \sum_{t=0}^{m-m_l} (-1)^t \binom{m_l + t}{t} \sum_{\mathbf{c}^{(m_l+t)} \in \mathcal{C}_{j,m_l+t}} \frac{m!}{\prod_{i=1}^{m_l+t}(l - n_{c_i})!(m - \nu_t)!}$$

$$\times \prod_{i=1}^{m_l+t} (n_{c_i})_{l-n_{c_i}} \frac{\left(\theta + n - \sum_{i=1}^{m_l+t} n_{c_i}\right)_{m-\nu_t}}{(\theta + n)_m},$$

where we set $\nu_t = \sum_{i=1}^{m_l+t}(l - n_{c_i})$. Analogously, Theorem 2 provides an expression for the probability distribution of $N_{l,m}^{(n)}$, i.e.

$$\mathbb{P}\left[N_{l,m}^{(n)} = m_l\right] = \frac{\theta^{m_l}}{t^{m_l}} \sum_{t=0}^{m-m_l} \left(-\frac{\theta}{l}\right)^t \frac{m!}{t!m_l!(m - lm_l - lt)!} \frac{(\theta + n)_{m-lm_l-lt}}{(\theta + n)_m}.$$

Similarly, according to Corollary 1 and Corollary 2, and using the limiting result for non–central generalized factorial coefficients

$$\lim_{\sigma \to 0} \frac{\mathscr{C}(n, k; \sigma, \gamma)}{\sigma^k} = \sum_{i=k}^{n} \binom{n}{i} |s(i, k)|(-\gamma)_{n-i},$$

the Bayesian estimators of the number of "old" and of "new" species of size l generated by $(X_1, \ldots, X_{n+m})$, conditional on $(X_1, \ldots, X_n)$, are given by

$$(23) \qquad \hat{O}_{l,m}^{(n)} = \sum_{t=1}^{l} \binom{m}{l - t} m_t(t)_{l-t} \frac{(\theta + n - t)_{m-(l-t)}}{(\theta + n)_m}$$

and

$$(24) \qquad \hat{N}_{l,m}^{(n)} = (l - 1)! \binom{m}{l} \frac{\theta}{(\theta + n + m - l)_l}.$$

In particular, from (23) and (24) the Bayesian estimator of the number of clusters of size l over an enlarged sample of size $n + m$, conditional on the partition structure of the n observed data, is given in the following proposition.

PROPOSITION 3. *If $(X_n)_{n\geq 1}$ is an exchangeable sequence associated to a Dirichlet process with parameter θ, then*

$$\hat{M}_{l,m}^{(n)} = \binom{m}{l} \frac{\theta (l - 1)!}{(\theta + n + m - l)_l} + \sum_{t=1}^{l} \binom{m}{l - t} m_t(t)_{l-t} \frac{(\theta + n - t)_{m-l+t}}{(\theta + n)_m}$$

for any $l \in \{1, \ldots, n + m\}$.

Finally, by combining (16) and (20) a simple limiting argument leads to show that, as $m \to +\infty$ and for any $l \geq 1$, $N_{l,m}^{(n)} \overset{d}{\longrightarrow} W_l^{(n)}$ and

$$(25) \qquad M_{l,m}^{(n)} \overset{d}{\longrightarrow} W_l^{(n)},$$

where $W_l^{(n)}$ is a random variable distributed according to a Poisson distribution with parameter $(\theta + n)/l$. Clearly, (25) reduces to (22) in the unconditional case corresponding to $n = 0$.

3.2. *The two–parameter Poisson–Dirichlet process.* The Pitman model with parameters (σ, θ) in (8), or $\mathrm{PD}(\sigma, \theta)$ process, stands out for its analytical tractability and for its modeling flexibility. In particular, within the species sampling context, the presence of the additional parameter $\sigma \in (0, 1)$, w.r.t. the simple Dirichlet model, allows to model more effectively both the clustering structure featured by the X_i's and the growth rate of K_n. Therefore, given its importance, we devote special attention to this process. A few additional asymptotic results that complement, for the specific case we are analyzing, those recalled in Section 2 for general Gibbs–type r.p.m.s are of particular interest.

3.2.1. *Distributional results.* Let us first state a result concerning the unconditional distribution of $M_{l,n}$, namely the number of clusters with frequency l in a sample of size n.

PROPOSITION 4. *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a $PD(\sigma, \theta)$ process with $\sigma \in (0, 1)$ and $\theta > -\sigma$. Then,*

$$(26) \quad \mathbb{P}[M_{l,n} = m_l] = \sum_{t=0}^{n-m_l} (-1)^t \frac{n!}{t! m_l! (n - lm_l - lt)!} \sigma^{m_l + t} \left(\frac{\theta}{\sigma} \right)_{m_l + t}$$

$$\times \left(\frac{(1-\sigma)_{l-1}}{l!} \right)^{m_l + t} \frac{(\theta + (m_l + t)\sigma)_{n - lm_l - lt}}{(\theta)_n}.$$

Hence, (26) provides the marginal distribution of the Pitman sampling formula (10), corresponding to $V_{n,j} = \sigma^j (\theta/\sigma)_j / (\theta)_n$, and, to the authors' knowledge, it is not explicitly reported in the literature.

Turning attention to the conditional case, one can easily derive the following results.

PROPOSITION 5. *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a $PD(\sigma, \theta)$ process with $\sigma \in (0, 1)$ and $\theta > -\sigma$. Then,*

$$(27) \quad \mathbb{P}\left[O_{l,m}^{(n)} = m_l \right] = \sum_{t=0}^{m-m_l} (-1)^t \binom{m_l + t}{t}$$

$$\times \sum_{\boldsymbol{c}^{(m_l+t)} \in \mathcal{C}_{j,m_l+t}} \binom{m}{l - n_{c_1}, \ldots, l - n_{c_{m_l+t}}, \sum_{i=1}^{m_l+t} (l - n_{c_i})} \prod_{i=1}^{m_l+t} (n_{c_i} - \sigma)_{l - n_{c_i}}$$

$$\times \ \frac{\left(\theta + n - \sum_{i=1}^{m_l+t} n_{c_i} + (m_l + t)\sigma\right)_{m - \sum_{i=1}^{m_l+t}(l - n_{c_i})}}{(\theta + n)_m}$$

for any $l \in \{1, \ldots, n\}$ and $m_l \in \{1, \ldots, n\}$ such that $m_l l \leq n$.

From (27) one can deduce a completely explicit expression for the Bayesian estimator of the number of "old" species with frequency l in the whole sample $X_1, \ldots, X_{n+m}$, namely

$$(28) \quad \hat{O}_{l,m}^{(n)} = \mathbb{E}[O_{l,m}^{(n)}] = \sum_{t=1}^{l} \binom{m}{l-t} m_t (t - \sigma)_{l-t} \frac{(\theta + n - t + \sigma)_{m-(l-t)}}{(\theta + n)_m},$$

which can be readily used in applications as will be shown in Section 4. In a similar fashion it is possible to deduce the distribution of the number of "new" species that will appear l times in $(X_{n+1}, \ldots, X_{n+m})$ conditional on the observations $(X_1, \ldots, X_n)$. Indeed one can show the following

PROPOSITION 6.   *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a $PD(\sigma, \theta)$ process with $\sigma \in (0, 1)$ and $\theta > -\sigma$. Then,*

$$(29) \ \ \mathbb{P}[N_{l,m}^{(n)} = m_l] = \sum_{t=0}^{m-m_l} (-1)^t \binom{m}{t, m_l, m - lm_l - lt} \prod_{i=0}^{m_l + t - 1} (\theta + j\sigma + i\sigma)$$

$$\times \left(\frac{(1-\sigma)_{l-1}}{l!}\right)^{m_l + t} \frac{(\theta + n + (m_l + t)\sigma)_{m - l(m_l + t)}}{(\theta + n)_m},$$

for any $n \geq 1$, $j = 1, \ldots, n$, $l \geq 1$ and $m_l \geq 1$ such that $m_l l \leq m$.

REMARK 2.   One can alternatively prove (29) by relying on the so–called quasi–conjugacy property of the two–parameter Poisson–Dirichlet process, a concept introduced in [20]. Indeed, it suffices to marginalize an *updated* Pitman sampling formula and (29) easily follows. Moreover, if $n = j = 0$ in (29) one recovers the marginal distribution of $M_{l,n}$ as described in (26) and, if one additionally sets $\sigma = 0$, the distribution of $M_{l,n}$ corresponding to the Ewens partition in (21) is obtained.

The Bayesian estimator for the number of "new" species with frequency l over the enlarged sample $n + m$ coincides with

$$(30) \qquad \hat{N}_{l,m}^{(n)} = \mathbb{E}[N_{l,m}^{(n)}] = \binom{m}{l} (1 - \sigma)_{l-1} (\theta + j\sigma) \frac{(\theta + n + \sigma)_{m-l}}{(\theta + n)_m}$$

for any $l \in \{1, \ldots, m\}$. Having determined $\hat{O}_{l,m}^{(n)}$ and $\hat{N}_{l,m}^{(n)}$, one finds out that a Bayesian estimator of the total number of species with frequency l among $(X_1, \ldots, X_{n+m})$, given $(X_1, \ldots, X_n)$, is given by

PROPOSITION 7. *If $(X_n)_{n \geq 1}$ is an exchangeable sequence with $\tilde{P}$ in* (2) *being the $PD(\sigma, \theta)$ process, for any $l = 1, \ldots, n+m$,*

$$(31) \qquad \hat{M}_{l,m}^{(n)} = \sum_{t=1}^{l} \binom{m}{l-t} m_t (i - \sigma)_{l-t} \frac{(\theta + n - t + \sigma)_{m-(l-t)}}{(\theta + n)_m}$$

$$+ \binom{m}{l}(1 - \sigma)_{l-1}(\theta + j\sigma)\frac{(\theta + n + \sigma)_{m-l}}{(\theta + n)_m}.$$

Of course, Theorem 3 allows a direct evaluation of $\hat{M}_{l,m}^{(n)}$ above and yields moments of any order $r \geq 1$ of $M_{l,m}^{(n)}$.

3.2.2. *Asymptotics.* We now study the asymptotic behaviour of $M_{l,m}^{(n)}$ and $N_{l,m}^{(n)}$, as $m \to \infty$. However, before proceeding, let us first recall a well–known result concerning the asymptotics of $M_{l,n}$ as n increases. To this end, let f_σ be the density function of a positive σ–stable random variable and Y_q, for any $q \geq 0$, a positive random variable with density function

$$f_{Y_q}(y) = \frac{\Gamma(q\sigma + 1)}{\sigma\Gamma(q+1)} y^{q-1/\sigma-1} f_\sigma\left(y^{-1/\sigma}\right).$$

Then, for any $l \geq 1$

$$\frac{M_{l,n}}{n^\sigma} \xrightarrow{d} \frac{\sigma(1-\sigma)_{(l-1)}}{l!} Y_{\theta/\sigma},$$

as $n \to +\infty$. See [23] for details. We now provide a new result concerning the limiting behaviour in the conditional case and, specifically, of $M_{l,m}^{(n)}$ and of $N_{l,m}^{(n)}$ as $m \to \infty$. It will be shown that they converge in distribution to the same random element that still depends on Y_q for a suitable choice of q.

THEOREM 4. *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a $PD(\sigma, \theta)$ process. For any $1 \leq j \leq n$ and $l \geq 1$, one has*

$$(32) \qquad \frac{N_{l,m}^{(n)}}{m^\sigma} \xrightarrow{d} \frac{\sigma(1-\sigma)_{l-1}}{l!} Z_{n,j}$$

as $m \to +\infty$, *where* $Z_{n,j} \overset{d}{=} B_{j+\theta/\sigma,n/\sigma-j}Y_{(\theta+n)/\sigma}$ *and* $B_{j+\theta/\sigma,n/\sigma-j}$ *is a beta random variable with parameters* $(j+\theta/\sigma, n/\sigma-j)$ *independent of* $Y_{(\theta+n)/\sigma}$. *Moreover*

$$(33) \qquad \frac{M_{l,m}^{(n)}}{m^{\sigma}} \overset{d}{\longrightarrow} \frac{\sigma(1-\sigma)_{l-1}}{l!} Z_{n,j}$$

as $m \to +\infty$.

The limit in (32) and (33) implies that K_n is asymptotically sufficient for predicting the conditional number of distinct species with frequency l to be generated by the additional sample $(X_{n+1}, \ldots, X_{n+m})$ as its size m increases. Such a limit involves the beta–tilted random variable $Z_{n,j}$, originally introduced in [8] by investigating the asymptotic behaviour of the conditional number of "new" distinct species $K_m^{(n)}$ generated by the additional sample as its size m increases. Specifically,

$$\frac{K_m^{(n)}}{m^{\sigma}} \to Z_{n,j},$$

almost surely, as $m \to +\infty$. It is worth noting that beta–tilted random variables of similar type have been recently object of a thorough investigation in [14] in the context of the so–called Lamperti–type laws.

REMARK 3.   Note that from (32) and (33) one obtains the unconditional result of [23] by setting $n = j = 0$. Moreover, one recovers a result in [8], which states that, conditional on $(X_1, \ldots, X_n)$, $m^{-\sigma} K_m^{(n)} \overset{d}{\to} Z_{n,j}$, as $m \to \infty$. Indeed, $K_m^{(n)} = \sum_{l=1}^{L_m^{(n)}} N_{l,m}^{(n)}$ and $L_m^{(n)}$ diverges as $m \to +\infty$: hence the limit in distribution for $K_m^{(n)}$ can be deduced from (32) upon noting that $\sum_{l \geq 1}(l!)^{-1}\sigma(1-\sigma)_{l-1} = 1$.

3.3. *The Gnedin model.*   Consider now the Gnedin model (9) with parameters $\zeta = 0$ and $\gamma \in [0,1)$. The corresponding random partition is representable as a mixture partitions of the type (8), however with parameters $(-1, \kappa)$, each of which generates a partition with a finite number of blocks κ. The mixing distribution for the total number of blocks is $p(\kappa) = \gamma\,(1-\gamma)_{\kappa-1}/\kappa!x$.

PROPOSITION 8.   *Let* $(X_n)_{n \geq 1}$ *be an exchangeable sequence associated to the Gnedin model with parameters* $(0, \gamma)$. *Then*

$$(34) \quad \mathbb{E}\left[(M_{l,n})_{[r]}\right] = \mathbb{1}_{\{rl\}}(n) \frac{r!\, l\, (\gamma)_{rl-r}(1-\gamma)_{r-1}}{(1+\gamma)_{rl-1}}$$

$$+ \mathbb{1}_{\{rl+1,\dots\}}(n) \frac{n(\gamma)_{rl-r}(1-\gamma)_r}{(1+\gamma)_{n-1}} \sum_{k=0}^{n-rl-1} \binom{n-rl-1}{k}$$

$$\times \frac{(r+k)!}{(1+k)!}\,(\gamma+rl-r)_{n-rl-1-k}(r+1-\gamma)_k.$$

From (34) one can determine the probability distribution of $M_{l,n}$. Indeed, if $n/l \notin \mathbb{N}$, then

$$\mathbb{P}[M_{l,n} = m_l] = \frac{\mathbb{1}_{\{1,\dots,n\}}(lm_l)\, n}{m_l!(1+\gamma)_{n-1}} \sum_{r=m_l}^{[n/l]} \frac{(-1)^{r-m_l}}{(r-m_l)!}\,(\gamma)_{rl+r}(1-\gamma)_r$$

$$\times \sum_{k=0}^{n-rl-1} \binom{n-rl-1}{k} \frac{(r+k)!}{(1+k)!}(\gamma+rl+r)_{n-rl-1-k}(r+1-\gamma)_k.$$

On the other hand, if $n/l \in \mathbb{N}$, then

$$\mathbb{P}[M_{l,n} = m_l] = \frac{\mathbb{1}_{\{1,\dots,n\}}(lm_l)\, n}{m_l!(1+\gamma)_{n-1}} \left\{ (-1)^{\frac{n}{l}-m_l} \frac{(\frac{n}{l}-1)!(\gamma)_{n-\frac{n}{l}}(1-\gamma)_{\frac{n}{l}-1}}{(\frac{n}{l}-m_l)!} \right.$$

$$+ \sum_{r=m_l}^{\frac{n}{l}-1} \frac{(-1)^{r-m_l}}{(r-m_l)!}\,(\gamma)_{rl+r}(1-\gamma)_r$$

$$\times \left. \sum_{k=0}^{n-rl-1} \binom{n-rl-1}{k} \frac{(r+k)!}{(1+k)!}(\gamma+rl+r)_{n-rl-1-k}(r+1-\gamma)_k \right\}.$$

Moreover, for any $l \geq 1$

$$(35) \qquad\qquad\qquad M_{l,n} \xrightarrow{\text{d}} 0,$$

as $n \to +\infty$. Note that the limiting result in (35) is not surprising since a Gnedin r.p.m. induces a random partition of $\mathbb{N}$ into an almost surely finite number of blocks even though with infinite expectation [11].

As for the posterior distribution of the number of clusters of size l, we now use the general results outlined in Section 2 to provide some explicit forms for the distribution of $O_{l,m}^{(n)}$ and $N_{l,m}^{(n)}$.

PROPOSITION 9.   *Let $(X_n)_{n\geq 1}$ be an exchangeable sequence associated to the Gnedin model with parameters $\zeta = 0$ and $\gamma \in [0,1)$. Then,*

$$\mathbb{P}\left[O_{l,m}^{(n)} = m_l\right] = \frac{\mathbb{1}_{\{1,\ldots,n\}}(lm_l)\, m!(m+n+j-r-rl-1)!}{(n)_m(\gamma+n)_m}$$

$$\times \sum_{r=m_l}^{[n/l]}(-1)^{r-m_l}\binom{r}{m_l}\sum_{\boldsymbol{c}^{(r)}\in\mathcal{C}_{j,r}}\frac{1}{(m-rl+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|)!}\prod_{i=1}^{r}\frac{(n_{c_i}-\sigma)_{l-n_{c_i}}}{(l-n_{c_i})!}$$

$$\times \sum_{k=0}^{m-rl+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|}\binom{m-rl+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|}{k}\frac{(j)_k\,(\gamma+n-j)_{m-k}}{(n-|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|+j-r-1+k)!}.$$

Moreover

$$\mathbb{P}\left[N_{l,m}^{(n)} = m_l\right] = \frac{\mathbb{1}_{\{1,\ldots,m\}}(lm_l)\, m!}{(n)_m(\gamma+n)_m}\sum_{r=m_l}^{[n/l]}(-1)^{r-m_l}\frac{(m-rl+n+j)!}{(r-m_l)!(m-rl)!}$$

$$\times \sum_{k=0}^{m-rl}\binom{m-rl}{k}\frac{(\gamma+n-j)_{m-r-k}(j)_{k+r}(j-\gamma)_{k+r}}{(n+j+k)!}.$$

One can further deduce the conditional expected values of $O_{l,m}^{(n)}$ and of $N_{l,m}^{(n)}$ which take on the following forms

$$\hat{O}_{l,m}^{(n)} = \frac{1}{(n)_m(\gamma+n)_m}\sum_{t=1}^{l}m_t\binom{m}{l-t}(t+1)_{l-t}\,(m+n+j-l-2)!$$

$$\times \sum_{k=0}^{m-l+t}\binom{m-l+t}{k}\frac{(\gamma+n-j)_{m-k}(j)_k}{(n+j-t-2+k)!}$$

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

As in previous examples, these quantities can, then, be used in order to provide a Bayesian estimator $\hat{M}_{l,m}^{(n)} = \hat{O}_{l,m}^{(n)} + \hat{N}_{l,m}^{(n)}$ of the number of species of size l over the enlarged sample of size $n+m$, conditional on the sample $(X_1,\ldots,X_n)$.

Finally, by combining Theorems 2 and 3 with the specific weights (9) it can be easily verified that for any $l \geq 1$

$$N_{l,m}^{(n)} \xrightarrow{\text{d}} 0 \qquad M_{l,m}^{(n)} \xrightarrow{\text{d}} 0,$$

as $m \to +\infty$. As in the unconditional case, these limits are not surprising due to the almost sure finiteness of the number of blocks of a random partition induced by the Gnedin model.

4. Genomic applications. A Bayesian nonparametric model (2), with $\tilde{P}$ being a Gibbs–type r.p.m. with $\sigma > 0$, is particularly suited for inferential problems with a large unknown number of species given it postulates an infinite number of species. These usually occur in genomic applications, such as the analysis of Expressed Sequence Tags (EST), Cap Analysis Gene Expression (CAGE) or Serial Analysis of Gene Expression (SAGE). See, e.g., [26, 19, 5]. The typical situation is as follows: a sample of size n sequenced from a genomic library is available and one would like to make predictions, over an enlarged sample of size $n + m$ and conditionally on the observed sample, of certain quantities of interest. The most obvious quantity is the number of distinct species to be observed in the enlarged sample, which represents a measure of the overall genes variety. The resulting Bayesian nonparametric estimators proposed in [17, 20] have already been integrated into the web server RichEst$^\copyright$ [5]. However estimators for the overall genes variety are certainly useful but necessarily need to be complemented by an effective analysis of the so–called "rare genes variety" (see e.g. [26]). Therefore, from an applied perspective it is important to devise estimators of the number of genes that appear only once, the so–called *unigenes* or, more generally, of the number of genes that are observed with frequency less than or equal to a specific abundance threshold τ. The results deduced in the present paper perfectly fit these needs. Indeed, conditional on an observed sample of size n, the quantity $\hat{M}_{1,m}^{(n)} = \mathbb{E}[M_{l,m}^{(n)}]$ is a Bayesian estimator of the number of genes that will appear only once in a sample of size $n + m$ and can be easily determined from Theorem 3. In a similar fashion, having fixed a threshold τ,

$$(36) \qquad \hat{M}_\tau^{(n)} = \sum_{l=1}^\tau \hat{M}_{l,m}^{(n)}$$

is a Bayesian estimator of the rare genes variety, namely the number of species appearing with frequency less than τ in a sample of size $n + m$.

Having laid out the framework and described the estimators to be used, we now test the proposed methodology on some real genomic data. To this end we deal with a widely used EST dataset obtained by sequencing a tomato–flower cDNA library (made from 0—3 mm buds of tomato flowers) from the Institute for Genomic Research Tomato Gene Index with library identifier T1526 [24]. The observed sample consists of $n = 2586$ ESTs with $j = 1825$

unique genes, whose frequencies can be summarized by

$$m_{i,2586} = 1434, 253, 71, 33, 11, 6, 2, 3, 1, 2, 2, 1, 1, 1, 2, 1, 1$$

with $i \in \{1, 2, \ldots, 14\} \cup \{16, 23, 27\}$, which means that we are observing 1434 genes which appear once, 253 genes which appear twice, etc.

As for the specific model (2) we adopt, $\tilde{P}$ is a $\mathrm{PD}(\sigma, \theta)$ process. The reason we rely on such a specification is two–fold: on the one hand it yields tractable estimators that can be exactly evaluated and, on the other, it is a very flexible model since it encompasses a wide range of partitioning structures according as to the value of σ. On the basis of our choice of the nonparametric prior, we only need to specify the parameter vector (σ, θ). This is achieved by adopting an empirical Bayes procedure [17]: we fix (σ, θ) so to maximize (8) corresponding to the observed sample $(j, n_1, \ldots, n_j)$, i.e.

$$(37) \qquad (\hat{\sigma}, \hat{\theta}) = \arg\max_{(\sigma, \theta)} \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{i=1}^{j}(1 - \sigma)_{n_i - 1}.$$

The quantities we wish to estimate are $N_\tau^{(n)} = \sum_{l=1}^{\tau} N_{l,m}$ and $O_\tau^{(n)} = \sum_{l=1}^{\tau} O_{l,m}$. These quantities identify the number of distinct genes with abundances not greater than τ or, in genomic terminology, with expression levels not greater than τ that are present among the "new" genes detected in the additional sample and the "old" genes observed in the basic sample, respectively. The overall number of rare distinct genes is easily recovered as $M_\tau^{(n)} = N_\tau^{(n)} + O_\tau^{(n)}$. The corresponding estimators can be deduced from (28), (30) and (31). In the present genomic context one can reasonably identify the rare genes as those presenting expression levels less than or equal to $\tau = 3, 4, 5$, which are the thresholds we employ for our analysis.

We first perform a cross–validation study for assessing the performance of our methodology when used to predict rare genes abundance. To this end 10 sub–samples of size 1000 have been drawn without replacement from the available 2586 EST sample. For each of the sub–samples we have generated, the corresponding values of (σ, θ) have been fixed according to (37). Predictions have, then, been performed for an additional sample of size $m = 1586$, which corresponds to the remaining observed genes. Table 1 below reports the true and estimated values for the $O_\tau^{(n)}$, $N_\tau^{(n)}$ and $M_\tau^{(n)}$ and shows the accurate performance of the proposed estimators. Such a result is a fortiori appreciable if one considers that predictions are made over an additional sample of size larger than 1.5 times the observed sample.

TABLE 1 AROUND HERE

We now deal with the whole dataset and provide estimates of rare genes abundance after additional sequencing. To this end, we consider, as possible sizes of the additional sample, $m \in \{250, 500, 750, 1000\}$. As for the prior specification of (σ, θ) the maximization in (37) leads to $(\hat{\sigma}, \hat{\theta}) = (0.612, 741)$. The resulting estimates of $O_\tau^{(n,j)}$, $N_\tau^{(n,j)}$ and $M_\tau^{(n,j)}$ are reported in Table 2.

<center>TABLE 2 AROUND HERE</center>

5. Proofs. We start by providing a lemma concerning the marginal frequency counts of the partition blocks induced by Gibbs–type random partition. In addition to the notation introduced in Section 2, we define the following shortened set notation

$$A_{n,m}(j, \boldsymbol{n}, s, k) := \{K_n = j, \boldsymbol{N} = \boldsymbol{n}, L_m^{(n)} = s, K_m^{(n)} = k\}$$

and

$$A_n(j, \boldsymbol{n}) := \{K_n = j, \boldsymbol{N} = \boldsymbol{n}\}.$$

for any $\boldsymbol{n} = (n_1, \ldots, n_j) \in \mathcal{D}_{n,j}$. Further additional notations will be introduced in the proofs when necessary.

LEMMA 1. *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence associated to a Gibbs–type r.p.m. For any $x \in \{1, \ldots, j\}$, let $\boldsymbol{q}^{(x)} = (q_1, \ldots, q_x)$ with $1 \leq q_1 < \cdots < q_x \leq j$ and define the vector of frequency counts $\boldsymbol{S}_{\boldsymbol{q}^{(x)}} := (S_{q_1}, \ldots, S_{q_x})$. Then,*

$$(38) \quad \mathbb{P}[\boldsymbol{S}_{\boldsymbol{q}^{(x)}} = \boldsymbol{s}_{\boldsymbol{q}^{(x)}} \mid A_{n,m}(j, \boldsymbol{n}, s, k)]$$

$$= \frac{(m-s)!}{(m-s-|\boldsymbol{s}_{\boldsymbol{q}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{q_i} - \sigma)_{s_{q_i}}}{s_{q_i}!}$$

$$\times \frac{(n - |\boldsymbol{n}_{\boldsymbol{q}^{(x)}}| - (j-x)\sigma)_{m-s-|\boldsymbol{s}_{\boldsymbol{q}^{(x)}}|}}{(n - j\sigma)_{m-s}}$$

for any vector $\boldsymbol{s}_{\boldsymbol{q}^{(x)}} = (s_{q_1}, \ldots, s_{q_x})$ of non-negative integers such that $|\boldsymbol{s}_{\boldsymbol{q}^{(x)}}| = \sum_{i=1}^{x} s_{q_i} \leq m - s$. Moreover, for any $y \in \{1, \ldots, k\}$, let $\boldsymbol{r}^{(y)} = (r_1, \ldots, r_y)$ with $1 \leq r_1 < \cdots < r_y \leq k$ and define the vector of frequency counts $\boldsymbol{S}_{\boldsymbol{r}^{(y)}}^* := (S_{j+r_1}, \ldots, S_{j+r_y})$. Then

$$(39) \quad \mathbb{P}\left[\boldsymbol{S}_{\boldsymbol{r}^{(y)}}^* = \boldsymbol{s}_{\boldsymbol{r}^{(y)}} \mid A_{n,m}(j, \boldsymbol{n}, s, k)\right]$$

$$= \frac{s!}{(s - |\boldsymbol{s}_{\boldsymbol{r}^{(y)}}|)!} \prod_{i=1}^{y} \frac{(1-\sigma)_{s_{j+r_i}}}{s_{j+r_i}!}$$

$$\times \frac{(k-y)!}{k!}\sigma^y \frac{\mathscr{C}\left(s-|\boldsymbol{s}_{\boldsymbol{r}^{(y)}}|,k-y;\sigma\right)}{\mathscr{C}\left(s,k;\sigma\right)}$$

for any vector $\boldsymbol{s}_{\boldsymbol{r}^{(y)}} = (s_{j+r_1},\ldots,s_{j+r_y})$ of positive integers such that $|\boldsymbol{s}_{\boldsymbol{r}^{(y)}}| = \sum_{i=1}^{y} s_{j+r_i} \le s$. Moreover, the random variables $\boldsymbol{S}_{\boldsymbol{q}^{(x)}}$ and $\boldsymbol{S}^*_{\boldsymbol{r}^{(y)}}$ are independent, conditionally on $(K_n, \boldsymbol{N}, L_m^{(n)}, K_m^{(n)})$.

PROOF. We start by recalling some useful conditional formulae for Gibbs–type random partitions recently obtained in [20]. In particular, from [20, Corollary 1] one has the conditional probability

$$(40) \quad \mathbb{P}\left[K_m^{(n)} = k, L_m^{(n)} = s \,|\, A_n(j,\boldsymbol{n})\right]$$

$$= \frac{V_{n+m,j+k}}{V_{n,j}}\binom{m}{s}(n-j\sigma)_{m-s}\frac{\mathscr{C}(s,k,\sigma)}{\sigma^k}.$$

On the other hand, for any vectors of non-negative integers $\boldsymbol{s}_{\boldsymbol{q}^{(j)}} = (s_1,\ldots,s_j)$ such that $|\boldsymbol{s}_{\boldsymbol{q}^{(j)}}| = m - s$, and for any vector of positive integers $\boldsymbol{s}_{\boldsymbol{r}^{(k)}} = (s_{j+1},\ldots,s_{j+k})$ such that $|\boldsymbol{s}_{\boldsymbol{r}^{(k)}}| = s$, [20, Equation (28)] yields the conditional probability

$$(41) \quad \mathbb{P}\left[\boldsymbol{S}_{\boldsymbol{q}^{(j)}} = \boldsymbol{s}_{\boldsymbol{q}^{(j)}}, \boldsymbol{S}^*_{\boldsymbol{r}^{(k)}} = \boldsymbol{s}_{\boldsymbol{r}^{(k)}}, L_m^{(n)} = s, K_m^{(n)} = k \,|\, A_n(j,\boldsymbol{n})\right]$$

$$= \frac{V_{n+m,j+k}}{V_{n,j}}\prod_{i=1}^{j}(n_i-\sigma)_{s_i}\prod_{\ell=1}^{k}(1-\sigma)_{s_{j+\ell}-1}.$$

A combination of (40) and (41) implies that

$$(42) \quad \mathbb{P}\left[\boldsymbol{S}_{\boldsymbol{q}^{(j)}} = \boldsymbol{s}_{\boldsymbol{q}^{(j)}}, \boldsymbol{S}^*_{\boldsymbol{r}^{(k)}} = \boldsymbol{s}_{\boldsymbol{r}^{(k)}} \,|\, A_{n,m}(j,\boldsymbol{n},s,k)\right]$$

$$= \frac{\sigma^k \prod_{i=1}^{j}(n_i-\sigma)_{s_{q_i}-1}\prod_{\ell=1}^{k}(1-\sigma)_{s_{j+r_\ell}-1}}{\binom{m}{s}(n-j\sigma)_{m-s}\mathscr{C}(s,k,\sigma)}.$$

Consider now the set $\mathcal{I}_{j,x} := \{1,\ldots,j\}\setminus\{q_1,\ldots,q_x\}$ and the corresponding partition set defined as follows

$$\mathcal{D}^{(0)}_{m-s-s^*,j-x} := \left\{(s_i, i \in \mathcal{I}_{j,x}) : s_i \ge 0 \text{ and } \sum_{i\in\mathcal{I}_{j,x}} s_i = m - s - s^*\right\},$$

where we set $s^* := \sum_{i=1}^{x} s_{q_i}$. In a similar vein, let us introduce the set $\mathcal{I}_{k,y} := \{1,\ldots,k\}\setminus\{r_1,\ldots,r_y\}$ and the corresponding partition set defined

as follows

$$\mathcal{D}_{s-s^{**},k-y} := \left\{ (s_{j+i}, i \in \mathcal{I}_{k,y}) : s_{j+i} > 0 \ \ \text{and} \ \ \sum_{i \in \mathcal{I}_{k,y}} s_{j+i} = s - s^{**} \right\},$$

where we set $s^{**} := \sum_{i=1}^{y} s_{j+r_i}$. By virtue of [4, Equation (2.6.1)] one can write

$$(43) \quad \frac{1}{(k-y)!} \sum_{\mathcal{D}_{s-s^{**},k}} s! \prod_{i=1}^{k} \frac{(1-\sigma)_{s_{j+i}-1}}{s_{j+i}!}$$

$$= \frac{s!}{(s-s^{**})! \prod_{i=1}^{y} s_{r_i}!} \frac{\mathscr{C}(s-s^{**}, k-y, \sigma)}{\sigma^{k-y}}$$

and, by virtue of [20, Lemma (A.1)], one can write

$$(44) \quad \sum_{\mathcal{D}_{m-s-s^{*},j-x}^{(0)}} \binom{m-s}{s_1,\ldots,s_j} \prod_{i=1}^{j} (1-\sigma)_{n_i+s_i-1}$$

$$= \frac{(m-s)!(n^* - (j-x)\sigma)_{m-s-s^*}}{(m-s-s^*)! \prod_{i=1}^{x} s_{q_i}!}$$

$$\prod_{i=1}^{x} (1-\sigma)_{n_{q_i}+s_{q_i}-1} \prod_{\ell \in \mathcal{I}_{j,x}} (1-\sigma)_{n_\ell-1}$$

where we set $n^* := \sum_{i \in \mathcal{I}_{j,x}} n_i = n - \sum_{i=1}^{x} n_{q_i}$. A simple application of the identities (43) and (44) to the conditional probability (42) proves both the conditional independence between $\boldsymbol{S}_{\boldsymbol{q}^{(x)}}$ and $\boldsymbol{S}^*_{\boldsymbol{r}^{(y)}}$ and the two expressions in (38) and (39). □

5.1. *Proof of Proposition 1.* For any $n \geq 1$ and $1 \leq j \leq n$ let $\mathcal{M}_{n,j}$ be the partition set of $\mathbb{N}_n$ containing all the vectors $\boldsymbol{m}_n = (m_1, \ldots, m_n) \in \{0, 1, \ldots, n\}^n$ such that $\sum_{i=1}^{n} m_i = j$ and $\sum_{i=1}^{n} im_i = n$. Hence, resorting to the probability distribution (10), one obtains for any $r \geq 1$

$$\mathbb{E}\left[(M_{l,n})_{[r]}\right] = n! \sum_{j=1}^{n} V_{n,j} \sum_{\boldsymbol{m}_n \in \mathcal{M}_{n,j}} (m_l)_{[r]} \prod_{i=1}^{n} \left(\frac{(1-\sigma)_{i-1}}{i!}\right)^{m_i} \frac{1}{m_i!}$$

$$= n! \sum_{j=1}^{n} V_{n,j} \sum_{\boldsymbol{m}_n \in \mathcal{M}_{n,j}} \left(\frac{(1-\sigma)_{l-1}}{l!}\right)^{m_l} \frac{1}{(m_l - r)!}$$

$$\times \prod_{1 \leq i \neq l \leq n} \left(\frac{(1-\sigma)_{i-1}}{i!} \right)^{m_i} \frac{1}{m_i!}$$

$$= n! \left(\frac{(1-\sigma)_{l-1}}{l!} \right)^r \sum_{j=1}^{n} V_{n,j}$$

$$\times \sum_{\boldsymbol{m}_{n-rl} \in \mathcal{M}_{n-rl,j-r}} \prod_{i=1}^{n-rl} \left(\frac{(1-\sigma)_{i-1}}{i!} \right)^{m_i} \frac{1}{m_i!}.$$

Finally, a direct application of [4, Equation (2.82)] implies the following identity

$$\sum_{\boldsymbol{m}_n \in \mathcal{M}_{n-rl,j-r}} \prod_{i=1}^{n} \left(\frac{(1-\sigma)_{i-1}}{i!} \right)^{m_i} \frac{1}{m_i!} = \frac{(n)_{[lr]}}{n! \sigma^{j-r}} \mathscr{C}(n-lr, j-r; \sigma)$$

and the proof is completed. □

5.2. *Proof of Theorem 1.* According to the definition of the random variable $O_{l,m}$ in (13), for any $r \geq 1$ one can write

$$\mathbb{E}\left[\left(O_{l,m}^{(n)} \right)^r \right] = \sum_{s=0}^{m} \sum_{k=0}^{s} \mathbb{P}\left[L_m^{(n)} = s, K_m^{(n)} = k \mid A_n(j, \boldsymbol{n}) \right]$$

$$\times \mathbb{E}\left[\left(\sum_{i=1}^{j} \mathbb{1}_l(n_i + S_i) \right)^r \;\middle|\; A_{n,m}(j, \boldsymbol{n}, s, k) \right].$$

It can be easily verified that a repeated application of the binomial expansion implies the following identity

$$(45) \quad \left(\sum_{i=1}^{j} \mathbb{1}_{\{l\}}(n_i + S_i) \right)^r = \sum_{x=1}^{j} \sum_{i_1=1}^{r-1} \sum_{i_2=1}^{i_2-1} \cdots \sum_{i_{x-1}=1}^{i_{x-2}-1} \binom{r}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x-2}}{i_{x-1}}$$

$$\times \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \prod_{t=1}^{x} \left(\mathbb{1}_{\{l\}}(n_{c_t} + S_{c_t}) \right)^{i_{x-t}-i_{x-t+1}}$$

provided $i_0 \equiv r$. Observe that the previous sum can be expressed in terms of Stirling numbers of the second kind $S(n,m)$; indeed, since $m!S(n,m)$ is the number of ways of distributing n distinguishable objects into m distinguishable groups, one has

$$(46) \quad \frac{1}{m!} \sum_{i_1=1}^{n-1} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{m-1}=1}^{i_{m-2}-1} \binom{n}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{m-2}}{i_{m-1}} = S(n,m),$$

for any $n \geq 1$ and $1 \leq m \leq n$. In particular, combining the identity (45) with (46) one obtains

$$(47) \quad \mathbb{E}\left[\left(O_{l,m}^{(n)}\right)^r \,\Big|\, L_m^{(n)} = s, K_m^{(n)} = k\right] = \sum_{x=1}^{j \wedge r} S(r,x)x!$$
$$\times \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \mathbb{P}\left[\boldsymbol{S}_{\boldsymbol{c}^{(x)}} = l\boldsymbol{1}_x - \boldsymbol{n}_{\boldsymbol{c}^{(x)}} \mid A_{n,m}(j, \boldsymbol{n}, s, k)\right]$$

where we set $\boldsymbol{1}_x := (1, \ldots, 1)$ and $\boldsymbol{n}_{\boldsymbol{c}^{(x)}} = (n_{c_1}, \ldots, n_{c_x})$. In (47) the bound $j \wedge r$ on the sum over the index x is motivated by the fact that $S(r,x) = 0$ if $x > r$. Hence, the identity (47) combined with (38) yields the following expression

$$(48) \quad \mathbb{E}\left[\left(O_{l,m}^{(n)}\right)^r \,\Big|\, L_m^{(n)} = s, K_m^{(n)} = k\right] = \sum_{x=1}^{j \wedge r} S(r,x)x!$$
$$\times \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \frac{(m-s)!}{(m-s-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{c_i} - \sigma)_{l-n_{c_i}}}{(l-n_{c_i})!}$$
$$\times \frac{(n - |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}| - (j-x)\sigma)_{m-s-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|}}{(n - j\sigma)_{m-s}}.$$

Observe that in (48) the sum over the index x, for $x = 1, \ldots, j \wedge r$, is equivalent to a sum over the index x for $x = 1, \ldots, r$. Indeed, if $j > r$ then the sum over the index x is non–null for $x = 1, \ldots, r$ because $S(r,x) = 0$ for any $x = r+1, \ldots, j$; on the other hand, if $j < r$ then the sum over the index x is non–null for $x = 1, \ldots, j$ because the set $\mathcal{C}_{j,x}$ is empty for any $x = j+1, \ldots, r$. Accordingly, resorting to [20, Corollary 1] one can rewrite the expected value above as

$$\mathbb{E}\left[\left(O_{l,m}^{(n)}\right)^r\right] = \sum_{s=0}^{m} \sum_{k=0}^{s} \frac{V_{n+m,j+k}}{V_{n,j}} \binom{m}{s} \frac{\mathscr{C}(s,k;\sigma)}{\sigma^k} \sum_{x=1}^{r} S(r,x)x!$$
$$\times \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \frac{(m-s)!}{(m-s-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{c_i} - \sigma)_{l-n_{c_i}}}{(l-n_{c_i})!}$$
$$\times \; (n - |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}| - (j-x)\sigma)_{m-s-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|}$$
$$= \sum_{x=1}^{r} S(r,x)x! \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \frac{m!}{(m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{c_i} - \sigma)_{l-n_{c_i}}}{(l-n_{c_i})!}$$

$$\times \sum_{k=0}^{m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|} \frac{V_{n+m,j+k}}{V_{n,j}} \sigma^{-k} \sum_{s=k}^{m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|} \binom{m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|}{s}$$

$$\times (n-|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|-(j-x)\sigma)_{m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|-s} \ \mathscr{C}(s,k;\sigma)$$

$$= \sum_{x=1}^{r} S(r,x)x! \sum_{\boldsymbol{c}^{(x)}\in\mathcal{C}_{j,x}} \frac{m!}{(m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{c_i}-\sigma)_{l-n_{c_i}}}{(l-n_{c_i})!}$$

$$\times \sum_{k=0}^{m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}\left(m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|,k;\sigma,-n+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|+(j-x)\sigma\right)}{\sigma^k}$$

where the last equality follows from [4, Equation(2.56)]. The proof of (14) is, thus, completed by using the relation between the r–th moment with the r–th factorial moment. $\qquad\square$

5.3. *Proof of Theorem 2.* The proof is along lines similar to the proof of Theorem 1. In particular, it can be easily verified that a repeated application of the binomial expansion implies the following identity

$$\left(\sum_{i=1}^{k} \mathbb{1}_{\{l\}}(S_{j+i})\right)^r = \sum_{y=1}^{k}\sum_{i_1=1}^{r-1}\sum_{i_2=1}^{i_2-1}\cdots\sum_{i_{y-1}=1}^{i_{y-2}-1} \binom{r}{i_1}\binom{i_1}{i_2}\cdots\binom{i_{y-2}}{i_{y-1}}$$

$$\times \sum_{\boldsymbol{c}^{(y)}\in\mathcal{C}_{k,y}} \prod_{t=1}^{y} \left(\mathbb{1}_{\{l\}}(S_{j+c_t})\right)^{i_{y-t}-i_{y-t+1}}.$$

Hence, according to the definition of the random variable $N_{l,m}$ in (13) and by combining the identity (46) with (39), one has

$$(49) \qquad \mathbb{E}\left[\left(N_{l,m}^{(n)}\right)^r \ \middle| \ L_m^{(n)}=s, K_m^{(n)}=k\right]$$

$$= \sum_{y=1}^{k} S(r,y)y!\binom{k}{y}\mathbb{P}\left[\boldsymbol{S}_{\boldsymbol{c}^{(y)}}^* = l\boldsymbol{1}_y \,\middle|\, A_{n,m}(j,\boldsymbol{n},s,k)\right]$$

$$= \sum_{y=1}^{k} S(r,y)\frac{s!}{(s-yl)!} \frac{[\sigma(1-\sigma)_{l-1}]^y}{(l!)^y} \frac{\mathscr{C}(s-yl,k-y;\sigma)}{\mathscr{C}(s,k;\sigma)}$$

where we set $\mathbf{1}_y := (1, \ldots, 1)$. Hence, (49) combined with (40) leads to the following expression

(50)

$$\mathbb{E}\left[\left(N_{l,m}^{(n)}\right)^r\right] = \sum_{s=0}^{m}\sum_{k=0}^{s} \frac{V_{n+m,j+k}}{V_{n,j}} \binom{m}{s}(n-j\sigma)_{m-s} \sum_{y=1}^{r\wedge k} S(r,y) \frac{s!}{(s-yl)!}$$

$$\times \frac{[\sigma(1-\sigma)_{l-1}]^y}{(l!)^y} \frac{\mathscr{C}(s-yl, k-y; \sigma)}{\sigma^k}.$$

In (50) note that the sum over the index y, for $y = 1, \ldots, k$, is equivalent to a sum over the index y for $y = 1, \ldots, r$. Indeed, if $k > r$ then the sum over the index y is non–null for $y = 1, \ldots, r$ because $S(r,y) = 0$ for any $y = r+1, \ldots, k$; on the other hand, if $k < r$ then the sum over the index y in non–null for $y = 1, \ldots, k$ because $\mathscr{C}(s-yl, k-y; \sigma) = 0$ for any $y = k+1, \ldots, r$. Basing on this, one can rewrite the expected value above as

$$\mathbb{E}\left[\left(N_{l,m}^{(n)}\right)^r\right] = \sum_{y=1}^{r} S(r,y) \frac{[(1-\sigma)_{l-1}]^y}{(l!)^y} \sum_{s=yl}^{m} \binom{m}{s}(n-j\sigma)_{m-s} \frac{s!}{(s-yl)!}$$

$$\times \sum_{k=y}^{s} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(s-yl, k-y; \sigma)}{\sigma^{k-y}}$$

$$= \sum_{y=1}^{r} S(r,y) \frac{[(1-\sigma)_{l-1}]^y}{(l!)^y} \sum_{s=0}^{m-yl} \binom{m}{s+yl}(n-j\sigma)_{m-s-yl} \frac{(s+yl)!}{(s)!}$$

$$\times \sum_{k=0}^{s+yl-y} \sigma^{-k} \frac{V_{n+m,j+k+y}}{V_{n,j}} \mathscr{C}(s,k;\sigma)$$

$$= \sum_{y=1}^{r} S(r,y) \frac{[(1-\sigma)_{l-1}]^y}{(l!)^y} \frac{m!}{(m-yl)!} \sum_{k=0}^{m-yl} \sigma^{-k} \frac{V_{n+m,j+k+y}}{V_{n,j}}$$

$$\times \sum_{s=k}^{m-yl} \binom{m-yl}{s}(n-j\sigma)_{m-yl-s} \mathscr{C}(s,k;\sigma)$$

$$= \sum_{y=1}^{r} S(r,y)\,[(1-\sigma)_{l-1}]^y\, \frac{m!}{(l!)^y\,(m-yl)!}$$

$$\times \sum_{k=0}^{m-yl} \sigma^{-k} \frac{V_{n+m,j+k+y}}{V_{n,j}} \frac{\mathscr{C}(m-yl,k;\sigma,-n+j\sigma)}{\sigma^k}.$$

The proof of (16) is, thus, completed by using the relation between the r–th moment with the r–th factorial moment. □

5.4. *Proof of Theorem 3.* The proof follows from conditional independence between the random variables $\boldsymbol{S}_{\boldsymbol{q}^{(x)}}$ and $\boldsymbol{S}_{\boldsymbol{r}^{(y)}}$, given $(K_n, \boldsymbol{N}_n, L_m^{(n)}, K_m^{(n)})$, as stated in Theorem 1. Indeed, according to the definition of the random variable $M_{l,m}$, for any $r \geq 1$ one can write

$$(51) \qquad \mathbb{E}\left[\left(M_{l,m}^{(n)}\right)^r\right]$$

$$= \sum_{t=0}^{r} \binom{r}{t} \sum_{s=0}^{m} \sum_{k=0}^{s} \alpha_t(l)\,\beta_{r-t}(l)\,\mathbb{P}[L_m^{(n)}=s,\,K_m^{(n)}=k \,|\, A_n(j,\boldsymbol{n})]$$

where

$$\alpha_t(l) := \mathbb{E}\left[\left(O_{l,m}^{(n)}\right)^t \,\middle|\, L_m^{(n)}=s,\,K_m^{(n)}=k\right]$$

$$= \sum_{x=1}^{j\wedge t} x! S(t,x) \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \mathbb{P}\left[\boldsymbol{S}_{\boldsymbol{c}^{(x)}} = l\boldsymbol{1}_x - \boldsymbol{n}_{\boldsymbol{c}^{(x)}} \,|\, A_{n,m}(j,\boldsymbol{n},s,k)\right].$$

and

$$\beta_{r-t}(l) := \mathbb{E}\left[\left(N_{l,m}^{(n)}\right)^{r-t} \,\middle|\, L_m^{(n)}=s,\,K_m^{(n)}=k\right]$$

$$= \sum_{y=1}^{k\wedge(r-t)} y! S(r-t,y) \sum_{\boldsymbol{c}^{(y)} \in \mathcal{C}_{k,y}} \mathbb{P}\left[\boldsymbol{S}_{\boldsymbol{c}^{(y)}}^* = l\boldsymbol{1}_y \,|\, A_{n,m}(j,\boldsymbol{n},s,k)\right].$$

In particular, by combining (51) with (48) and (49) one has

$$\mathbb{E}\left[\left(M_{l,m}^{(n)}\right)^r\right]$$

$$= \sum_{t=0}^{r} \binom{r}{t} \sum_{s=0}^{m} \sum_{k=0}^{s} \mathbb{P}[L_m^{(n)}=s,\,K_m^{(n)}=k \,|\, A_n(j,\boldsymbol{n})]$$

$$\times \sum_{x=1}^{t} S(t,x)x! \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \frac{(m-s)!}{(m-s-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{c_i}-\sigma)_{l-n_{c_i}}}{(l-n_{c_i})!}$$

$$\times \frac{(n - |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}| - (j - x)\sigma)_{m-s-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|}}{(n - j\sigma)_{m-s}}$$

$$\times \sum_{y=1}^{r-t} S(r - t, y) \frac{s!}{(s - yl)!} \frac{[\sigma(1 - \sigma)_{l-1}]^y}{(l!)^y} \frac{\mathscr{C}(s - yl, k - y; \sigma)}{\sigma^k}.$$

Using the same arguments applied in the last part of Theorems 2 and 1, the expression (51) combined with (40) leads to the following

$$
(52) \qquad \mathbb{E}\left[\left(M_{l,m}^{(n)}\right)^r\right]
$$

$$
= \sum_{t=0}^{r} \binom{r}{t} \sum_{x=1}^{t} S(t, x) \sum_{y=1}^{r-t} S(r - t, y) x! \frac{[(1 - \sigma)_{l-1}]^y}{(l!)^y}
$$

$$
\times \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \frac{m!}{(m - xl - yl + |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{c_i} - \sigma)_{l-c_i}}{(l - n_{c_i})!}
$$

$$
\times \sum_{k=0}^{m-xl-yl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|} \frac{V_{n+m,j+k+y}}{V_{n,j}}
$$

$$
\times \frac{\mathscr{C}(m - xl - yl + |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|, k; \sigma, -n + |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}| + (j - x)\sigma)}{\sigma^k}.
$$

The expression in (52) can be further simplified by applying well–known properties of the Stirling numbers of the second kind. In particular, according to the identity

$$
S(r, y + x) \binom{y + x}{x} = \sum_{t=x}^{r-y} \binom{r}{t} S(t, x) S(r - t, y)
$$

(see [4, Chapter 2]) one can write

$$
\mathbb{E}\left[\left(M_{l,m}^{(n)}\right)^r\right]
$$

$$
= \sum_{x=0}^{r} \sum_{y=0}^{r-x} S(r, y + x) \binom{y + x}{x} x! \frac{[(1 - \sigma)_{l-1}]^y}{(l!)^y}
$$

$$
\times \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \frac{m!}{(m - xl - yl + |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{c_i} - \sigma)_{l-c_i}}{(l - n_{c_i})!}
$$

$$\times \sum_{k=0}^{m-xl-yl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|} \frac{V_{n+m,j+k+y}}{V_{n,j}}$$

$$\times \frac{\mathscr{C}(m-xl-yl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|,k;\sigma,-n+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|+(j-x)\sigma)}{\sigma^k}$$

$$= \sum_{y=0}^{r} S(r,y) \sum_{x=0}^{y} \binom{y}{x} x! \frac{[(1-\sigma)_{l-1}]^{y-x}}{(l!)^{y-x}}$$

$$\times \sum_{\boldsymbol{c}^{(x)}\in\mathcal{C}_{j,x}} \frac{m!}{(m-yl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{c_i}-\sigma)_{l-c_i}}{(l-n_{c_i})!}$$

$$\times \sum_{k=0}^{m-yl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|} \frac{V_{n+m,j+k+y-x}}{V_{n,j}}$$

$$\times \frac{\mathscr{C}(m-yl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|,k;\sigma,-n+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|+(j-x)\sigma)}{\sigma^k}.$$

The proof of (20) is, thus, completed by using the relation between the r–th moment with the r–th factorial moment. $\qquad\square$

5.5. *Proofs for the Dirichlet process.*

5.5.1. *Proof of Proposition 2 and 3.* The distribution of $M_{l,n}$ is determined by its factorial moments as

$$\mathbb{P}[M_{l,n}=m_l] = \frac{\mathbb{1}_{\{1,\dots,n\}}(m_l l)}{m_l!} \sum_{k=m_l}^{n} \frac{(-1)^{k-m_l}}{(k-m_l)!} \mathbb{E}\left[(M_{l,n})_{[k]}\right]$$

$$= \frac{n!}{m_l!(\theta)_n} \sum_{k=m_l}^{[n/l]} \frac{(-1)^{k-m_l}}{(k-m_l)!} \frac{(\theta)_{[n-kl]}}{l^k(n-kl)!}$$

and, from this, (26) easily follows. On the other hand, Proposition 3 is a trivial consequence of (23) and (24). $\qquad\square$

5.6. *Proofs for the Pitman model.*

5.6.1. *Proof of Proposition 4.* This again follows from the application of the sieve formula, as discussed in the proof of Proposition 2. $\qquad\square$

5.6.2. *Proof of Proposition 5.* From Theorem 1 one finds that

$$
\mathbb{E}\left[\left(O_{l,m}^{(n)}\right)_{[r]}\right] = \frac{r!\,m!}{(\theta+n)_m} \sum_{\boldsymbol{c}^{(r)} \in \mathcal{C}_{j,r}} \frac{1}{(m-rl+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|)!} \prod_{i=1}^{r} \frac{(n_{c_i}-\sigma)_{l-n_{c_i}}}{(l-n_{c_i})!}
$$

$$
\times \sum_{k=0}^{m-rl+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|} \left(\frac{\theta}{\sigma}+j\right)_k \mathscr{C}(m-rl+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|, k; \sigma, -n+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|+(j-r)\sigma).
$$

By definition

$$
\sum_{k=0}^{n} \mathscr{C}(n,k;\sigma,\gamma)\,(t)_k = (\sigma t - \gamma)_n
$$

and this entails

$$
\mathbb{E}\left[\left(O_{l,m}^{(n)}\right)_{[r]}\right] = \frac{r!\,m!}{(\theta+n)_m} \sum_{\boldsymbol{c}^{(r)} \in \mathcal{C}_{j,r}} \frac{1}{(m-rl+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|)!} \prod_{i=1}^{r} \frac{(n_{c_i}-\sigma)_{l-n_{c_i}}}{(l-n_{c_i})!}
$$

$$
\times \ (\theta+n-|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|+r\sigma)_{m-rl+|\boldsymbol{n}_{\boldsymbol{c}^{(r)}}|}.
$$

The usual application of the sieve formula yields (27).   □

5.6.3. *Proof of Proposition 6.* Follows from Theorem 2, along the same lines as in the proof of Proposition 3.4.   □

5.6.4. *Proof of Theorem 4.* Our strategy will consist in examining the asymptotic beahviour of the r–th moments of $N_{l,m}^{(n)}$ and of $M_{l,m}^{(n)}$, for any $r \geq 1$, as m increases. To this end it is worth referring to the following decomposition that implicitly follows from the proof of Theorem 3. Indeed it can be seen that

$$
\mathbb{E}[(M_{l,m}^{(n)})^r] = \mathbb{E}[(O_{l,m}^{(n)})^r] + \mathbb{E}[(N_{l,m}^{(n)})^r] + \sum_{i=1}^{r-1} \binom{r}{i} \mathcal{B}^{(i)}(\sigma, n, j, \boldsymbol{n}, m),
$$

where

$$
\mathbb{E}[(O_{l,m}^{(n)})^r] = \frac{m!}{(\theta+n)_m} \sum_{x=1}^{j \wedge r} x!\,S(r,x) \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \prod_{r=1}^{x} \frac{(n_{c_r}-\sigma)_{l-n_{c_r}}}{(l-n_{c_r})!}
$$

$$
\times \ \frac{(\theta+n-|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|+x\sigma)_{m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|}}{(m-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!}
$$

$$\mathbb{E}[(N_{l,m}^{(n)})^r] = \frac{m!}{(\theta+n)_m} \sum_{y=1}^{[m/l]\wedge r} S(r,y) \frac{\sigma^y[(1-\sigma)_{l-1}]^y}{(l!)^y} \left(j+\frac{\theta}{\sigma}\right)_y$$

$$\times \frac{(\theta+n+y\sigma)_{m-ly}}{(m-yl)!}$$

$$\mathcal{B}^{(i)}(\sigma,n,j,\boldsymbol{n},m) = \frac{m!}{(\theta+n)_m} \sum_{x=1}^{j\wedge i} x! S(i,x) \sum_{\boldsymbol{c}^{(x)}\in\mathcal{C}_{j,x}} \prod_{r=1}^{x} \frac{(n_{c_r}-\sigma)_{l-n_{c_r}}}{(l-n_{c_r})!}$$

$$\times \sum_{y=1}^{m\wedge(r-i)} S(r-i,y) \frac{\sigma^y[(1-\sigma)_{l-1}]^y}{(l!)^y} \left(j+\frac{\theta}{\sigma}\right)_y$$

$$\times \frac{(\theta+n-|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|+\sigma x)_{m-yl-xl+n_{c_i}+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|}}{(m-yl-xl+|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|)!}.$$

By virtue of Stirling's approximation formula one has, as $m \to +\infty$,

$$\mathbb{E}[(O_{l,m}^{(n)})^r] \sim m^{-\theta-n+1}\,\Gamma(\theta+n) \sum_{x=1}^{j\wedge r} \sum_{\boldsymbol{c}^{(x)}\in\mathcal{C}_{j,x}} x!\,S(r,x)$$

$$\times \frac{m^{\theta+n-|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|-1+x\sigma}}{\Gamma(\theta+n-|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|+x\sigma)} \prod_{t=1}^{x} \frac{(n_{c_t}-\sigma)_{l-n_{c_t}}}{(l-n_{c_t})!}$$

where $a_n \sim b_n$ means that $a_n/b_n \to 1$, as $n \to \infty$. The term that asymptotically dominates the right–hand side of the asymptotic equivalence above, as $m \to \infty$, can be bounded by

$$m^{(j\wedge r)\sigma-|\boldsymbol{n}_{\boldsymbol{c}^{(j\wedge r)}}|} \frac{\Gamma(\theta+n)\,(j\wedge r)!\,S(r,(j\wedge r))}{\Gamma(\theta+n-|\boldsymbol{n}_{\boldsymbol{c}^{(j\wedge r)}}|+(j\wedge r)\sigma)} \prod_{t=1}^{(j\wedge r)} \frac{(n_{c_t}-\sigma)_{l-n_{c_t}}}{(l-n_{c_t})!}.$$

Since $|\boldsymbol{n}_{\boldsymbol{c}^{(j\wedge r)}}| \geq 1$, one has

$$\lim_{m\to\infty} \frac{\mathbb{E}[(O_{l,m}^{(n)})^r]}{m^{r\sigma}} = 0.$$

In a similar fashion note that, as $m \to \infty$, the following asymptotic equivalence holds true

$$\mathbb{E}[(N_{l,m}^{(n)})^r] \sim \Gamma(\theta+n)\,m^{1-\theta-n} \sum_{y=1}^{r} S(r,y) \frac{\sigma^y[(1-\sigma)_{l-1}]^y}{(l!)^y}$$

$$\times \; \frac{\left(j + \frac{\theta}{\sigma}\right)_y}{\Gamma(\theta + n + y\sigma)} \; m^{\theta + n + y\sigma - 1}$$

which, in turn, yields

$$\lim_{m \to +\infty} \frac{\mathbb{E}[(N_{l,m}^{(n)})^r]}{m^{r\sigma}} = \left(\frac{\sigma(1-\sigma)_{l-1}}{l!}\right)^r \frac{\Gamma(\theta + n)\left(j + \frac{\theta}{\sigma}\right)_r}{\Gamma(\theta + n + r\sigma)}.$$

Finally, still as $m \to \infty$,

$$\mathcal{B}^i(\sigma, n, j, \boldsymbol{n}, m) \sim \frac{\Gamma(\theta + n)}{m^{\theta + n - 1}} \sum_{x=1}^{j \wedge i} x! S(i, x) \sum_{\boldsymbol{c}^{(x)} \in \mathcal{C}_{j,x}} \prod_{t=1}^{x} \frac{(n_{c_t} - \sigma)_{l - n_{c_t}}}{(l - n_{c_t})!}$$

$$\times \sum_{y=1}^{r-i} S(r - i, y) \frac{\sigma^y [(1-\sigma)_{l-1}]^y}{(l!)^y}$$

$$\times \; \frac{\left(j + \frac{\theta}{\sigma}\right)_y}{\Gamma(\theta + n - |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}| + x\sigma)} \; m^{\theta + n - 1 + x\sigma - |\boldsymbol{n}_{\boldsymbol{c}^{(x)}}|}$$

and, since $|\boldsymbol{n}_{\boldsymbol{c}^{(x)}}| \geq 1$ for any $x = 1, \dots, (j \wedge i)$, one has

$$\lim_{m \to \infty} \frac{1}{m^{r\sigma}} \, \mathcal{B}^i(\sigma, n, j, \boldsymbol{n}, m) = 0$$

for any $i = 1, \dots, r - 1$. These limiting relations plainly lead to conclude that

$$\lim_{m \to +\infty} \mathbb{E}\left[m^{-r\sigma} \left(M_{l,m}^{(n)}\right)^r\right] = \left(\frac{\sigma(1-\sigma)_{l-1}}{l!}\right)^r \frac{\Gamma(\theta + n)\left(j + \frac{\theta}{\sigma}\right)_r}{\Gamma(\theta + n + r\sigma)}$$

$$= \left(\frac{\sigma(1-\sigma)_{l-1}}{l!}\right)^r \mathbb{E}[Z_{n,j}^r].$$

According to [8, Proposition 2], the distribution of the random variable $Z_{n,j}$ is uniquely characterized by the moment sequence $(\mathbb{E}[(Z_{n,j})^r])_{r \geq 1}$. Similar arguments lead to determine the limiting distribution of the random variable $N_{l,m}^{(n)}/m_\sigma$, as $m \to +\infty$. □

5.7. *Proofs for the Gnedin model.*

5.7.1. *Proof of Propositions 8 and 9.* The proof of (34) follows from (11) and (9), after noting that $\mathscr{C}(n, k; -1) = (-1)^k n! (n-1)! / [k! (k-1)! (n-k)!]$. As for the determination of the distributions of $O_{l,m}^{(n)}$ and $N_{l,m}^{(n)}$ one uses the fact that $\mathscr{C}(n, k; -1, \gamma) = (-1)^k \binom{n - \gamma - 1}{n - k} n! / k!$ along with the results stated in Theorems 1 and 2. □

References.

[1] ARRATIA, R., BARBOUR, A.D. and TAVARÉ, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2**, 519–535.

[2] ARRATIA, R., BARBOUR, A.D. and TAVARÉ, S. (2003). *Logarithmic combinatorial structures: a probabilistic approach.* EMS Monograph in Mathematics.

[3] BARBOUR, A.D. (1992). Refined approximations for the Ewens sampling formula. *Random Structure Algorithms* **3**, 267–276.

[4] CHARALAMBIDES, C.A. (2005). *Combinatorial methods in discrete distributions.* Hoboken, NJ: Wiley.

[5] DURDEN, C. and DONG, Q. (2009). RICHEST - a web server for richness estimation in biological data. *Bioinformation* **3**, 296–298.

[6] EWENS, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112.

[7] EWENS, W.J. and TAVARÉ, S. (1998). The Ewens Sampling Formula. In *Encyclopedia of Statistical Science*, Vol. 2 update. (Eds. Kotz, S., Read, C.B. and Banks, D.L.), pp. 230–234, Wiley, New York.

[8] FAVARO, S., LIJOI, A., MENA, R.H. and PRÜNSTER, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B* **71**, 993–1008.

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

[10] GNEDIN, A.V. and PITMAN, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **325**, 83–102.

[11] GNEDIN, A.V. (2010). A species sampling model with finitely many types. *Elect. Comm. in Probab.* **15**, 79–88.

[12] GRIFFITHS, R.C. and SPANÒ, D. (2007). Record indices and age-ordered frequencies in exchangeable Gibbs partitions. *Electron. J. Probab.* **12**,1101–1130.

[13] HO, M.W., JAMES, L.F. and LAU, J.W. (2007). Gibbs partitions (EPPF's) derived from a stable subordinator are Fox H and Meijer G transforms. *MatharXiv* preprint, arXiv:0708.0619v2.

[14] JAMES, L.F. (2010). Lamperti–type laws. *Ann. Appl. Probab.* **20**, 1303-1340.

[15] KINGMAN, J.F.C. (1978). The representation of partition structures. *J. London Math. Soc.* **18**, 374–380.

[16] KINGMAN, J.F.C. (1982). The coalescent. *Stochast. Process. Appl.* **13**, 235–248.

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

[18] LIJOI, A., MENA, R.H. AND PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 715-740.

[19] LIJOI A., MENA, R.H. AND PRÜNSTER, I. (2007b). A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinformatics*, **8**: 339.

[20] LIJOI, A., PRÜNSTER, I. and WALKER, S.G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18**, 1519–1547.

[21] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields.* **102**, 145–158.

[22] PITMAN, J. (2003). Poisson-Kingman partitions. *Science and Statistics: A Festschrift for Terry Speed* (D.R. Goldstein, ed.) *Lecture Notes Monograph Series* **40** 1–34. IMS, Beachwood, OH.

[23] PITMAN, J. (2006). *Combinatorial stochastic processes.* Ecole d'Eté de Probabilités

de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875, Springer, New York.

[24] QUACKENBUSH, J., CHO, J., LEE, D., LIANG, F., HOLT, I., KARAMYCHEVA, S., PARVIZI, B., PERTEA, G., SULTANA, R. and WHITE, J. (2000). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29**, 159–164.

[25] SCHWEINSBERG, J. (2010). The number of small blocks in exchangeable random partitions. *ALEA Lat. Am. J. Probab. Math. Stat.* **7**, 217–242.

[26] VALEN, E. (2009). Deciphering Transcriptional Regulation - Computational Approaches. *Ph.D. Thesis*, Bioinformatics Centre, University of Copenhagen.

N.		$\tau = 3,\ n = 1000$			$\tau = 4,\ n = 1000$			$\tau = 5,\ n = 1000$		
		$O_\tau^{(n)}$	$N_\tau^{(n)}$	$M_\tau^{(n)}$	$O_\tau^{(n)}$	$N_\tau^{(n)}$	$M_\tau^{(n)}$	$O_\tau^{(n)}$	$N_\tau^{(n)}$	$M_\tau^{(n)}$
1	est.	750	1010	1759	777	1014	1791	793	1016	1809
	true	767	991	1758	793	998	1791	803	999	1802
2	est.	739	1006	1744	765	1010	1775	781	1011	1792
	true	753	1005	1758	785	1006	1791	794	1008	1802
3	est.	730	1003	1733	755	1007	1762	770	1008	1779
	true	742	1016	1758	772	1019	1791	783	1019	1802
4	est.	765	1043	1807	789	1047	1836	804	1048	1852
	true	772	986	1758	800	991	1791	811	991	1802
5	est.	741	971	1712	771	976	1748	788	978	1766
	true	761	997	1758	788	1003	1791	797	1005	1802
6	est.	758	1027	1785	784	1031	1816	800	1033	1833
	true	770	988	1758	798	993	1791	809	993	1802
7	est.	739	997	1735	766	1002	1768	783	1003	1786
	true	758	1000	1758	787	1004	1791	796	1006	1802
8	est.	734	984	1719	763	989	1752	780	991	1770
	true	747	1011	1758	779	1012	1791	790	1012	1802
9	est.	729	969	1698	759	974	1733	777	975	1752
	true	747	1011	1758	779	1012	1791	789	1013	1802
10	est.	757	1020	1777	784	1025	1809	800	1026	1826
	true	774	984	1758	799	992	1791	807	995	1802

Table 1: *Cross–validation study based on sub–samples of size 1000 and prediction on the remaining $m = 1586$ data. The reported estimated and true quantities are the number of rare genes (i.e. with expression levels less than or equal to τ, for $\tau = 3, 4, 5$) among the "old" genes ($O_\tau^{(n)}$), the "new" genes ($N_\tau^{(n)}$) and all genes ($M_\tau^{(n)}$).*

m	$\tau = 3$ $n = 2586,\ j = 1825$			$\tau = 4$ $n = 2586,\ j = 1825$			$\tau = 5$ $n = 2586,\ j = 1825$		
	$\hat{O}_\tau^{(n)}$	$\hat{N}_\tau^{(n)}$	$\hat{M}_\tau^{(n)}$	$\hat{O}_\tau^{(n)}$	$\hat{N}_\tau^{(n)}$	$\hat{M}_\tau^{(n)}$	$\hat{O}_\tau^{(n)}$	$\hat{N}_\tau^{(n)}$	$\hat{M}_\tau^{(n)}$
250	1745	138	1882	1782	138	1920	1798	138	1935
500	1730	272	2002	1773	272	2045	1793	272	2064
750	1715	402	2117	1763	402	2165	1787	403	2189
1000	1700	529	2229	1753	530	2283	1780	530	2310

Table 2: *Estimates for an additional sample corresponding to $m \in \{250, 500, 750, 1000\}$ given the observed EST dataset of size $n = 2586$ with $j = 1825$ distinct genes: estimates for the number of rare genes (i.e. with expression levels less than or equal to τ, for $\tau = 3, 4, 5$) among the "old" genes ($O_\tau^{(n)}$), the "new" genes ($N_\tau^{(n)}$) and all genes ($M_\tau^{(n)}$).*

Stefano Favaro and Igor Prünster
Dipartimento di Statistica e Matematica Applicata
Università degli Studi di Torino
Corso Unione Sovietica 218/bis, I–10134 Torino, Italy
E-mail: stefano.favaro@unito.it; igor@econ.unito.it

Antonio Lijoi
Dipartimento di Economia Politica
e Metodi Quantitativi
Università degli Studi di Pavia
Via San Felice 5, 27100 Pavia, Italy
E-mail: lijoi@unipv.it