Posterior analysis of rare variants in Gibbs-type species sampling models

Journal of Multivariate Analysis 131 (2014) 79–98

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Posterior analysis of rare variants in Gibbs-type species sampling models

Oriana Cesari [a], Stefano Favaro [b,*,1], Bernardo Nipoti [b,1]

[a] Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy

[b] University of Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy

ARTICLE INFO

Article history:
Received 21 November 2012
Available online 26 June 2014

AMS 2000 subject classifications:
60G57
62G05
62F15

Keywords:
Bayesian nonparametric inference
Asymptotic credible intervals
Exchangeable random partition
Gibbs-type random probability measure
Index of diversity
Sampling formula
Species sampling problem
Rare variant
Two parameter Poisson–Dirichlet process

ABSTRACT

Species sampling problems have a long history in ecological and biological studies and a number of statistical issues, including the evaluation of species richness, are still to be addressed. In this paper, motivated by Bayesian nonparametric inference for species sampling problems, we consider the practically important and technically challenging issue of developing a comprehensive posterior analysis of the so-called rare variants, namely those species with frequency less than or equal to a given abundance threshold. In particular, by adopting a Gibbs-type prior, we provide an explicit expression for the posterior joint distribution of the frequency counts of the rare variants, and we investigate some of its statistical properties. The proposed results are illustrated by means of two novel applications to a benchmark genomic dataset.

1. Introduction

Suppose that statistical units drawn from a population are representative of different species. Their labels are denoted by $\hat{X}_i$ and their respective proportions in the population by $\tilde{p}_i$, for $i \geq 1$. Therefore, models for species sampling problems can be usefully embedded in the framework of discrete random probability measures, $\tilde{P} = \sum_{i\geq 1} \tilde{p}_i \delta_{\hat{X}_i}$, where δ_a denotes the point mass at a. Discrete random probability measures emerge as remarkable tools for theoretical and applied analysis in, e.g., population genetics, ecology, genomics, mathematical physics, machine learning. The most celebrated example of discrete random probability measure is the Dirichlet process introduced by Ferguson [14] and whose random masses $\tilde{p}_i$ are obtained either by normalizing the jumps of a Gamma completely random measure or by means of a stick-breaking procedure. This process has been also popularized under the name of (one parameter) Poisson–Dirichlet process and characterized in terms of the distribution of its ranked random masses by Kingman [22]. The reader is referred to Lijoi and Prünster [28] for an up-to-date review of classes of discrete random probability measures generalizing the Dirichlet process.

* Corresponding author.
E-mail addresses: oriana.cesari@carloalberto.org (O. Cesari), stefano.favaro@unito.it (S. Favaro), bernardo.nipoti@unito.it (B. Nipoti).
[1] Also affiliated to Collegio Carlo Alberto, Moncalieri, Italy.

http://dx.doi.org/10.1016/j.jmva.2014.06.017
0047-259X/

In this paper our attention will be focused on statistical issues related to species sampling problems: these will be addressed by a Bayesian nonparametric approach. We consider data from a population whose species composition is directed by a discrete random probability measure $\tilde{P}$ with distribution Π, i.e.

$$\begin{aligned} X_i \mid \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P} \quad i = 1, \ldots, n \\ \tilde{P} &\sim \Pi, \end{aligned} \tag{1}$$

for any $n \geq 1$. According to de Finetti's representation theorem, $(X_i)_{i\geq 1}$ is exchangeable and Π takes on the interpretation of a prior distribution over the composition of the population. Since $\tilde{P}$ is discrete, we expect ties in a sample $(X_1, \ldots, X_n)$ from $\tilde{P}$. Precisely we expect $K_n \leq n$ distinct observations, or species, with frequencies $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$ such that $\sum_{1\leq i\leq K_n} N_i = n$. Accordingly, the sample induces a random partition of $\{1, \ldots, n\}$, in the sense that any index $i \neq j$ belongs to the same partition set if and only if $X_i = X_j$. We denote by $p_j^{(n)}(n_1, \ldots, n_j)$ the function corresponding to the probability of any particular partition of $\{1, \ldots, n\}$ having $K_n = j$ blocks with frequencies $\mathbf{N}_n = (n_1, \ldots, n_j)$. This function is known as the exchangeable partition probability function (EPPF), a concept introduced in Pitman [34] as a development of earlier results in Kingman [23]. See Pitman [36] for a comprehensive account on exchangeable random partitions.

Under the framework (1), with $\tilde{P}$ being in the class of the Gibbs-type random probability measures by Pitman [35], Lijoi et al. [25] introduced a novel Bayesian nonparametric methodology for making inferences on quantities related to an additional unobserved sample $(X_{n+1}, \ldots, X_{n+m})$ from $\tilde{P}$, given an observed sample $(X_1, \ldots, X_n)$. A particularly important example is represented by the estimation of the number of new species that will be observed in the additional sample. See Lijoi et al. [29], Favaro et al. [12], Favaro et al. [11] and Bacallado et al. [1] for estimators of other features related to species richness under Gibbs-type priors. This class of priors stands out for both mathematical tractability and flexibility. Indeed, apart from the Dirichlet process, the class of the Gibbs-type random probability measures includes as special cases the two parameter Poisson–Dirichlet process, also known as Pitman–Yor process, and the normalized generalized Gamma process. We refer to Perman et al. [33], Pitman and Yor [37] and Ishwaran and James [18] for details on the two parameter Poisson–Dirichlet process, and to James [19], Pitman [35], Lijoi et al. [27] and James [20] for details on the normalized generalized Gamma process. Gibbs-type priors also stand out for being particularly suited in the context of inferential problems with a large unknown number of species, which typically occur in several genomic applications. See, e.g., Lijoi et al. [26], Guindani et al. [16] and De Blasi et al. [5].

Motivated by the goal of performing Bayesian nonparametric inference for species sampling problems, in this paper we develop a comprehensive posterior analysis of the so-called rare variants, namely the species with frequency less than or equal to a given abundance threshold τ. Ecological and biological literature have always devoted special attention to rare variants. In ecology, for instance, conservation of biodiversity represents 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. See, e.g., Usher [39] and Magurran [31] for detailed surveys on measurements of biodiversity, conservation of populations, commonness and rarity of species. On the other hand in genetics one is typically interested in the number of individuals with rare genes, the reasons being that rare genes of a specific type may be associated with a deleterious disease. See, e.g., Elandt-Johnson [7] and Laird and Lange [24] for a detailed account on the role of rare variants in genetics.

Under the statistical framework (1), with $\tilde{P}$ being a Dirichlet process, Joyce and Tavaré [21] first investigated the prior distribution of the rare variants, namely the joint distribution of the frequency counts of the rare variants induced by an initial sample $(X_1, \ldots, X_n)$ from $\tilde{P}$. In particular they mainly focused on the study of the asymptotic behavior, for a large sample size n, of such a prior distribution. In this paper we derive the prior distribution of the rare variants under the more general assumption of $\tilde{P}$ being a Gibbs-type random probability measure. Furthermore, following ideas set forth in Lijoi et al. [25], we derive and investigate the posterior distribution of the rare variants. Such a posterior distribution corresponds to the conditional joint distribution of the frequency counts of the rare variants induced by an additional sample $(X_{n+1}, \ldots, X_{n+m})$, given $(X_1, \ldots, X_n)$. Precisely, this is as a suitable convolution of: (i) the joint posterior distribution of the new rare variants that are generated from the additional sample and do not coincide with rare variants already detected in the initial sample; (ii) the joint posterior distribution of the old rare variants that arise by updating, via the additional sample, the rare variants already detected in the initial sample. Our distributional results are derived by generalizing some of the combinatorial techniques originally developed in Favaro et al. [12], where special cases of the results in this paper have been presented. After submitting the first version of this paper we learnt that the posterior distribution of the rare variants have been recently obtained independently, and by means of different techniques, in Cerquetti [2]. For additional distributional results on rare variants we refer to the M.Sc. Thesis of Cesari [3], from which the main contributions of the present papers are taken.

Our prior and posterior distributional results admit several applications, not necessarily related to the study of the rare variants. In this paper we focus on two representative applications, which will be illustrated under the assumption of a two parameter Poisson–Dirichlet prior. Firstly, we devise a novel methodology to approximately quantify the uncertainty of a Bayesian nonparametric estimator for the number of rare species. This estimator has been recently introduced in Favaro et al. [12] and the problem of evaluating its accuracy is of great importance in several applied contexts. To this end, we

introduce approximate credible intervals that exploit the knowledge of the asymptotic behavior of the posterior distribution of rare variants. Secondly, we study, both a priori and a posteriori, the correlation between variants of different order. Explicit expressions for the correlations can be obtained by a direct application of our results on the prior and posterior distributions of number of rare variants. Our analysis of the prior correlation aims at gaining further insight on the role of the parameters characterizing the two parameter Poisson–Dirichlet process. Both applications are illustrated by means of the analysis of a well-known benchmark genomic dataset.

The paper is structured as follows. In Section 2 we recall the definition of Gibbs-type exchangeable random partition and we present some preliminary results on the prior distribution of rare variants. In Section 3 we derive and investigate the posterior distribution of rare variants under the general framework of Gibbs-type prior and in the special case of the two parameter Poisson–Dirichlet prior. Section 4 contains two representative applications of our prior and posterior distributions for rare variants. Proofs are deferred to the Appendix.

2. Gibbs-type exchangeable random partitions

In this section we review some sampling properties of Gibbs-type random probability measures. See the monograph by Pitman [36] and references therein for a comprehensive account on these sampling properties. For any $x > 0$ and any positive integer n, throughout the paper we use $(x)_{(n)}$ and $(x)_{[n]}$ to denote the rising factorial and falling factorial, respectively. Moreover, for any $\alpha \in (0, 1)$, let f_α be the density function of a positive α-stable random variable and let $S_{\alpha,c}$, for any $c > -1$, be a random variable with density function

$$f_{S_{\alpha,c}}(y) = \frac{\Gamma(c\alpha + 1)}{\alpha\Gamma(c + 1)} y^{c-1-1/\alpha} f_\alpha\left(y^{-1/\alpha}\right). \tag{2}$$

The random variable $S_{\alpha,c}^{-1/\alpha}$ is the so-called polynomially tilted positive α-stable random variable. See, e.g., Pitman [36] for a detailed account. See also Devroye [6] for exact sampling methods for $S_{\alpha,c}$. Finally, we denote by Z_a a random variable distributed according to a Poisson distribution with parameter a.

Gnedin and Pitman [15] characterized the EPPF induced by a Gibbs-type prior in terms of a distribution with a suitable product form, a feature which is crucial for guaranteeing mathematical tractability. Specifically, they showed that a sample $(X_1, \ldots, X_n)$ from a Gibbs-type random probability measure induces an exchangeable random partition with an EPPF of the form

$$p_j^{(n)}(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1 - \alpha)_{(n_i - 1)}, \tag{3}$$

for any $\alpha < 1$ and any collection of nonnegative weights $(V_{n,j})_{j \le n, n \ge 1}$ which satisfy the recursion $V_{n,j} = V_{n+1,j+1} + (n - j\alpha)V_{n+1,j}$, with the initial condition $V_{1,1} = 1$. A random partition distributed according to (3) is termed Gibbs-type exchangeable random partition. From (3), the distribution of the number K_n of blocks in a Gibbs-type exchangeable random partition corresponds to

$$\mathbb{P}[K_n = j] = V_{n,j} \frac{\mathscr{C}(n, j; \alpha)}{\alpha^j},$$

with $\mathscr{C}(n, j; \alpha) = (j!)^{-1} \sum_{i=0}^{j} \binom{j}{i} (-1)^i (-i\alpha)_{(n)}$ being the generalized factorial coefficient. See Charalambides [4] for details. In Example 1 we recall the Ewens sampling model and the Ewens–Pitman sampling model, which are two noteworthy examples of Gibbs-type exchangeable random partitions introduced in Ewens [8] and Pitman [34]. Precisely, they correspond to the exchangeable random partitions induced by a sample $(X_1, \ldots, X_n)$ from the Dirichlet process and the two parameter Poisson–Dirichlet process, respectively.

For any fixed $\alpha < 1$, the backward recursion of the weights $V_{n,j}$'s cannot be solved in a unique way. The solutions form a convex set where each element is the distribution of an exchangeable random partition. Theorem 12 in Gnedin and Pitman [15] describes the extreme points of such a convex set. Let

$$c_n(\alpha) = \begin{cases} 1 & \text{if } \alpha \in (-\infty, 0) \\ \log(n) & \text{if } \alpha = 0 \\ n^\alpha & \text{if } \alpha \in (0, 1), \end{cases}$$

for any $n \ge 1$. Then, for every Gibbs-type exchangeable random partition there exists a positive and almost surely finite random variable W_α such that

$$\frac{K_n}{c_n(\alpha)} \xrightarrow{\text{a.s.}} W_\alpha,$$

as $n \to +\infty$. A Gibbs-type exchangeable random partition is a unique mixture over $\varkappa$ of extreme exchangeable random partitions for which $W_\alpha = \varkappa$ almost surely. For $\alpha \in (-\infty, 0)$ the extremes are Ewens–Pitman sampling models with parameter $(\alpha, -\alpha\varkappa)$; for $\alpha = 0$ the extremes are Ewens sampling models with parameter $\varkappa \ge 0$; for $\alpha \in (0, 1)$ the Ewens–Pitman sampling models are not extremes. We refer to Section 6.1 in Pitman [35] for details on W_α.

Example 1. For any $\alpha \in [0, 1)$ and $\theta > -\alpha$ the Ewens–Pitman sampling model is a Gibbs-type exchangeable random partition with nonnegative weights of the form

$$V_{n,j} = \frac{\prod_{i=0}^{j-1}(\theta + i\alpha)}{(\theta)_{(n)}}. \tag{4}$$

The Ewens sampling model with parameter $\theta > 0$ corresponds to the special case $\alpha = 0$. Moreover,

$$\frac{K_n}{c_n(\alpha)} \xrightarrow{\text{a.s.}} \begin{cases} S_{\alpha,\theta/\alpha} & \text{if } \alpha \in (0, 1) \\ \theta & \text{if } \alpha = 0 \end{cases} \tag{5}$$

as $n \to +\infty$, where $S_{\alpha,\theta/\alpha}$ is a random variable with density function (2). See Pitman [36] for details on the weight (4) and on the limiting behavior of K_n in (11).

If $M_{l,n}$ is the number of species with frequency l in a sample $(X_1, \ldots, X_n)$ from a Gibbs-type random probability measure, then a change of variables in (3) yields the distribution of $\mathbf{M}_n = (M_{1,n}, \ldots, M_{n,n})$. Specifically, the Gibbs-type sampling formula determines the distribution of $\mathbf{M}_n$ and it corresponds to

$$p_n(m_1, \ldots, m_n) = n! V_{n,j} \prod_{i=1}^{n} \left(\frac{(1-\alpha)_{(i-1)}}{i!} \right)^{m_i} \frac{1}{m_i!} \tag{6}$$

with $(m_1, \ldots, m_n) \in \{0, 1, \ldots, n\}^n$ such that $\sum_{i=1}^n i m_i = n$ and $\sum_{i=1}^n m_i = j$. The Ewens–Pitman sampling formula, introduced by Pitman [34], is recovered as a special case of (6) by substituting $V_{n,j}$ with the expression in (4), i.e.

$$p_n(m_1, \ldots, m_n) = n! \frac{\prod_{i=0}^{j-1}(\theta + i\alpha)}{(\theta)_{(n)}} \prod_{i=1}^{n} \left(\frac{(1-\alpha)_{(i-1)}}{i!} \right)^{m_i} \frac{1}{m_i!}. \tag{7}$$

Of course the celebrated Ewens sampling formula, introduced in the seminal paper by Ewens [8], is a special case of (7) and it can be recovered by letting $\alpha \to 0$. We refer to Pitman [36] for a review on the Ewens–Pitman sampling formula and to Ewens and Tavaré [10] for a review on the Ewens sampling formula.

Let q be a positive integer and let $\mathbf{l} = (l_1, \ldots, l_q)$ be distinct positive integers. Moreover, let $\mathbf{r} = (r_1, \ldots, r_q)$ be a vector of positive integers. In the next theorem we derive the mixed falling factorial moment of order $\mathbf{r}$ of the random variable

$$\mathbf{M}_{\mathbf{l},n} = (M_{l_1,n}, M_{l_2,n}, \ldots, M_{l_q,n}).$$

Our result generalizes Theorem 1 in Favaro et al. [12], where the falling factorial moment of the random variable $M_{l,n}$ was derived. See also Ewens and Tavaré [10] for some moment formulae of $M_{l,n}$ under the framework of the Ewens sampling formula. Henceforth we agree that, for any positive integer n, $\sum_{i=0}^{-n} \equiv 0$.

Theorem 1. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a Gibbs-type prior. Then, for any $1 \leq q \leq n$, $r_i \geq 1$ and $1 \leq l_i \neq l_j \leq n$ with $i \neq j$*

$$\mathbb{E}\left[\prod_{i=1}^{q} (M_{l_i,n})_{[r_i]} \right] = H_\alpha(q, n, \mathbf{l}, \mathbf{r}) \sum_{j=0}^{n-\sum_{i=1}^q l_i r_i} V_{n,j+\sum_{i=1}^q r_i} \frac{\mathscr{C}\left(n - \sum_{i=1}^q l_i r_i, j; \alpha \right)}{\alpha^j}, \tag{8}$$

where

$$H_\alpha(q, n, \mathbf{l}, \mathbf{r}) = \frac{n!}{\left(n - \sum_{i=1}^q l_i r_i \right)!} \mathbb{1}_{\{0,1,\ldots,n\}}\left(n - \sum_{i=1}^q l_i r_i \right) \times \prod_{i=1}^{q} \left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!} \right)^{r_i}.$$

We recall that the falling factorial moment (8) characterizes the distribution of the random variable $\mathbf{M}_{\mathbf{l},n}$. Specifically, by means of standard arguments involving probability generating functions, (8) leads to an explicit expression for the distribution of the random variable $\mathbf{M}_{\mathbf{l},n}$. If $q = \tau < n$ and $l_i = i$ for $i = 1, \ldots, \tau$, then (8) provides the distribution of the rare variants, namely the joint distribution of the frequency counts $(M_{1,n}, \ldots, M_{\tau,n})$. In the next corollary we present a collection of results which can be derived by Theorem 1 under the assumption of a two parameter Poisson–Dirichlet prior. In Section 4 these results will be applied to a benchmark genomic dataset.

Corollary 1. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a two parameter Poisson–Dirichlet prior. Then, for any $\alpha \in [0, 1)$ and $\theta > -\alpha$ one has*

(i) *for any $1 \leq \tau \leq n$ and $r_i \geq 1$*

$$\mathbb{E}\left[\prod_{i=1}^{\tau}(M_{i,n})_{[r_i]}\right] = H_{\alpha}^{(\tau)}(n, \mathbf{r})\, \alpha^{\sum_{i=1}^{\tau} r_i} \left(\frac{\theta}{\alpha}\right)_{(\sum_{i=1}^{\tau} r_i)} \frac{\left(\theta + \alpha \sum_{i=1}^{\tau} r_i\right)_{(n-\sum_{i=1}^{\tau} i r_i)}}{(\theta)_{(n)}}; \tag{9}$$

(ii) *for any $1 \leq l_1 \neq l_2 \leq n$*

$$\mathbb{E}\left[M_{l_1,n} M_{l_2,n}\right] = H_{\alpha}(2, n, (l_1, l_2), (1, 1))\, \alpha^2 \left(\frac{\theta}{\alpha}\right)\left(\frac{\theta}{\alpha} + 1\right) \frac{(\theta + 2\alpha)_{(n-l_1-l_2)}}{(\theta)_n}; \tag{10}$$

(iii) *for any $\tau < n$ and $\leq l_i \neq l_j < n$ with $i \neq j$, as $n \to +\infty$*

$$\frac{1}{c_n(\alpha)} \sum_{i=1}^{\tau} M_{l_i,n} \xrightarrow{w} \begin{cases} \left(\sum_{i=1}^{\tau} \dfrac{\alpha(1-\alpha)_{(l_i-1)}}{l_i!}\right) S_{\alpha,\theta/\alpha} & \text{if } \alpha \in (0, 1) \\ \sum_{i=1}^{\tau} Z_{\theta/l_i} & \text{if } \alpha = 0. \end{cases} \tag{11}$$

Besides providing the distribution of the rare variants under a Gibbs-type prior, Theorem 1 is also a fundamental tool for deriving the distribution of the so-called sampling indexes of diversity. Sampling indexes of diversity are quantitative measures expressed as linear combinations of the $M_{i,n}$'s. They reflect the species richness of a population, and simultaneously take into account how evenly the individuals are distributed among those species. The value of a diversity index increases both when the number of types increases and when evenness increases. See Magurran [30], Magurran [31] and Magurran [32] for detailed accounts of indexes of diversity and their sampling versions. A prototype of the sampling index of diversity has the form

$$D = \sum_{i=1}^{n} c(i) M_{i,n}, \tag{12}$$

where the $c(i)$'s are deterministic weights. Noteworthy examples of sampling indexes of diversity are the sampling version of the Simpson index, which corresponds to the choice $c(i) = i^2/n^2$, and the sampling version of the Shannon entropy, which corresponds to the choice $c(i) = -(i/n)\log(i/n)$. In population genetics these two sampling indexes are known as the Watterson and Ewens tests of neutrality, and they are used as a good statistic for testing departures from selective neutrality in the direction of heterozygote advantage or disadvantage. See the monograph by Ewens [9] and references therein.

Intuitively, the problem of determining the distribution of the random variable D reduces to Theorem 1. Indeed the falling factorial moment of order r of D can be written in terms of the factorial moment 1. Specifically, by means of standard combinatorial manipulations of the falling factorials one has

$$\begin{aligned} \mathbb{E}[(D)_{[r]}] = & \sum_{(r_1,\ldots,r_n)\in\mathcal{D}_{n,r}} \binom{r}{r_1, \ldots, r_n} \times \sum_{v_1=0}^{r_1} (-1)^{v_1 - r_1} \mathscr{C}(r_1, v_1; c(1)) \sum_{v_2=0}^{r_2} (-1)^{v_2 - r_2} \mathscr{C}(r_2, v_2; c(2)) \\ & \times \cdots \times \sum_{v_n=0}^{r_n} (-1)^{v_n - r_n} \mathscr{C}(r_n, v_n; c(n)) \mathbb{E}\left[\prod_{i=1}^{n} (M_{i,n})_{[v_i]}\right], \end{aligned} \tag{13}$$

where $\mathcal{D}_{n,r} = \{(r_1, \ldots, r_n) : r_i \geq 0 \text{ and } \sum_{i=1}^{n} r_i = r\}$ and $\mathbb{E}[\prod_{i=1}^{n}(M_{i,n})_{[v_i]}]$ is given by 1 with $q = n$ and $l_i = i$, for any $i = 1, \ldots, n$. Eq. (13) thus leads, for the above choices of $c(i)$'s, to the distributions of the sampling versions of the Simpson index and of the Shannon entropy under a Gibbs-type prior.

3. Posterior analysis of rare variants

As already mentioned in Section 2, there exists a consolidated literature with plenty of results on unconditional properties of species sampling models and exchangeable random partitions. On the other hand, the investigation of the conditional properties of these models, given a sample generated by them, is more recent and many issues, which are essential in Bayesian nonparametric inference for species sampling problems, are still to be addressed. In this section we assume that the observations are modeled under the statistical framework (1) with $\tilde{P}$ being in the class of Gibbs-type random probability measures. Then, given an initial observed sample $(X_1, \ldots, X_n)$, we provide a comprehensive study on the posterior distribution of the rare variants.

Hereafter we resort to the notation set forth in Section 2 and we introduce some further quantities in order to describe the distribution of the random partition structure induced by an additional sample $(X_{n+1}, \ldots, X_{n+m})$, given that the initial sample $(X_1, \ldots, X_n)$ has been observed. Firstly, if we denote by $X_1^*, \ldots, X_{K_n}^*$ the labels identifying the K_n species in $(X_1, \ldots, X_n)$, then

$$L_m^{(n)} = \sum_{i=1}^{m} \prod_{j=1}^{K_n} \mathbb{1}_{\{X_{n+i} \neq X_j^*\}} \tag{14}$$

is the number of observations in $(X_{n+1}, \ldots, X_{n+m})$ which generate a certain number, say $K_m^{(n)}$, of species not coinciding with species already observed in the initial sample. These species will be referred to as new species. In particular, if $X_{K_n+1}^*, \ldots, X_{K_n+K_m^{(n)}}^*$ are the labels identifying the $K_m^{(n)}$ new species, then

$$S_{K_n+i} = \sum_{j=1}^{m} \mathbb{1}_{\{X_{n+j} = X_{K_n+i}^*\}}, \tag{15}$$

are their corresponding frequencies, for any $i = 1, \ldots, K_m^{(n)}$. Analogously, let

$$S_i = \sum_{j=1}^{m} \mathbb{1}_{\{X_{n+j} = X_i^*\}} \tag{16}$$

be the number of observations in $(X_{n+1}, \ldots, X_{n+m})$ which coincides with the ith species observed in $(X_1, \ldots, X_n)$, for any $i = 1, \ldots, K_n$. These species will be referred to as old species. See Lijoi et al. [29] and Favaro et al. [12] for a description of the random variables (14)–(16) by means of conditional partition probability functions when data are generated by Gibbs-type priors.

The random variables (15) and (16) play a crucial role in deriving the posterior distribution of rare variants. By means of (15) and (16) we define $N_{l,m}^{(n)}$ and $O_{l,m}^{(n)}$ as

$$\mathbb{P}\left[N_{l,m}^{(n)} = x\right] = \mathbb{P}\left[\sum_{i=1}^{K_m^{(n)}} \mathbb{1}_{\{S_{K_n+i}=l\}} = x \mid K_n = j, \boldsymbol{N}_n = (n_1, \ldots, n_j)\right] \tag{17}$$

and

$$\mathbb{P}\left[O_{l,m}^{(n)} = x\right] = \mathbb{P}\left[\sum_{i=1}^{K_n} \mathbb{1}_{\{N_i+S_i=l\}} = x \mid K_n = j, \boldsymbol{N}_n = (n_1, \ldots, n_j)\right]. \tag{18}$$

Given the random partition $(K_n, \boldsymbol{N}_n)$ induced by $(X_1, \ldots, X_n)$: (i) $N_{l,m}^{(n)}$ is the conditional number of species with frequency l among the new species in $(X_{n+1}, \ldots, X_{n+m})$; (ii) $O_{l,m}^{(n)}$ is the conditional number of species with frequency l among the old species in the whole sample $(X_1, \ldots, X_{n+m})$. Hence,

$$M_{l,m}^{(n)} = N_{l,m}^{(n)} + O_{l,m}^{(n)} \tag{19}$$

is the conditional number of species with frequency l in $(X_1, \ldots, X_{n+m})$. Note that (19) is a posterior counterpart of the random variable $M_{l,n}$ introduced in Section 2.

Let q be a positive integer and let $\mathbf{l} = (l_1, \ldots, l_q)$ be distinct positive integers. Moreover, let $\mathbf{r} = (r_1, \ldots, r_q)$ be positive integers and consider the sets $C_0 = \{0\}$ and $C_{r_i} = \{(c_{i,1}, \ldots, c_{i,r_i}) : N_{c_{i,t}} \leq l_i \ \forall\ t \text{ and } 1 \leq c_{i,t} < c_{i,h} \leq j \text{ if } t < h\}$. Denoting by $\mathbf{c}^{(r_i)} = (c_{i,1}, \ldots, c_{i,r_i})$ an element of the set C_{r_i}, let $\mathcal{C}_{\mathbf{r}} = \{(\mathbf{c}_1^{(r_1)}, \ldots, \mathbf{c}_q^{(r_q)}) : \mathbf{c}_i^{(r_i)} \in C_{r_i} \text{ and } c_{i,t} \neq c_{i',h}\ \forall\ t, h, \text{if } i \neq i'\}$. Finally, we denote by $\mathbf{c}^{(\mathbf{r})} = (\mathbf{c}_1^{(r_1)}, \ldots, \mathbf{c}_q^{(r_q)})$ an element of the set $\mathcal{C}_{\mathbf{r}}$. In the next theorem we derive the mixed falling factorial moment of order $\mathbf{r}$ of the random variable

$$\mathbf{O}_{\mathbf{l},m}^{(n)} = (O_{l_1,m}^{(n)}, O_{l_2,m}^{(n)}, \ldots, O_{l_q,m}^{(n)}).$$

The factorial moment of $O_{l,m}^{(n)}$ has been first derived in Theorem 2 in Favaro et al. [12]. Such a result represented a fundamental tool for computing the posterior distribution, and the corresponding Bayesian nonparametric estimator, of the number of old species with frequency l induced by the additional sample. The next theorem provides a generalization of Theorem 2 in Favaro et al. [12].

Theorem 2. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a Gibbs-type prior. Then, for any $1 \leq q \leq n+m$, $r_i \geq 1$ and $1 \leq l_i \neq l_j \leq n+m$ with $i \neq j$,*

$$\mathbb{E}\left[\prod_{i=1}^{q}(O_{l_i,m}^{(n)})_{[r_i]}\right] = \sum_{\mathbf{c}^{(\mathbf{r})}\in\mathcal{C}_{\mathbf{r}}} I_\alpha\left(q, m, \mathbf{l}, \mathbf{r}, \mathbf{n}, \mathbf{c}^{(\mathbf{r})}\right) \times \sum_{k=0}^{m-\sum_{i=1}^{q}(l_i r_i - |\mathbf{n}_{\mathbf{c}_i^{(r_i)}}|)} \frac{V_{n+m,j+k}}{V_{n,j}} \times \frac{\mathscr{C}\left(m - \sum_{i=1}^{q}(l_i r_i - |\mathbf{n}_{\mathbf{c}_i^{(r_i)}}|), k; \alpha, -n + \sum_{i=1}^{q}|\mathbf{n}_{\mathbf{c}_i^{(r_i)}}| + \alpha\left(j - \sum_{i=1}^{q} r_i\right)\right)}{\alpha^k}, \tag{20}$$

with

$$I_\alpha\left(q, m, \mathbf{l}, \mathbf{r}, \mathbf{n}, \mathbf{c}^{(\mathbf{r})}\right) = \frac{m!}{\left(m - \sum_{i=1}^{q}(l_i r_i - |\mathbf{n}_{\mathbf{c}_i^{(r_i)}}|)\right)!} \mathbb{1}_{\{0,1,\ldots,m\}}\left(m - \sum_{i=1}^{q}(l_i r_i - |\mathbf{n}_{\mathbf{c}_i^{(r_i)}}|)\right) \prod_{i=1}^{q} r_i! \prod_{t=1}^{r_i} \frac{(n_{c_{i,t}} - \alpha)_{(l_i - n_{c_{i,t}})}}{(l_i - n_{c_{i,t}})!}.$$

If $q = \tau < n+m$ and $l_i = i$ for $i = 1, \ldots, \tau$, then Eq. (20) characterizes the posterior distribution of the so-called old rare variants. Precisely, by old rare variants we mean species with frequencies $1, 2, \ldots, \tau$ among the old species in the whole sample $(X_1, \ldots, X_{n+m})$. In the next corollary we present a collection of results which can be derived by a direct application of Theorem 2 under the assumption of a two parameter Poisson–Dirichlet prior.

Corollary 2. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a two parameter Poisson–Dirichlet prior. Then, for any $\alpha \in [0, 1)$ and $\theta > -\alpha$ one has*

(i) *for any $1 \leq \tau \leq n+m$ and $r_i \geq 1$*

$$\mathbb{E}\left[\prod_{i=1}^{\tau}(O_{i,m}^{(n)})_{[r_i]}\right] = \sum_{\mathbf{c}^{(\mathbf{r})}\in\mathcal{C}_{\mathbf{r}}} I_\alpha\left(\tau, m, (1, \ldots, \tau), \mathbf{r}, \mathbf{n}, \mathbf{c}^{(\mathbf{r})}\right) \times \frac{\left(\theta + n - \sum_{i=1}^{\tau}|\mathbf{n}_{\mathbf{c}_i^{(r_i)}}| + \alpha\sum_{i=1}^{\tau} r_i\right)_{\left(m - \sum_{i=1}^{\tau}(i r_i - |\mathbf{n}_{\mathbf{c}_i^{(r_i)}}|)\right)}}{(\theta + n)_{(m)}}; \tag{21}$$

(ii) *for any $1 \leq l_1 \neq l_2 \leq n+m$*

$$\mathbb{E}\left[O_{l_1,m}^{(n)} O_{l_2,m}^{(n)}\right] = \sum_{\mathbf{c}^{(1,1)}\in\mathcal{C}_{(1,1)}} I_\alpha\left(2, m, (l_1, l_2), (1, 1), \mathbf{n}, \mathbf{c}^{(1,1)}\right) \times \frac{(\theta + n + 2\alpha - |\mathbf{n}_{\mathbf{c}_1^{(1)}}| - |\mathbf{n}_{\mathbf{c}_2^{(1)}}|)_{(m - l_1 + |\mathbf{n}_{\mathbf{c}_1^{(1)}}| - l_2 + |\mathbf{n}_{\mathbf{c}_2^{(1)}}|)}}{(\theta + n)_{(m)}}; \tag{22}$$

(iii) *for any $1 \leq \tau < n+m$ and $1 \leq l_i \neq l_j < n+m$ with $i \neq j$, as $m \to +\infty$*

$$\frac{1}{c_m(\alpha)} \sum_{i=1}^{\tau} O_{l_i,m}^{(n)} \xrightarrow{w} 0. \tag{23}$$

With regard to the new species generated by the additional sample, in the next theorem we establish a result analogous to Theorem 2 for the random variable

$$\mathbf{N}_{\mathbf{l},m}^{(n)} = (N_{l_1,m}^{(n)}, N_{l_2,m}^{(n)}, \ldots, N_{l_q,m}^{(n)}).$$

The falling factorial moment of the random variable $N_{l,m}^{(n)}$ has been first derived in Theorem 3 in Favaro et al. [12]. Such a result led to the posterior distribution, and to the corresponding Bayesian nonparametric estimator, of the number of new species with frequency l induced by the additional sample. The next theorem provides a generalization of Theorem 3 in Favaro et al. [12].

Theorem 3. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a Gibbs-type prior. Then, for any $1 \leq q \leq m$, $r_i \geq 1$ and $1 \leq l_i \neq l_j \leq m$ with $i \neq j$,*

$$\mathbb{E}\left[\prod_{i=1}^{q}(N_{l_i,m}^{(n)})_{[r_i]}\right] = J_\alpha\left(q, m, \mathbf{l}, \mathbf{r}\right) \times \sum_{k=0}^{m - \sum_{i=1}^{q} l_i r_i} \frac{V_{n+m,j+k+\sum_{i=1}^{q} r_i}}{V_{n,j}} \frac{\mathscr{C}\left(m - \sum_{i=1}^{q} l_i r_i, k; \alpha, -n + j\alpha\right)}{\alpha^k}, \tag{24}$$

with

$$J_\alpha(q, m, \mathbf{l}, \mathbf{r}) = \frac{m!}{\left(m - \sum_{i=1}^{q} l_i r_i\right)!} \mathbb{1}_{\{0,1,\ldots,m\}}\left(m - \sum_{i=1}^{q} l_i r_i\right) \times \prod_{i=1}^{q}\left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{r_i}.$$

If $q = \tau < m$ and $l_i = i$ for $i = 1, \ldots, \tau$, then Eq. (20) characterizes the posterior distribution of the so-called new rare variants. Precisely, by new rare variants we mean species with frequency $1, 2, \ldots, \tau$ among the new species in $(X_{n+1}, \ldots, X_{n+m})$. Similarly to Corollary 2, in the next corollary we present a collection of results which can be derived from Theorem 2 under the assumption of a two parameter Poisson–Dirichlet prior. Before stating the next corollary, let us introduce a nonnegative random variable $S^{(n,j)}_{\alpha,\theta}$, for any $\alpha \in (0, 1)$ and $\theta > -\alpha$, and for any $n \geq 1$ and $j \leq n$. Specifically we define

$$S^{(n,j)}_{\alpha,\theta} = B_{j+\theta/\alpha, n/\alpha - j} S_{\alpha,(\theta+n)/\alpha} \tag{25}$$

where $B_{j+\theta/\alpha,n/\alpha-j}$ and $Z_{\alpha,(\theta+n)/\alpha}$ are two independent random variables distributed according to a Beta distribution and a distribution with density function (2)

Corollary 3. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a two parameter Poisson–Dirichlet prior. Then, for any $\alpha \in [0, 1)$ and $\theta > -\alpha$ one has*

(i) *for any $1 \leq \tau \leq n$ and $r_i \geq 1$,*

$$\mathbb{E}\left[\prod_{i=1}^{\tau}(N^{(n)}_{i,m})_{[r_i]}\right] = J_\alpha(\tau, m, (1, \ldots, \tau), \mathbf{r}) \times \alpha^{\sum_{i=1}^{\tau} r_i}\left(\frac{\theta}{\alpha} + j\right)_{(\sum_{i=1}^{\tau} r_i)} \frac{\left(\theta + n + \alpha \sum_{i=1}^{\tau} r_i\right)_{(m - \sum_{i=1}^{\tau} i r_i)}}{(\theta + n)_{(m)}}; \tag{26}$$

(ii) *for any $1 \leq l_1 \neq l_2 \leq m$*

$$\mathbb{E}\left[N^{(n)}_{l_1,m} N^{(n)}_{l_2,m}\right] = J_\alpha(2, m, (l_1, l_2), (1, 1)) \times \alpha^2\left(\frac{\theta}{\alpha} + j\right)\left(\frac{\theta}{\alpha} + j + 1\right) \frac{(\theta + n + 2\alpha)_{(m-l_1-l_2)}}{(\theta + n)_{(m)}}; \tag{27}$$

(iii) *for any $1 \leq \tau < n + m$ and $1 \leq l_i \neq l_j < n + m$ with $i \neq j$, as $m \to +\infty$*

$$\frac{1}{c_m(\alpha)} \sum_{i=1}^{\tau} N^{(n)}_{l_i,m} \xrightarrow{w} \begin{cases} \left(\sum_{i=1}^{\tau} \frac{\alpha(1-\alpha)_{(l_i-1)}}{l_i!}\right) S^{(n,j)}_{\alpha,\theta} & \text{if } \alpha \in (0, 1) \\ \sum_{i=1}^{\tau} Z_{\theta/l_i} & \text{if } \alpha = 0. \end{cases} \tag{28}$$

A suitable combination of the results stated in Theorems 2 and 3 provides the mixed falling factorial moment of order $\mathbf{r}$ of the random variable

$$\mathbf{M}^{(n)}_{\mathbf{l},m} := (M^{(n)}_{l_1,m}, M^{(n)}_{l_2,m}, \ldots, M^{(n)}_{l_q,m}).$$

As for $O^{(n)}_{l,m}$ and $N^{(n)}_{l,m}$, the falling factorial moment of $M^{(n)}_{l,m}$ has been derived in Theorem 4 in Favaro et al. [12] and applied to determine the posterior distribution, and the corresponding Bayesian nonparametric estimator, of the number of species with frequency l induced by the additional sample. The next theorem provides a generalization of Theorem 4 in Favaro et al. [12].

Theorem 4. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a Gibbs-type prior. Then, for any $1 \leq q \leq n + m$, $r_i \geq 1$ and $1 \leq l_i \neq l_j \leq n + m$ with $i \neq j$,*

$$\mathbb{E}\left[\prod_{i=1}^{q}(M^{(n)}_{l_i,m})_{[r_i]}\right] = \sum_{x_1=0}^{r_1} \cdots \sum_{x_q=0}^{r_q} \sum_{\mathbf{c}^{(\mathbf{x})} \in \mathcal{C}_{\mathbf{x}}} \tilde{H}_\alpha\left(q, m, \mathbf{l}, \mathbf{r}, \mathbf{x}, \mathbf{n}, \mathbf{c}^{(\mathbf{x})}\right) \times \sum_{k=0}^{m - \sum_{i=1}^{q}(l_i r_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)} \frac{V_{n+m, j+k+\sum_{i=1}^{q}(r_i - x_i)}}{V_{n,j}}$$
$$\times \frac{\mathscr{C}\left(m - \sum_{i=1}^{q}(l_i r_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|), k; \alpha, -n + \sum_{i=1}^{q} |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}| + \alpha\left(j - \sum_{i=1}^{q} x_i\right)\right)}{\alpha^k}, \tag{29}$$

with

$$\tilde{H}_\alpha\left(q, m, \mathbf{l}, \mathbf{r}, \mathbf{x}, \mathbf{n}, \mathbf{c}^{(\mathbf{x})}\right) = \frac{m!}{\left(m - \sum_{i=1}^{q}(l_i r_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)!}\mathbb{1}_{\{0,1,\ldots,m\}}\left(m - \sum_{i=1}^{q}(l_i r_i - |\mathbf{n}_{\mathbf{c}_i^{(r_i)}}|)\right) \times \prod_{i=1}^{q}\frac{r_i!}{(r_i - x_i)!}\left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{r_i - x_i}\prod_{t=1}^{x_i}\frac{(n_{c_{i,t}} - \alpha)_{(l_i - n_{c_{i,t}})}}{(l_i - n_{c_{i,t}})!}.$$

Theorem 4 represents the posterior counterpart of Theorem 1. If $q = \tau < n + m$ and $l_i = i$ for $i = 1, \ldots, \tau$, then (29) characterizes the posterior distribution of the rare variants. Similarly to Corollaries 2 and 3, in the next corollary we present a collection of results which can be derived by Theorem 4 under the assumption of a two parameter Poisson–Dirichlet prior. In Section 4 these results will be applied to a benchmark genomic dataset.

Corollary 4. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a two parameter Poisson–Dirichlet prior. Then, for any $\alpha \in [0, 1)$ and $\theta > -\alpha$ one has*

(i) *for any $1 \leq \tau \leq n + m$ and $r_i \geq 1$,*

$$\mathbb{E}\left[\prod_{i=1}^{\tau}(M_{i,m}^{(n)})_{[r_i]}\right] = \sum_{x_1=0}^{r_1}\cdots\sum_{x_\tau=0}^{r_\tau}\alpha^{\sum_{i=1}^{\tau}(r_i - x_i)}\sum_{\mathbf{c}^{(\mathbf{x})}\in\mathcal{C}_{\mathbf{x}}}\tilde{H}_\alpha\left(\tau, m, (1, \ldots, \tau), \mathbf{r}, \mathbf{x}, \mathbf{n}, \mathbf{c}^{(\mathbf{x})}\right) \times \left(\frac{\theta}{\alpha} + j\right)_{\left(\sum_{i=1}^{\tau}(r_i - x_i)\right)}\frac{\left(\theta + n + \alpha\sum_{i=1}^{\tau}r_i - \sum_{i=1}^{\tau}|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|\right)_{\left(m - \sum_{i=1}^{\tau}(ir_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)}}{(\theta + n)_{(m)}}; \tag{30}$$

(ii) *for any $1 \leq l_1 \neq l_2 \leq n + m$*

$$\mathbb{E}\left[M_{l_1,m}^{(n)}M_{l_2,m}^{(n)}\right] = \sum_{x_1\in\{0,1\}}\sum_{x_2\in\{0,1\}}\alpha^{2-x_1-x_2}\times\sum_{\mathbf{c}^{(x_1,x_2)}\in\mathcal{C}_{(x_1,x_2)}}\tilde{H}_\alpha\left(2, m, (l_1, l_2), (1, 1), (x_1, x_2), \mathbf{n}, \mathbf{c}^{(x_1,x_2)}\right) \times \left(\frac{\theta}{\alpha} + j\right)_{(2-x_1-x_2)}\frac{(\theta + n + 2\alpha - |\mathbf{n}_{\mathbf{c}_1^{(x_1)}}| - |\mathbf{n}_{\mathbf{c}_2^{(x_2)}}|)_{(m - l_1 + |\mathbf{n}_{\mathbf{c}_1^{(x_1)}}| - l_2 + |\mathbf{n}_{\mathbf{c}_2^{(x_2)}}|)}}{(\theta + n)_{(m)}}; \tag{31}$$

(iii) *for any $1 \leq \tau < n + m$ and $1 \leq l_i \neq l_j < n + m$ with $i \neq j$, as $m \to +\infty$*

$$\frac{1}{c_m(\alpha)}\sum_{i=1}^{\tau}M_{l_i,m}^{(n)} \xrightarrow{w} \begin{cases}\left(\sum_{i=1}^{\tau}\frac{\alpha(1-\alpha)_{(l_i-1)}}{l_i!}\right)S_{\alpha,\theta}^{(n,j)} & \text{if } \alpha \in (0, 1)\\ \sum_{i=1}^{\tau}Z_{\theta/l_i} & \text{if } \alpha = 0.\end{cases} \tag{32}$$

We conclude by pointing out the usefulness of Theorem 4 for determining the posterior distribution of sampling diversity indexes. Indeed, along lines similar to those presented in Section 2, Theorem 4 leads to an explicit expression for

$$\mathbb{P}[D^{(n)} = x] = \mathbb{P}\left[\sum_{i=1}^{n+m}c(i)M_{i,m}^{(n)} = x \mid K_n = j, \mathbf{N}_n = (n_1, \ldots, n_j)\right]. \tag{33}$$

The distribution (33) takes on the interpretation of the posterior counterpart of the distribution of the sampling diversity index D. Hence, according to suitable specification of the $c(i)$'s one obtains the posterior distributions of the Simpson index and of the sampling version of the Shannon entropy under a Gibbs-type prior.

4. Illustrations

We present two novel statistical applications which exploit the results stated in Theorem 1, Corollary 1, Theorem 4 and Corollary 4. Firstly, we devise a new methodology for deriving approximated credible intervals for Bayesian nonparametric estimators of the number of rare species. Secondly, we study the correlation, both a priori and a posteriori, between frequency counts of different orders. Although these quantities can be studied for generic Gibbs-type random probability measures, for illustrative purposes we specify our analysis to the two parameter Poisson–Dirichlet process, with $\alpha \in (0, 1)$.

The applications will be illustrated by means of the analysis of a real genomic dataset. To be more specific, we refer to a widely used EST dataset obtained by sequencing a tomato-flower cDNA library (made from 0 to 3 mm buds of tomato flowers) from the Institute for Genomic Research Tomato Gene Index with library identifier T1526. See Quackenbush [38] and references therein for details. The observed sample consists of $n = 2586$ ESTs and features $j = 1825$ unique genes whose frequencies can be summarized as follows

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

where $i \in \{1, 2, \ldots, 14\} \cup \{16, 23, 27\}$. In order to specify the values of θ and α characterizing the Poisson–Dirichlet process, we adopt the empirical Bayes procedure undertaken in Lijoi et al. [25] and Favaro et al. [12]. This consists in choosing, for the parameter vector (α, θ), the values that maximize the EPPF with respect to the observed sample. Under the assumption of a two parameter Poisson–Dirichlet prior, this EPPF is obtained by plugging (4) in (3). For the dataset we are considering, this method leads to $\hat{\alpha} = 0.612$ and $\hat{\theta} = 741$.

4.1. Asymptotic credible intervals

Let $R^{(n)}_{\tau,m} = \sum_{i=1}^{\tau} M^{(n)}_{i,m}$ be the number of species with frequency less than or equal to τ in the enlarged sample. Recall that the additional sample is assumed to be not observed. We consider the problem of making Bayesian nonparametric inference on $R^{(n)}_{\tau,m}$. The Bayesian nonparametric estimator of $R^{(n)}_{\tau,m}$ coincides with

$$\hat{R}^{(n)}_{\tau,m} = \sum_{i=1}^{\tau} \hat{M}^{(n)}_{i,m}, \tag{34}$$

where $\hat{M}^{(n)}_{i,m} = \mathbb{E}[M^{(n)}_{i,m}]$ is the estimated number of species with frequency i. A closed form expression for $\hat{M}^{(n)}_{i,m}$ was first derived in Favaro et al. [12] and can be obtained, as a special case, from (29). Here we present a novel methodology to approximately quantify the uncertainty of the estimator $\hat{R}^{(n)}_{\tau,m}$. To this end, we observe that fluctuation (32) provides a useful tool for approximating the distribution of the random variable $R^{(n)}_{\tau,m}$, under the two-parameter Poisson–Dirichlet assumption. We aim at exploiting such limiting result in order to construct asymptotic credible intervals for the estimators.

First, we observe that the same limiting result as in (32) would clearly hold true for any scaling factor $r(m)$ such that $r(m) \approx c_m(\alpha) = m^{\alpha}$. Numerical investigations show that, as soon as θ and n are not overwhelmingly smaller than m,

$$m^{\alpha} \sum_{i=1}^{\tau} \frac{\alpha(1-\alpha)_{(i-1)}}{i!} \mathbb{E}[S^{(n,j)}_{\alpha,\theta}] \tag{35}$$

can be far from the exact estimator $\hat{R}^{(n)}_{\tau,m}$. For this reason we introduce the scaling $r^*_{\tau}(m) \approx m^{\alpha}$ such that $\hat{R}^{(n)}_{\tau,m} = r^*_{\tau}(m) \sum_{i=1}^{\tau} (\alpha(1-\alpha)_{(i-1)}/i!)\mathbb{E}[S^{(n,j)}_{\alpha,\theta}]$, and we define

$$\breve{R}^{(n)}_{\tau,m} = r^*_{\tau}(m) \sum_{i=1}^{\tau} \frac{\alpha(1-\alpha)_{(i-1)}}{i!} \mathbb{E}[S^{(n,j)}_{\alpha,\theta}]. \tag{36}$$

To keep the exposition as simple as possible we do not provide the expression for $r^*_{\tau}(m)$. See Favaro et al. [13] for a similar approach in Bayesian nonparametric inference for the number of new species generated by the additional sample.

In order to obtain asymptotic credible intervals for $\hat{R}^{(n)}_{\tau,m}$, we evaluate appropriate quantiles of the distribution of the limiting random variable $S^{(n,j)}_{\alpha,\theta}$ in (25). Let s_1 and s_2 be quantiles of the distribution of $S^{(n,j)}_{\alpha,\theta}$ such that (s_1, s_2) is the 95% credible interval. Then, according to (32) and (36), one has

$$\left(r^*_{\tau}(m) \sum_{i=1}^{\tau} \frac{\alpha(1-\alpha)_{(i-1)}}{i!} s_1,\ r^*_{\tau}(m) \sum_{i=1}^{\tau} \frac{\alpha(1-\alpha)_{(i-1)}}{i!} s_2 \right) \tag{37}$$

is a 95% asymptotic credible interval for $\hat{R}^{(n)}_{\tau,m}$. In order to determine the quantiles s_1 and s_2, we devised an algorithm for sampling the limiting random variable $S^{(n,j)}_{\alpha,\theta}$. To this end, we combine the algorithm proposed in Favaro et al. [13] with the fast rejection algorithm for sampling from an exponentially tilted positive α-stable random variable. See Hofert [17] for details.

As for the analysis of the tomato dataset, we compare the estimated numbers of rare species and the associated uncertainty, under different choices for the threshold parameter τ and different sizes of the additional unobserved sample. Specifically, in Fig. 1 we have plotted $\hat{R}^{(n)}_{1,m}$, $\hat{R}^{(n)}_{2,m}$ and $\hat{R}^{(n)}_{5,m}$, as a function of the size m of the additional unobserved sample in $[0, 3000]$. Each estimate is endowed with asymptotic 95% credible intervals, obtained by applying the described methodology. We conclude this section by observing that the same methodology can be adopted when interest is, more in general, on the uncertainty of the estimated total number of species that appear in the enlarged sample of size $n+m$ with frequency l in $\{l_1, \ldots, l_\tau\}$, with $\{l_1, \ldots, l_\tau\}$ being distinct integers such that $l_i \in \{1, \ldots, m+m\}$ for every i.

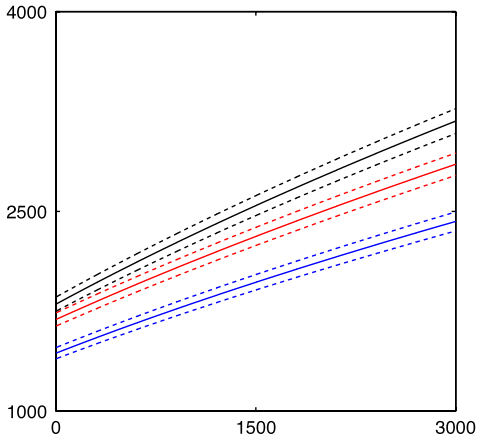


Fig. 1. Tomato dataset. Exact estimates $\hat{R}^{(n)}_{1,m}$ (blue solid curves), $\hat{R}^{(n)}_{2,m}$ (red solid curves) and $\hat{R}^{(n)}_{5,m}$ (black solid curves), together with asymptotic 95% credible intervals (dashed curves), as a function of the size m of the additional sample. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Correlation

We now investigate the relation between the number of variants of two specified frequencies l_i and l_j by computing their prior and posterior correlation. More specifically, as for the unconditional analysis, we are interested in the quantity

$$\mathrm{Corr}(M_{l_i,n}, M_{l_j,n}) \tag{38}$$

for some $l_i \neq l_j$ in $\{1, \ldots, n\}$, while, when the posterior analysis is concerned, we consider an observed sample of size n and focus our attention on the quantity

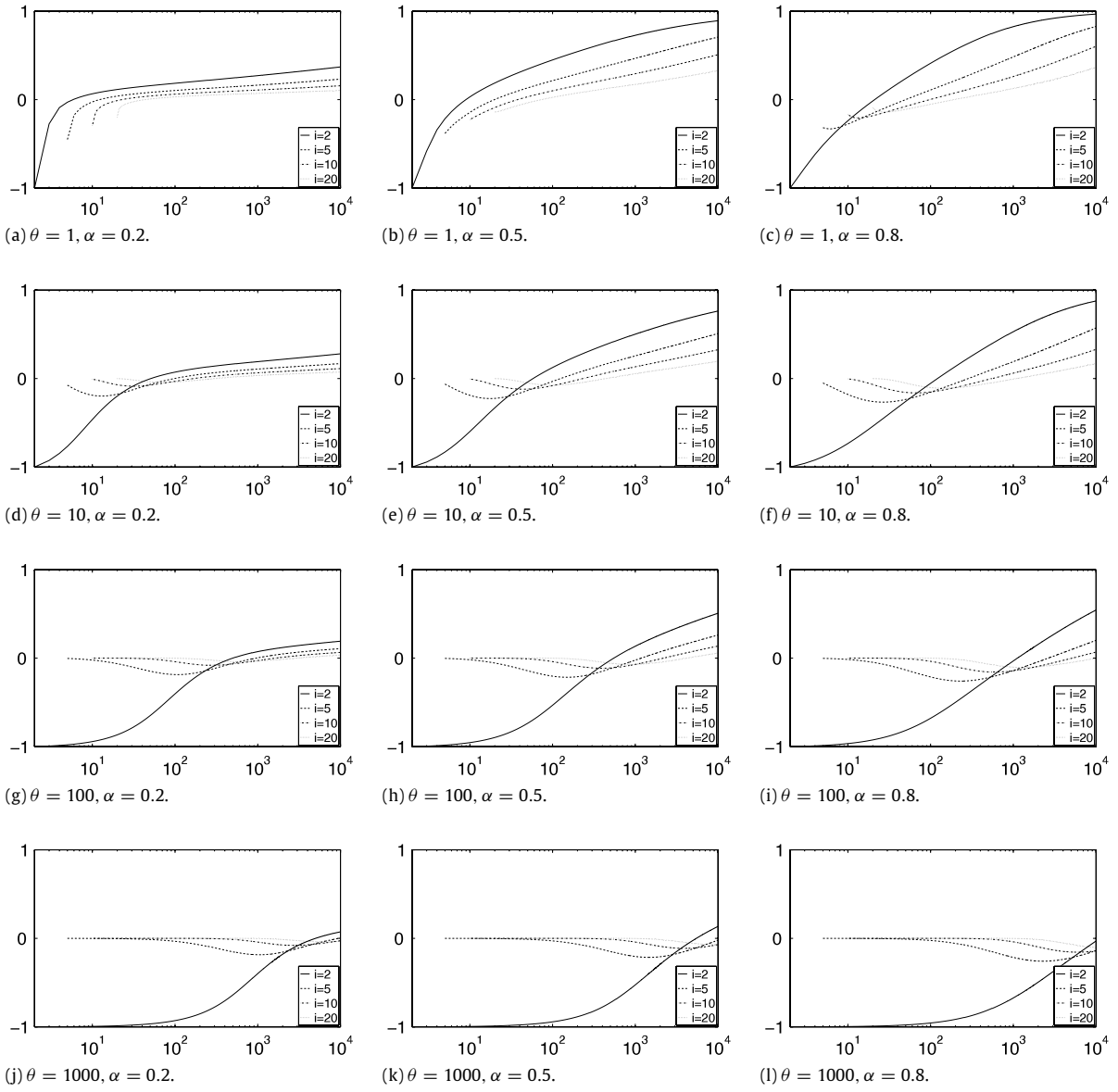
$$\mathrm{Corr}(M^{(n)}_{l_i,m}, M^{(n)}_{l_j,m}), \tag{39}$$

where $l_i \neq l_j$ are in $\{1, \ldots, n+m\}$. In this case, (9) and (30), and in particular (10) and (31), along with the relation between the rth moment and the rth factorial moment, are all needed to obtain explicit expressions for (38) and (39).

The analysis we carry out is two-fold. On the one side we investigate the role of the parameters θ and α in determining the prior correlation between $M_{l_i,n}$ and $M_{l_j,n}$. On the other side, after tuning these parameters by means of the aforementioned empirical Bayes procedure, we apply the proposed estimators to the tomato dataset, by investigating $\mathrm{Corr}(M^{(n)}_{l_1,m}, M^{(n)}_{l_2,m})$, conditionally on an observed sample of size n. In Fig. 2 we show the correlation between $M_{1,n}$ and $M_{i,n}$, as a function of the sample size n, for i that ranges in $\{2, 5, 10, 20\}$. The parameter θ is fixed as being equal to 1, 10, 100 and 1000 in first, second, third and fourth row respectively, while α is set equal to 0.2, 0.5 and 0.8 in first, second and third column respectively. First of all we notice that the correlation between $M_{1,n}$ and $M_{i,n}$ is not defined when $n < i$ since, in that case, the distribution of $M_{i,n}$ would be degenerate at 0.

By investigating the plots, we recognize some common patterns such as a negative correlation for every choice of i, when n is small. This is in accordance with the intuition since, for example, it is immediate to check that $\mathrm{Corr}(M_{1,n}, M_{2,n}) = -1$ when $n = 2$, given that, after two observations, the only two possible realizations of the vector $(M_{1,n}, M_{2,n})$ are $\{(2, 0), (0, 1)\}$. Another feature that characterizes every plot in Fig. 2 is that, the smaller is i the larger is the range of values taken by $\mathrm{Corr}(M_{1,n}, M_{2,n})$. This means that the prior distribution favors larger correlations, positive or negative, between number of variants of contiguous frequencies and penalizes the same quantity for variants with frequencies that differ significantly. As for the effect of the parameters, it is apparent that α has a role in determining the absolute value of the correlations: for n sufficiently large, small values of α lead to a small correlation, while a large α determines possibly high correlations. This effect is particularly evident when the parameter θ is small (compare, e.g., Figs. 2(a) and (c)). At the same time, θ has an impact on the sign of the correlation, since small values of θ, for n sufficiently large, give rise to positive correlations, while negative ones are generated by large values of θ. This can be appreciated by comparing, for example, Figs. 2(c) and (l).

In Fig. 3 we show how the posterior correlation of $(M^{(n)}_{1,m}, M^{(n)}_{i,m})$, for i that ranges in $\{2, 5, 10, 20\}$, varies with the additional sample size m. When m is small, $M^{(n)}_{1,m}$ and $M^{(n)}_{2,m}$ turn out to be negatively correlated. Such correlation becomes less significant when the additional sample size grows, and eventually positive when m approaches 10^4. On the other side, according to the data, $M^{(n)}_{1,m}$ and $M^{(n)}_{i,m}$ can be approximately considered uncorrelated if $i = 5, 10, 20$, regardless to the size of the additional sample. Finally, the correlation of $(M^{(n)}_{1,m}, M^{(n)}_{20,m})$ is not defined if $m < 4$ since, in that case, according to the observed sample, the distribution of $M^{(n)}_{20,m}$ would be degenerate at 0.



(a) $\theta = 1, \alpha = 0.2$. (b) $\theta = 1, \alpha = 0.5$. (c) $\theta = 1, \alpha = 0.8$.

(d) $\theta = 10, \alpha = 0.2$. (e) $\theta = 10, \alpha = 0.5$. (f) $\theta = 10, \alpha = 0.8$.

(g) $\theta = 100, \alpha = 0.2$. (h) $\theta = 100, \alpha = 0.5$. (i) $\theta = 100, \alpha = 0.8$.

(j) $\theta = 1000, \alpha = 0.2$. (k) $\theta = 1000, \alpha = 0.5$. (l) $\theta = 1000, \alpha = 0.8$.

Fig. 2. $\mathrm{Corr}(M_{1,n}, M_{i,n})$ as a function of n, under the two parameter Poisson–Dirichlet process, for different values of θ and α. Logarithmic scale for the x-axis.

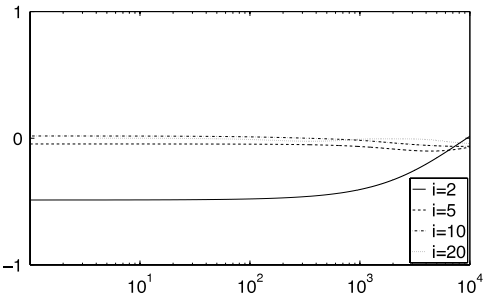


Fig. 3. Tomato dataset. $\mathrm{Corr}(M^{(n)}_{1,m}, M^{(n)}_{2,m})$ as a function of m, for $i = 2, 5, 10, 20$, under the two parameter Poisson–Dirichlet process. The observed sample size is $n = 2586$ and the parameters are set so that $\hat{\theta} = 741$ and $\hat{\alpha} = 0.612$. Logarithmic scale for the x-axis.

Acknowledgments

The authors are grateful to an anonymous referee for valuable remarks and suggestions that have led to a substantial improvement of the paper. Stefano Favaro is supported by the European Research Council (ERC) through StG N-BNP 306406.

Appendix

We start by recalling Lemma 1 in Favaro et al. [12]. This lemma provides an explicit expression for the conditional distributions of the random variables (15) and (16), given the initial observed sample. In addition to the notation introduced in Section 3, 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

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

Further additional notations will be introduced during the proofs when necessary.

Lemma 1 (*Favaro et al. [12]*). *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a Gibbs-type prior. For any integer $1 \leq x \leq j$, let $\boldsymbol{q}^{(x)} = (q_1, \ldots, q_x)$ with $1 \leq q_1 < \cdots < q_x \leq j$ and let $\boldsymbol{S}_{\boldsymbol{q}^{(x)}} = (S_{q_1}, \ldots, S_{q_x})$. Then,*

$$\mathbb{P}[\boldsymbol{S}_{\boldsymbol{q}^{(x)}} = \boldsymbol{s}_{\boldsymbol{q}^{(x)}} \mid A_{n,m}(j, \boldsymbol{n}, s, k)] = \frac{(m-s)!}{(m-s-|\boldsymbol{s}_{\boldsymbol{q}^{(x)}}|)!} \prod_{i=1}^{x} \frac{(n_{q_i}-\alpha)_{(s_{q_i})}}{s_{q_i}!} \times \frac{(n-|\boldsymbol{n}_{\boldsymbol{q}^{(x)}}|-(j-x)\alpha)_{(m-s-|\boldsymbol{s}_{\boldsymbol{q}^{(x)}}|)}}{(n-j\alpha)_{(m-s)}} \tag{A.1}$$

for any vector $\boldsymbol{s}_{\boldsymbol{q}^{(x)}} = (s_{q_1}, \ldots, s_{q_x})$ of nonnegative integers such that $|\boldsymbol{s}_{\boldsymbol{q}^{(x)}}| = \sum_{i=1}^{x} s_{q_i} \leq m-s$. Moreover, for any integer $1 \leq y \leq k$, let $\boldsymbol{p}^{(y)} = (p_1, \ldots, p_y)$ with $1 \leq p_1 < \cdots < p_y \leq k$ and let $\boldsymbol{S}^*_{\boldsymbol{p}^{(y)}} := (S_{j+p_1}, \ldots, S_{j+p_y})$. Then,

$$\mathbb{P}\left[\boldsymbol{S}^*_{\boldsymbol{p}^{(y)}} = \boldsymbol{s}_{\boldsymbol{p}^{(y)}} \mid A_{n,m}(j, \boldsymbol{n}, s, k)\right] = \frac{s!}{(s-|\boldsymbol{s}_{\boldsymbol{p}^{(y)}}|)!} \prod_{i=1}^{y} \frac{(1-\alpha)_{(s_{j+p_i})}}{s_{j+p_i}!} \times \frac{(k-y)!}{k!}\alpha^y \frac{\mathscr{C}\left(s-|\boldsymbol{s}_{\boldsymbol{p}^{(y)}}|, k-y; \alpha\right)}{\mathscr{C}(s,k;\alpha)} \tag{A.2}$$

for any vector $\boldsymbol{s}_{\boldsymbol{p}^{(y)}} = (s_{j+p_1}, \ldots, s_{j+p_y})$ of positive integers such that $|\boldsymbol{s}_{\boldsymbol{p}^{(y)}}| = \sum_{i=1}^{y} s_{j+p_i} \leq s$. Finally, $\boldsymbol{S}_{\boldsymbol{q}^{(x)}}$ and $\boldsymbol{S}^*_{\boldsymbol{p}^{(y)}}$ are independent, conditionally on $(K_n, \boldsymbol{N}_n, L_m^{(n)}, K_m^{(n)})$.

Proof of Theorem 1. For any pair of integers $n \geq 1$ and $1 \leq j \leq n$ we denote by $\mathcal{M}_{n,j}$ the partition set of $\{1, \ldots, n\}$ which contains all the vectors $\mathbf{m} = (m_1, \ldots, m_n) \in \{0, 1, \ldots, n\}^n$ such that $\sum_{i=1}^{n} m_i = j$ and $\sum_{i=1}^{n} i m_i = n$. Hence, resorting to the Gibbs-type sampling formula (6), we can write

$$\mathbb{E}\left[\prod_{i=1}^{q}(M_{l_i,n})_{[r_i]}\right] = n! \sum_{j=1}^{n} V_{n,j} \sum_{\mathbf{m}\in\mathcal{M}_{n,j}} \prod_{i=1}^{q}(m_{l_i})_{[r_i]} \prod_{i=1}^{n}\left(\frac{(1-\alpha)_{(i-1)}}{i!}\right)^{m_i}\frac{1}{m_i!}.$$

Note that, according to the definition of the set $\mathcal{M}_{n,j}$, the sum over the index j is different from zero when $\sum_{i=1}^{q} r_i \leq j \leq n - \sum_{i=1}^{q}(l_i r_i - r_i)$. Accordingly, we have

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{q}(M_{l_i,n})_{[r_i]}\right] &= n! \sum_{j=\sum_{i=1}^{q} r_i}^{n-\sum_{i=1}^{q} l_i r_i + \sum_{i=1}^{q} r_i} V_{n,j} \\
&\quad \times \sum_{\mathbf{m}\in\mathcal{M}_{n,j}} \prod_{i=1}^{q}\left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{m_{l_i}} \frac{1}{(m_{l_i}-r_i)!} \prod_{i\notin\{l_1,\ldots,l_q\}}\left(\frac{(1-\alpha)_{(i-1)}}{i!}\right)^{m_i}\frac{1}{m_i!} \\
&= n! \prod_{i=1}^{q}\left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{r_i} \sum_{j=\sum_{i=1}^{q} r_i}^{n-\sum_{i=1}^{q} l_i r_i + \sum_{i=1}^{q} r_i} V_{n,j} \\
&\quad \times \sum_{\mathbf{m}\in\mathcal{M}_{n-\sum_{i=1}^{q} l_i r_i, j-\sum_{i=1}^{q} r_i}} \prod_{i=1}^{n}\left(\frac{(1-\alpha)_{(i-1)}}{i!}\right)^{m_i}\frac{1}{m_i!}.
\end{aligned} \tag{A.3}$$

Then, a direct application of Equation (2.82) in Charalambides [4] leads to the identity

$$\sum_{\mathbf{m}\in\mathcal{M}_{n-\sum_{i=1}^{q} l_i r_i, j-\sum_{i=1}^{q} r_i}} \prod_{i=1}^{n} \left(\frac{(1-\alpha)_{(i-1)}}{i!}\right)^{m_i} \frac{1}{m_i!} = \frac{\left(\sum_{i=1}^{q} l_i r_i\right)!}{n!} \binom{n}{\sum_{i=1}^{q} l_i r_i} \frac{\mathscr{C}\left(n-\sum_{i=1}^{q} l_i r_i, j-\sum_{i=1}^{q} r_i; \alpha\right)}{\alpha^{j-\sum_{i=1}^{q} r_i}}, \tag{A.4}$$

with $\mathscr{C}(n, k; \alpha)$ being the generalized factorial coefficient. The proof is completed by combining (A.3) with (A.4) and by means of standard algebra for falling factorials. □

Proof of Corollary 1. We present a sketch of the proof. Of course the starting point is Eq. (8). In particular, we set $q = \tau$ and, according to (4), we set

$$V_{n,j+\sum_{i=1}^{\tau} r_i} = \frac{\alpha^{j+\sum_{i=1}^{\tau} r_i}}{(\theta)_{(n)}} \left(\frac{\theta}{\alpha}\right)_{(\sum_{i=1}^{\tau} r_i)} \left(\frac{\theta}{\alpha} + \sum_{i=1}^{\tau} r_i\right)_{(j)}.$$

The sum over k in the resulting expression is solved by standard algebra on rising factorials and by means of Equation (2.49) in Charalambides [4]. Then (9) follows by setting $l_i = i$ for $i = 1, \ldots, \tau$, (10) follows by setting $\tau = 2$ and $r_1 = r_2 = 1$, and (11) follows by applying the Stirling approximation, for large m. □

Proof of Theorem 2. Using the definition of the random variable $O_{l,m}^{(n)}$ in (18), we can write

$$\mathbb{E}\left[\prod_{i=1}^{q} (O_{l_i,m}^{(n)})^{r_i}\right] = \sum_{s=0}^{m} \sum_{k=0}^{s} \mathbb{P}[K_m^{(n)} = k, L_m^{(n)} = s \mid A_n(j, \mathbf{n})] \times \mathbb{E}\left[\prod_{i=1}^{q} \left(\sum_{t=1}^{K_n} \mathbb{1}_{\{l_i\}}(N_t + S_t) \mid A_{n,m}(j, \mathbf{n}, s, k)\right)^{r_i}\right] \tag{A.5}$$

where

$$\mathbb{P}[K_m^{(n)} = k, L_m^{(n)} = s \mid A_n(j, \mathbf{n})] = \frac{V_{n+m,j+k}}{V_{n,j}} \binom{m}{s} (n - j\alpha)_{(m-s)} \frac{\mathscr{C}(s, k; \alpha)}{\alpha^k}. \tag{A.6}$$

See Lijoi et al. [29] for details on (A.6). The proof is along lines similar to the proof of Theorem 2 in Favaro et al. [12]. By a repeated application of the Binomial theorem, we rewrite the power function on the right-hand side of (A.5) as

$$\begin{aligned}\left(\sum_{t=1}^{K_n} \mathbb{1}_{\{l_i\}}(N_t + S_t)\right)^{r_i} &= \sum_{x_i=1}^{K_n \wedge r_i} \sum_{i_1=1}^{r_i-1} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x_i-1}=1}^{i_{x_i-2}-1} \sum_{i_{x_i}=0}^{0} \binom{r_i}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x_i-2}}{i_{x_i-1}} \\ &\quad \times \sum_{\mathbf{c}_i^{(x_i)} \in A_{x_i}} (\mathbb{1}_{\{l_i\}}(N_{c_{i,1}} + S_{c_{i,1}}))^{r_i - i_1} (\mathbb{1}_{\{l_i\}}(N_{c_{i,2}} + S_{c_{i,2}}))^{i_1 - i_2} \cdots \\ &\quad \times \cdots (\mathbb{1}_{\{l_i\}}(N_{c_{i,x_i-1}} + S_{c_{i,x_i-1}}))^{i_{x_i-2} - i_{x_i-1}} (\mathbb{1}_{\{l_i\}}(N_{c_{i,x_i}} + S_{c_{i,x_i}}))^{i_{x_i-1}}\end{aligned} \tag{A.7}$$

where

$$A_{x_i} = \{\mathbf{c}_i^{(x_i)} = (c_{i,1}, \ldots, c_{i,x_i}) \in \{1, \ldots, K_n\}^{x_i} : c_{i,t} < c_{i,s} \text{ if } t < s\}$$

is the set of all vectors $\mathbf{c}_i^{(x_i)}$ of length x_i that can be defined with elements of the set $\{1, \ldots, j\}$ and without repetitions. If $j < r_i$ then the set A_{x_i} is empty for any $x_i > j$, for $i = 1, \ldots, q$. Therefore, as a matter of convenience, we set the upper bound of the sum over x_i to be r_i. Since the power function on the right-hand side of (A.7) is written as a function of terms with exponent different from zero, we ignore the exponents of the indicator functions. Then, we have

$$\begin{aligned}\left(\sum_{t=1}^{K_n} \mathbb{1}_{\{l_i\}}(N_t + S_t)\right)^{r_i} &= \sum_{x_i=1}^{r_i} \sum_{i_1=1}^{r_i-1} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x_i-1}=1}^{i_{x_i-2}-1} \sum_{i_{x_i}=0}^{0} \binom{r_i}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x_i-2}}{i_{x_i-1}} \sum_{\mathbf{c}_i^{(x_i)} \in A_{x_i}} \prod_{i=1}^{x_i} \mathbb{1}_{\{l_i\}}(N_{c_{i,t}} + S_{c_{i,t}}) \\ &= \sum_{x_i=1}^{r_i} S(r_i, x_i) x_i! \sum_{\mathbf{c}_i^{(x_i)} \in A_{x_i}} \prod_{t=1}^{x_i} \mathbb{1}_{\{l_i\}}(N_{c_{i,t}} + S_{c_{i,t}})\end{aligned} \tag{A.8}$$

where $S(n, k)$ is the Stirling number of the second kind. See Charalambides [4] for details. The last equality of (A.8) is obtained by means of the well-known identity

$$\sum_{i_1=1}^{r_i-1} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x_i-1}=1}^{i_{x_i-2}-1} \sum_{i_{x_i}=0}^{0} \binom{r_i}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x_i-2}}{i_{x_i-1}} = x_i! S(r_i, x_i)$$

for any $r_i \geq 1$ and $1 \leq x_i \leq r_i$. Indeed, according to the combinatorial meaning of Stirling number of the second kind, the factor $x_i! S(r_i, x_i)$ in (A.8) is the number of ways of distributing r_i distinguishable objects into x_i distinguishable groups. From (A.8), and by means of standard algebra, we can write

$$\mathbb{E}\left[\prod_{i=1}^{q}\left(\sum_{t=1}^{K_n}\mathbb{1}_{\{l_i\}}(N_t+S_t)\mid A_{n,m}(j,\mathbf{n},k,s)\right)^{r_i}\right]=\sum_{x_1=1}^{r_1}\cdots\sum_{x_q=1}^{r_q}\prod_{i=1}^{q}S(r_i,x_i)x_i!$$
$$\times\sum_{\mathbf{c}^{(\mathbf{x})}\in\mathcal{C}_{\mathbf{x}}}\mathbb{P}[S_{\mathbf{c}_1^{(x_1)}}=l_1\mathbf{1}_{x_1}-\mathbf{n}_{\mathbf{c}_1^{(x_1)}},\ldots,S_{\mathbf{c}_q^{(xq)}}=l_q\mathbf{1}_{x_q}-\mathbf{n}_{\mathbf{c}_q^{(xq)}}\mid A_{n,m}(j,\mathbf{n},s,k)] \tag{A.9}$$

where $\mathbf{x}=(x_1,\ldots,x_q)$, $\mathbf{1}_{x_i}=(1,\ldots,1)$, $\mathbf{n}_{\mathbf{c}_i^{(x_i)}}=(n_{c_{i,1}},\ldots,n_{c_{i,x_i}})$ and where we set

$$\mathcal{C}_{\mathbf{x}}=\{\mathbf{c}^{(\mathbf{x})}=(\mathbf{c}_1^{(x_1)},\ldots,\mathbf{c}_q^{(x_q)}):\mathbf{c}_i^{(x_i)}\in C_{x_i}\ \text{ and } c_{i,t}\neq c_{i',h}\ \forall\, t,h, \text{if } i\neq i'\}$$

with $C_{x_i}=\{\mathbf{c}^{(x_i)}=(c_{i,1},\ldots,c_{i,x_i}):N_{c_{i,t}}\leq l_i\ \forall\ t \text{ and } 1\leq c_{i,t}<c_{i,h}\leq K_n \text{ if } t<h\}$. See Cesari [3] for further details about the derivation of the set $\mathcal{C}_{\mathbf{x}}$. A similar construction will be considered in the proof of Theorem 3. The conditional probability in (A.9) can be obtained from (A.1) in Lemma 1. Specifically, by combining (A.9) with (A.1) we can write the expected value in (A.9) as

$$\mathbb{E}\left[\prod_{i=1}^{q}\left(\sum_{t=1}^{K_n}\mathbb{1}_{\{l_i\}}(N_t+S_t)\mid A_{n,m}(j,\mathbf{n},k,s)\right)^{r_i}\right]$$
$$=\sum_{x_1=1}^{r_1}\cdots\sum_{x_q=1}^{r_q}\prod_{i=1}^{q}S(r_i,x_i)x_i!\times\sum_{\mathbf{c}^{(\mathbf{x})}\in\mathcal{C}_{\mathbf{x}}}\frac{(m-s)!}{\left(m-s-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)!}\prod_{i=1}^{q}\prod_{t=1}^{x_i}\frac{(n_{c_{i,t}}-\alpha)_{(l_i-n_{c_{i,t}})}}{(l_i-n_{c_{i,t}})!}$$
$$\times\frac{\left(n-\sum_{i=1}^{q}|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|-\alpha\left(j-\sum_{i=1}^{q}x_i\right)\right)_{\left(m-s-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)}}{(n-j\alpha)_{(m-s)}}. \tag{A.10}$$

The proof in completed by marginalizing the last expression over $(K_m^{(n)}, L_m^{(n)})$, with respect to the conditional distribution in (A.6), and by means of standard algebra involving rising factorials. In particular, by combining (A.5) with (A.10),

$$\mathbb{E}\left[\prod_{i=1}^{q}(O_{l_i,m}^{(n)})^{r_i}\right]=\sum_{k=0}^{m}\alpha^{-k}\frac{V_{n+m,j+k}}{V_{n,j}}\sum_{s=k}^{m}\mathscr{C}(s,k;\alpha)\sum_{x_1=1}^{r_1}\cdots\sum_{x_q=1}^{r_q}\prod_{i=1}^{q}S(r_i,x_i)x_i!$$
$$\times\sum_{\mathbf{c}^{(\mathbf{x})}\in\mathcal{C}_{\mathbf{x}}}\prod_{i=1}^{q}\prod_{t=1}^{x_i}\frac{(n_{c_{i,t}}-\alpha)_{(l_i-n_{c_{i,t}})}}{(l_i-n_{c_{i,t}})!}\times\frac{m!}{\left(m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)!}\binom{m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)}{s}$$
$$\times\left(n-\sum_{i=1}^{q}|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|-\alpha\left(j-\sum_{i=1}^{q}x_i\right)\right)_{\left(m-s-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)}.$$

Note that the sum over s has lower bound k and upper bound $m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)$. Indeed, according to the definition of Binomial coefficient, one obtains

$$\binom{m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)}{s}=0$$

for any $s>m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)$. Hence, we can write the last expression as

$$\sum_{x_1=1}^{r_1}\cdots\sum_{x_q=1}^{r_q}\prod_{i=1}^{q}S(r_i,x_i)x_i!\times\sum_{\mathbf{c}^{(\mathbf{x})}\in\mathcal{C}_{\mathbf{x}}}\prod_{i=1}^{q}\prod_{t=1}^{x_i}\frac{(n_{c_{i,t}}-\alpha)_{(l_i-n_{c_{i,t}})}}{(l_i-n_{c_{i,t}})!}$$
$$\times\frac{m!}{\left(m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)!}\sum_{k=0}^{m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)}\alpha^{-k}\frac{V_{n+m,j+k}}{V_{n,j}}$$

$$\times \sum_{s=k}^{m-\sum_{i=1}^{q}(l_i x_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)} \binom{m - \sum_{i=1}^{q}(l_i x_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)}{s} \mathscr{C}(s, k; \alpha)$$

$$\times \left(n - \sum_{i=1}^{q} |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}| - \alpha\left(j - \sum_{i=1}^{q} x_i\right)\right)_{\left(m-s-\sum_{i=1}^{q}(l_i x_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)}$$

(by means of Equation (2.56) in Charalambides [4])

$$= \sum_{x_1=1}^{r_1} \cdots \sum_{x_q=1}^{r_q} \prod_{i=1}^{q} S(r_i, x_i) x_i! \times \sum_{\mathbf{c}^{(\mathbf{x})} \in \mathcal{C}_{\mathbf{x}}} \prod_{i=1}^{q} \prod_{t=1}^{x_i} \frac{(n_{c_{i,t}} - \alpha)_{(l_i - n_{c_{i,t}})}}{(l_i - n_{c_{i,t}})!}$$

$$\times \frac{m!}{\left(m - \sum_{i=1}^{q}(l_i x_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)!} \sum_{k=0}^{m-\sum_{i=1}^{q}(l_i x_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)} \alpha^{-k} \frac{V_{n+m,j+k}}{V_{n,j}}$$

$$\times \mathscr{C}\left(m - \sum_{i=1}^{q}(l_i x_i - |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|), k; \alpha, -n + \sum_{i=1}^{q} |\mathbf{n}_{\mathbf{c}_i^{(x_i)}}| + \alpha\left(j - \sum_{i=1}^{q} x_i\right)\right),$$

with $\mathscr{C}(n, k; \alpha)$ being the generalized factorial coefficient. Eq. (20) follows from the last expression by the relation between the mixed moment of order $\mathbf{r}$ and the mixed falling factorial moment of order $\mathbf{r}$. See Charalambides [4] for details. □

Proof of Corollary 2. We present a sketch of the proof. Of course the starting point is Eq. (20). In particular, we set $q = \tau$ and, according to (4), we set

$$\frac{V_{n+m,j+k}}{V_{n,j}} = \frac{\alpha^k}{(\theta + n)_{(m)}} \left(\frac{\theta}{\alpha} + j\right)_{(k)}.$$

The sum over k in the resulting expression is solved by standard algebra on rising factorials and by means of Equation (2.49) in Charalambides [4]. Then (21) follows by setting $l_i = i$ for $i = 1, \ldots, \tau$, (22) follows by setting $\tau = 2$ and $r_1 = r_2 = 1$, and (23) follows by applying the Stirling approximation, for large m. □

Proof of Theorem 3. Using the definition of the random variable $N_{l,m}^{(n)}$ in (17), we can write

$$\mathbb{E}\left[\prod_{i=1}^{q} (N_{l_i,m}^{(n)})^{r_i}\right] = \sum_{s=0}^{m} \sum_{k=0}^{s} \mathbb{P}[K_m^{(n)} = k, L_m^{(n)} = s \mid A_n(j, \mathbf{n})] \mathbb{E}\left[\prod_{i=1}^{q} \left(\sum_{t=1}^{K_m^{(n)}} \mathbb{1}_{\{l_i\}}(S_{K_n+t}) \mid A_{n,m}(j, \mathbf{n}, s, k)\right)^{r_i}\right], \tag{A.11}$$

where the distribution of $(K_m^{(n)}, L_m^{(n)}) \mid A_n(j, \mathbf{n})$ in (A.11) is given by (A.6). Then, the proof is along lines similar to the proof of Theorem 3 in Favaro et al. [12]. Specifically, by means of a repeated application of the Binomial theorem, we rewrite the power function on the right-hand side of (A.11) as follows

$$\left(\sum_{t=1}^{K_m^{(n)}} \mathbb{1}_{\{l_i\}}(S_{K_n+t})\right)^{r_i} = \sum_{y_i=1}^{K_m^{(n)} \wedge r_i} \sum_{i_1=1}^{r_i-1} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{y_i-1}=1}^{i_{y_i-2}-1} \sum_{i_{y_i}=0}^{0} \binom{r_i}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{y_i-2}}{i_{y_i-1}}$$

$$\times \sum_{\mathbf{d}_i^{(y_i)} \in B_{y_i}} (\mathbb{1}_{\{l_i\}}(S_{K_n+d_{i,1}}))^{r_i - i_1} (\mathbb{1}_{\{l_i\}}(S_{K_n+d_{i,2}}))^{i_1 - i_2} \cdots$$

$$\times \cdots (\mathbb{1}_{\{l_i\}}(S_{K_n+d_{i,y_i-1}}))^{i_{y_i-2} - i_{y_i-1}} (\mathbb{1}_{\{l_i\}}(S_{K_n+d_{i,y_i}}))^{i_{y_i-1}}, \tag{A.12}$$

where

$$B_{y_i} = \{\mathbf{d}_i^{(y_i)} = (d_{i,1}, \ldots, d_{i,y_i}) \in \{1, \ldots, K_m^{(n)}\}^{y_i} : d_{i,t} < d_{i,s}, t \neq s\}$$

is the set of all vectors $\mathbf{d}_i^{(y_i)}$ of length y_i that can be defined with elements of the set $\{1, \ldots, k\}$ and without repetitions. As in the proof of Theorem 2, as a matter of convenience, we set the upper bound of the sum over y_i to be r_i. Because the power function on the right-hand side of (A.12) is written as a function of terms with exponent different from zero, we can ignore

the exponents of the indicator functions. Hence, we write the last expression as follows

$$\left(\sum_{t=1}^{K_m^{(n)}} \mathbb{1}_{\{l_i\}}(S_{K_n+t})\right)^{r_i} = \sum_{y_i=1}^{r_i}\sum_{i_1=1}^{r_i-1}\sum_{i_2=1}^{i_1-1}\cdots\sum_{i_{y_i-1}=1}^{i_{y_i-2}-1}\sum_{i_{y_i}=0}^{0}\binom{r_i}{i_1}\binom{i_1}{i_2}\cdots\binom{i_{y_i-2}}{i_{y_i-1}}\sum_{\mathbf{d}_i^{(y_i)}\in B_{y_i}}\prod_{i=1}^{y_i}\mathbb{1}_{\{l_i\}}(S_{K_n+d_{i,t}})$$

$$= \sum_{y_i=1}^{r_i} S(r_i, y_i)y_i! \sum_{\mathbf{d}_i^{(y_i)}\in B_{y_i}}\prod_{t=1}^{y_i}\mathbb{1}_{\{l_i\}}(S_{K_n+d_{i,t}}). \tag{A.13}$$

See the proof of Theorem 2 for details on the last equality of (A.13). According to the identity (A.13), and by means of standard algebraic manipulations, we have

$$\mathbb{E}\left[\prod_{i=1}^{q}\left(\sum_{t=1}^{K_m^{(n)}}\mathbb{1}_{\{l_i\}}(S_{K_n+t}) \mid A_{n,m}(j,\mathbf{n},k,s)\right)^{r_i}\right]$$

$$= \sum_{y_1=1}^{r_1}\cdots\sum_{y_q=1}^{r_q}\prod_{i=1}^{q} S(r_i,y_i)y_i! \times \sum_{\mathbf{d}(\mathbf{y})\in\mathcal{B}_{\mathbf{y}}}\mathbb{P}[S^*_{\mathbf{d}_1^{(y_1)}} = l_1\mathbf{1}_{y_1},\ldots,S^*_{\mathbf{d}_q^{(y_q)}} = l_q\mathbf{1}_{y_q} - \mathbf{n}_{\mathbf{d}_q^{(y_q)}} \mid A_{n,m}(j,\mathbf{n},s,k)] \tag{A.14}$$

where $\mathbf{y} = (y_1,\ldots,y_q)$, $\mathbf{1}_{y_i} = (1,\ldots,1)$, $\mathbf{n}_{\mathbf{d}_i^{(y_i)}} = (n_{d_{i,1}},\ldots,n_{d_{i,y_i}})$ and where we set

$$\mathcal{B}_{\mathbf{y}} = \{\mathbf{d}^{(\mathbf{y})} = (\mathbf{d}_1^{(y_1)},\ldots,\mathbf{d}_q^{(y_q)}) : \mathbf{d}_i^{(y_i)} \in B_{y_i},\ d_{i,t}\neq d_{i',h},\ 1\le i\neq i'\le q\}.$$

See Cesari [3] for further details about the derivation of the set $\mathcal{B}_{\mathbf{y}}$. The conditional probability in (A.14) is obtained from (A.2) in Lemma 1. Specifically, by combining (A.14) with (A.2) we can write the expected value in (A.14) as

$$\mathbb{E}\left[\prod_{i=1}^{q}\left(\sum_{t=1}^{K_m^{(n)}}\mathbb{1}_{\{l_i\}}(S_{K_n+t}) \mid A_{n,m}(j,\mathbf{n},k,s)\right)^{r_i}\right]$$

$$= \sum_{y_1=1}^{r_1}\cdots\sum_{y_q=1}^{r_q}\prod_{i=1}^{q} S(r_i,y_i)y_i! \times \sum_{\mathbf{d}(\mathbf{y})\in\mathcal{B}_{\mathbf{y}}}\frac{s!}{\left(s-\sum_{i=1}^{q}l_iy_i\right)!}\prod_{i=1}^{q}\prod_{t=1}^{y_i}\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\frac{\left(k-\sum_{i=1}^{q}y_i\right)!}{k!}$$

$$\times\,\alpha^{\sum_{i=1}^{q}y_i}\frac{\mathscr{C}\left(s-\sum_{i=1}^{q}l_iy_i,\,k-\sum_{i=1}^{q}y_i;\,\alpha\right)}{\mathscr{C}(s,k;\alpha)}$$

$$= \sum_{y_1=1}^{r_1}\cdots\sum_{y_q=1}^{r_q}\prod_{i=1}^{q} S(r_i,y_i)\frac{s!}{\left(s-\sum_{i=1}^{q}l_iy_i\right)!}\prod_{i=1}^{q}\left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{y_i}\times\alpha^{\sum_{i=1}^{q}y_i}\frac{\mathscr{C}\left(s-\sum_{i=1}^{q}l_iy_i,\,k-\sum_{i=1}^{q}y_i;\,\alpha\right)}{\mathscr{C}(s,k;\alpha)}, \tag{A.15}$$

where the last equality is obtained by using the cardinality of the set $\mathcal{B}_{\mathbf{y}}$; this cardinality is

$$\binom{k}{y_1,\ldots,y_q,k-\sum_{i=1}^{q}y_i}.$$

The proof is completed by marginalizing the last expression over $(K_m^{(n)}, L_m^{(n)})$, with respect to the conditional distribution in (A.6), and by means of standard algebra involving rising factorials. In particular, by combining (A.11) with (A.15),

$$\mathbb{E}\left[\prod_{i=1}^{q}(N_{l_i,m}^{(n)})^{r_i}\right] = \sum_{k=0}^{m}\alpha^{-k}\frac{V_{n+m,j+k}}{V_{n,j}}\sum_{s=k}^{m}\binom{m}{s}(n-j\alpha)_{(m-s)}\sum_{y_1=1}^{r_1}\cdots\sum_{y_q=1}^{r_q}\prod_{i=1}^{q}S(r_i,y_i)\prod_{i=1}^{q}\left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{y_i}$$

$$\times\,\frac{s!}{\left(s-\sum_{i=1}^{q}l_iy_i\right)!}\alpha^{\sum_{i=1}^{q}y_i}\mathscr{C}\left(s-\sum_{i=1}^{q}l_iy_i,\,k-\sum_{i=1}^{q}y_i;\,\alpha\right).$$

Note that the sum over k has lower bound zero and upper bound $m - \sum_{i=1}^{q} l_i y_i$ while the sum over s has lower bound k and upper bound $m - \sum_{i=1}^{q} l_i y_i$. Indeed, according to the definition of generalized factorial coefficient, one has

$$\mathscr{C}\left(s - \sum_{i=1}^{q} l_i y_i, k - \sum_{i=1}^{q} y_i; \alpha\right) = 0$$

for any k and s such that $s - \sum_{i=1}^{q} l_i y_i < k - \sum_{i=1}^{q} y_i$. Hence, we can write the last expression as

$$\begin{aligned}
&\sum_{y_1=1}^{r_1} \cdots \sum_{y_q=1}^{r_q} \prod_{i=1}^{q} S(r_i, y_i) \prod_{i=1}^{q} \left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{y_i} \times \frac{m!}{\left(m - \sum_{i=1}^{q} l_i y_i\right)!} \sum_{k=0}^{m-\sum_{i=1}^{q} l_i y_i} \alpha^{-k} \frac{V_{n+m,j+k+\sum_{i=1}^{q} y_i}}{V_{n,j}} \\
&\quad \times \sum_{s=k}^{m-\sum_{i=1}^{q} l_i y_i} \binom{m - \sum_{i=1}^{q} l_i y_i}{s} (n - j\alpha)_{\left(m-\sum_{i=1}^{q} l_i y_i - s\right)} \mathscr{C}(s, k; \alpha) \\
&\quad \text{(by means of Equation (2.56) in Charalambides [4])} \\
&= \sum_{y_1=1}^{r_1} \cdots \sum_{y_q=1}^{r_q} \prod_{i=1}^{q} S(r_i, y_i) \prod_{i=1}^{q} \left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{y_i} \\
&\quad \times \frac{m!}{\left(m - \sum_{i=1}^{q} l_i y_i\right)!} \sum_{k=0}^{m-\sum_{i=1}^{q} l_i y_i} \alpha^{-k} \frac{V_{n+m,j+k+\sum_{i=1}^{q} y_i}}{V_{n,j}} \mathscr{C}\left(m - \sum_{i=1}^{q} l_i y_i, k; \alpha, -n + j\alpha\right),
\end{aligned}$$

with $\mathscr{C}(n, k; \alpha)$ being the generalized factorial coefficient. Eq. (24) follows from the last expression by the relation between the mixed moment of order $\mathbf{r}$ and the mixed falling factorial moment of order $\mathbf{r}$. See Charalambides [4] for details. □

Proof of Corollary 3. We present a sketch of the proof. Of course the starting point is Eq. (24). In particular, we set $q = \tau$ and, according to (4), we set

$$\frac{V_{n+m,j+k+\sum_{i=1}^{\tau} r_i}}{V_{n,j}} = \frac{\alpha^{k+\sum_{i=1}^{\tau} r_i}}{(\theta + n)_{(m)}} \left(\frac{\theta}{\alpha} + j\right)_{\left(\sum_{i=1}^{\tau} r_i\right)} \left(\frac{\theta}{\alpha} + j + \sum_{i=1}^{\tau} r_i\right)_{(k)}.$$

The sum over k in the resulting expression is solved by standard algebra on rising factorials and by means of Equation (2.49) in Charalambides [4]. Then (26) follows by setting $l_i = i$ for $i = 1, \ldots, \tau$, (27) follows by setting $\tau = 2$ and $r_1 = r_2 = 1$, and (23) follows by applying the Stirling approximation, for large m. □

Proof of Theorem 4. We provide a sketch of the proof because it combines arguments applied in the proofs of Theorems 2 and 3. Using the definition of $M_{l,m}^{(n)}$ in (19),

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{q} (M_{l_i,m}^{(n)})^{r_i}\right] &= \sum_{s=0}^{m} \sum_{k=0}^{s} \mathbb{P}[K_m^{(n)} = k, L_m^{(n)} = s \mid A_n(j, \mathbf{n})] \\
&\quad \times \mathbb{E}\left[\prod_{i=1}^{q} \left(\sum_{t=1}^{K_n} \mathbb{1}_{\{l_i\}}(N_t + S_t) + \sum_{t=1}^{K_m^{(n)}} \mathbb{1}_{\{l_i\}}(S_{K_n+t}) \mid A_{n,m}(j, \mathbf{n}, s, k)\right)^{r_i}\right],
\end{aligned} \tag{A.16}$$

where the distribution of $(K_m^{(n)}, L_m^{(n)}) \mid A_n(j, \mathbf{n})$ in (A.11) is given by (A.6). We proceed along lines similar to the proof of Theorem 4 in Favaro et al. [12]. In particular, we apply the Vandermonde identity to expand the power function on the right-hand side of (A.16) and, combining (A.8) and (A.13), we have

$$\begin{aligned}
&\left(\sum_{t=1}^{K_n} \mathbb{1}_{\{l_i\}}(N_t + S_t) + \sum_{t=1}^{K_m^{(n)}} \mathbb{1}_{\{l_i\}}(S_{K_n+t})\right)^{r_i} \\
&= \sum_{x_i=0}^{r_i} \sum_{y_i=0}^{r_i - x_i} S(r_i, x_i + y_i)(x_i + y_i)! \prod_{t=1}^{x_i} \mathbb{1}_{\{l_i\}}(N_{c_{i,t}} + S_{c_{i,t}}) \prod_{t=1}^{y_i} \mathbb{1}_{\{l_i\}}(S_{K_n+d_{i,t}}).
\end{aligned} \tag{A.17}$$

According to (A.17), and by means on standard algebraic manipulations, we can write

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=1}^{q}\left(\sum_{t=1}^{K_n}\mathbb{1}_{\{l_i\}}(N_t+S_t)+\sum_{t=1}^{K_m^{(n)}}\mathbb{1}_{\{l_i\}}(S_{K_n+t})\mid A_{n,m}(j,\mathbf{n},s,k)\right)^{r_i}\right]\\
&=\sum_{x_1=0}^{r_1}\cdots\sum_{x_q=0}^{r_q}\sum_{y_1=0}^{r_1-x_1}\cdots\sum_{y_q=0}^{r_q-x_q}\prod_{i=1}^{q}S(r_i,x_i+y_i)(x_i+y_i)!\\
&\quad\times\sum_{\mathbf{c}^{(\mathbf{x})}\in\mathcal{C}_{\mathbf{x}}}\sum_{\mathbf{d}^{(\mathbf{y})}\in\mathcal{B}_{\mathbf{y}}}\mathbb{P}[S_{\mathbf{c}_1^{(x_1)}}=l_1\mathbf{1}_{x_1}-\mathbf{n}_{\mathbf{c}_1^{(x_1)}},\ldots,S_{\mathbf{c}_q^{(xq)}}=l_q\mathbf{1}_{x_q}-\mathbf{n}_{\mathbf{c}_q^{(xq)}},\\
&\quad S^*_{\mathbf{d}_1^{(y_1)}}=l_1\mathbf{1}_{y_1},\ldots,S^*_{\mathbf{d}_q^{(yq)}}=l_q\mathbf{1}_{y_q}-\mathbf{n}_{\mathbf{d}_q^{(yq)}}\mid A_{n,m}(j,\mathbf{n},s,k)].
\end{aligned}\tag{A.18}$$

As in the proofs of Theorems 2 and 3, the conditional probability in (A.18) is obtained from (A.1) and (A.2) in Lemma 1. Specifically, by combining (A.18) with (A.1) and (A.2) in Lemma 1 we can write the expected value in (A.18) as

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=1}^{q}\left(\sum_{t=1}^{K_n}\mathbb{1}_{\{l_i\}}(N_t+S_t)+\sum_{t=1}^{K_m^{(n)}}\mathbb{1}_{\{l_i\}}(S_{K_n+t})\mid A_{n,m}(j,\mathbf{n},s,k)\right)^{r_i}\right]\\
&=\sum_{x_1=0}^{r_1}\cdots\sum_{x_q=0}^{r_q}\sum_{y_1=0}^{r_1-x_1}\cdots\sum_{y_q=0}^{r_q-x_q}\prod_{i=1}^{q}S(r_i,x_i+y_i)(x_i+y_i)!\\
&\quad\times\sum_{\mathbf{c}^{(\mathbf{x})}\in\mathcal{C}_{\mathbf{x}}}\sum_{\mathbf{d}^{(\mathbf{y})}\in\mathcal{B}_{\mathbf{y}}}\frac{(m-s)!}{\left(m-s-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)!}\prod_{i=1}^{q}\prod_{t=1}^{x_i}\frac{(n_{c_{i,t}}-\alpha)_{(l_i-n_{c_{i,t}})}}{(l_i-n_{c_{i,t}})!}\\
&\quad\times\frac{\left(n-\sum_{i=1}^{q}|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|-\alpha\left(j-\sum_{i=1}^{q}x_i\right)\right)_{\left(m-s-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)\right)}}{(n-j\alpha)_{(m-s)}}\\
&\quad\times\frac{s!}{\left(s-\sum_{i=1}^{q}l_iy_i\right)!}\prod_{i=1}^{q}\prod_{t=1}^{y_i}\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\frac{\left(k-\sum_{i=1}^{q}y_i\right)!}{k!}\times\alpha^{\sum_{i=1}^{q}y_i}\frac{\mathscr{C}\left(s-\sum_{i=1}^{q}l_iy_i,\,k-\sum_{i=1}^{q}y_i;\,\alpha\right)}{\mathscr{C}(s,k;\alpha)}.
\end{aligned}\tag{A.19}$$

The proof in completed by marginalizing the last expression over $(K_m^{(n)}, L_m^{(n)})$ with respect to the conditional distribution in (A.6) and by means of standard algebra involving rising factorials. In particular, combining (A.16) with (A.19) and along lines similar to the last part of the proofs of Theorems 2 and 3,

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{q}(M_{l_i,m}^{(n)})^{r_i}\right]&=\sum_{x_1=0}^{r_1}\cdots\sum_{x_q=0}^{r_q}\sum_{y_1=0}^{r_1-x_1}\cdots\sum_{y_q=0}^{r_q-x_q}\prod_{i=1}^{q}S(r_i,x_i+y_i)\frac{(x_i+y_i)!}{y_i!}\\
&\quad\times\prod_{i=1}^{q}\left(\frac{(1-\alpha)_{(l_i-1)}}{l_i!}\right)^{y_i}\sum_{\mathbf{c}^{(\mathbf{x})}\in\mathcal{C}_{\mathbf{x}}}\prod_{i=1}^{q}\prod_{t=1}^{x_i}\frac{(n_{c_{i,t}}-\alpha)_{(l_i-n_{c_{i,t}})}}{(l_i-n_{c_{i,t}})!}\\
&\quad\times\frac{m!}{\left(m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)-\sum_{i=1}^{q}l_iy_i\right)!}\times\sum_{k=\sum_{i=1}^{q}y_i}^{m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)-\sum_{i=1}^{q}l_iy_i}\alpha^{-k}\frac{V_{n+m,j+k+\sum_{i=1}^{q}y_i}}{V_{n,j}}\\
&\quad\times\mathscr{C}\left(m-\sum_{i=1}^{q}(l_ix_i-|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|)-\sum_{i=1}^{q}l_iy_i,\,k;\,\alpha,\,-n+\sum_{i=1}^{q}|\mathbf{n}_{\mathbf{c}_i^{(x_i)}}|+\alpha\left(j-\sum_{i=1}^{q}x_i\right)\right)
\end{aligned}$$

with $\mathscr{C}(n,k;\alpha)$ being the generalized factorial coefficient. Eq. (29) follows from the last expression by the relation between the mixed moment of order $\mathbf{r}$ and the mixed falling factorial moment of order $\mathbf{r}$. See Charalambides [4] for details. □

Proof of Corollary 4. We present a sketch of the proof. Of course the starting point is Eq. (29). In particular, we set $q = \tau$ and, according to (4), we set

$$\frac{V_{n+m,j+k+\sum_{i=1}^{\tau}(r_i-x_i)}}{V_{n,j}} = \frac{\alpha^{k+\sum_{i=1}^{\tau}(r_i-x_i)}}{(\theta+n)_{(m)}} \left(\frac{\theta}{\alpha}+j\right)_{(\sum_{i=1}^{\tau}(r_i-x_i))} \left(\frac{\theta}{\alpha}+j+\sum_{i=1}^{\tau}(r_i-x_i)\right)_{(k)}.$$

The sum over k in the resulting expression is solved by standard algebra on rising factorials and by means of Equation (2.49) in Charalambides [4]. Then (30) follows by setting $l_i = i$ for $i = 1, \ldots, \tau$, (31) follows by setting $\tau = 2$ and $r_1 = r_2 = 1$, and (32) follows by applying the Stirling approximation, for large m. □

References

[1] S. Bacallado, S. Favaro, L. Trippa, Looking-backward probabilities for Gibbs-type exchangeable random partitions, Bernoulli (2013) in press.
[2] A. Cerquetti, Marginals of multivariate Gibbs distributions with applications in Bayesian species sampling, Electron. J. Stat. 7 (2013) 697–716.
[3] O. Cesari, The posterior distribution of rare variants in Gibbs-type species sampling models (M.Sc. thesis), University of Torino and Collegio Carlo Alberto, 2012. Available at http://web.econ.unito.it/favaro/CesariMScThesis.pdf.
[4] C.A. Charalambides, Combinatorial Methods in Discrete Distributions, Wiley, 2005.
[5] P. De Blasi, S. Favaro, A. Lijoi, R.H. Mena, I. Prünster, M. Ruggiero, Are Gibbs-type priors the most natural generalization of the Dirichlet process? IEEE Trans. Pattern Anal. Mach. Intell. (2014) in press.
[6] L. Devroye, Random variate generation for exponentially and polynomially tilted stable distributions, ACM Trans. Model. Comput. Simul. 19 (2009) 4.
[7] R.C. Elandt-Johnson, Probability Models and Statistical Methods in Genetics, John-Wiley and Sons Inc., 1971.
[8] W.J. Ewens, The sampling theory of selectively neutral alleles, Theor. Popul. Biol. 3 (1972) 87–112.
[9] W.J. Ewens, Mathematical Population Genetics, Springer, 2004.
[10] W.J. Ewens, S. Tavaré, The Ewens sampling formula, in: S. Kotz, C.B. Read, D.L. Banks (Eds.), Encyclopedia of Statistical Science, vol. 2 Update, Wiley, New York, 1998.
[11] S. Favaro, A. Lijoi, I. Prünster, A new estimator of the discovery probability, Biometrics 68 (2012) 1188–1196.
[12] S. Favaro, A. Lijoi, I. Prünster, Conditional formulae for Gibbs-type exchangeable random partitions, Ann. Appl. Probab. 23 (2013) 1721–1754.
[13] S. Favaro, A. Lijoi, I. Prünster, R.H. Mena, Bayesian nonparametric inference for species variety with a two parameter Poisson–Dirichlet process prior, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2009) 993–1008.
[14] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Statist. 1 (1973) 209–230.
[15] A. Gnedin, J. Pitman, Exchangeable Gibbs partitions and Stirling triangles, J. Math. Sci. 138 (2006) 5674–5685.
[16] M. Guindani, N. Sepulveda, C.D.M. Paulino, P. Müller, A Bayesian semiparametric model for the analysis of sequence counts data, J. R. Stat. Soc. Ser C. Appl. Stat. 63 (2014) 385–404.
[17] M. Hofert, Efficiently sampling nested Archimedean copulas, Comput. Statist. Data Anal. 55 (2011) 57–70.
[18] H. Ishwaran, L.F. James, Gibbs sampling methods for stick-breaking priors, J. Amer. Statist. Assoc. 96 (2001) 161–173.
[19] L.F. James, Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics, 2002. Preprint arXiv:math/0205093.
[20] L.F. James, Stick-breaking PG(α, ζ)-generalized Gamma processes, 2013. Preprint arXiv:math/13086570.
[21] P. Joyce, Tavaré, The distribution of rare alleles, J. Math. Biol. 33 (1995) 602–618.
[22] J.F.C Kingman, Random discrete distributions, J. R. Stat. Soc. Ser. B Stat. Methodol. 37 (1975) 1–22.
[23] J.F.C Kingman, The representation of partition structures, J. Lond. Math. Soc. 18 (1978) 374–380.
[24] N.M. Laird, C. Lange, The Fundamentals of Modern Statistical Genetics, Springer, 2010.
[25] A. Lijoi, R.H. Mena, I. Prünster, Bayesian nonparametric estimation of the probability of discovering a new species, Biometrika 94 (2007) 769–786.
[26] A. Lijoi, R.H. Mena, I. Prünster, A Bayesian nonparametric method for prediction in EST analysis, BMC Bioinform. 8 (2007) 339.
[27] A. Lijoi, R.H. Mena, I. Prünster, Controlling the reinforcement in Bayesian nonparametric mixture models, J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2007) 715–740.
[28] A. Lijoi, I. Prünster, Models beyond the Dirichlet process, in: N.L. Hjort, C.C. Holmes, P. Müller, S.G. Walker (Eds.), Bayesian Nonparametrics, Cambridge University Press, Cambridge, 2010.
[29] A. Lijoi, I. Prünster, S.G. Walker, Bayesian nonparametric estimators derived from conditional Gibbs structures, Ann. Appl. Probab. 18 (2008) 1519–1547.
[30] A.E. Magurran, Ecological Diversity and Its Measurement, Princeton University Press, 1988.
[31] A.E. Magurran, Measuring Biological Diversity, Wiley-Blackwell, 2003.
[32] A.E. Magurran, Biological Diversity: Frontiers in Measurement and Assessment, Oxford University Press, 2011.
[33] M. Perman, J. Pitman, M. Yor, Size-biased sampling of Poisson point processes and excursions, Probab. Theory Related Fields 92 (1992) 21–39.
[34] J. Pitman, Exchangeable and partially exchangeable random partitions, Probab. Theory Related Fields 102 (1995) 145–158.
[35] J. Pitman, Poisson–Kingman partitions, in: D.R. Goldstein (Ed.), Science and Statistics: A Festschrift for Terry Speed, in: Lecture Notes Monograph Series, vol. 40, IMS, Beachwood, OH, 2003, pp. 1–34.
[36] J. Pitman, Combinatorial stochastic processes., in: Ecole d'Eté de Probabilités de Saint-Flour XXXII, in: Lecture Notes in Mathematics, N. 1875, Springer, New York, 2006.
[37] J. Pitman, M. Yor, The two parameter Poisson–Dirichlet distribution derived from a stable subordinator, Ann. Probab. 27 (1997) 1870–1902. Institute of Mathematical Statistics, Hayward.
[38] J. Quackenbush, J. Cho, D. Lee, F. Liang, I. Holt, S. Karamycheva, B. Parvizi, G. Pertea, R. Sultana, J. White, The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species, Nucl. Acids Res. 29 (2000) 159–164.
[39] M.B. Usher, Biological Management and Conservation, Chapman and Hall, 1973.