Bernoulli **21**(1), 2015, 1–37
DOI: 10.3150/13-BEJ559

Looking-backward probabilities for Gibbs-type exchangeable random partitions

SERGIO BACALLADO[1], STEFANO FAVARO[2,4] and LORENZO TRIPPA[3]

[1]*Department of Statistics, Stanford University, Sequoia Hall, Stanford, CA 94305, USA. E-mail: sergio.bacallado@gmail.com*
[2]*Department of Economics and Statistics, University of Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy. E-mail: stefano.favaro@unito.it*
[3]*Harvard School of Public Health and Dana-Faber Cancer Institute, 450 Brookline Avenue CLSB 11039 Boston, MA 02215, USA. E-mail: ltrippa@jimmy.harvard.edu*
[4]*Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy*

Gibbs-type random probability measures and the exchangeable random partitions they induce represent the subject of a rich and active literature. They provide a probabilistic framework for a wide range of theoretical and applied problems that are typically referred to as species sampling problems. In this paper, we consider the class of looking-backward species sampling problems introduced in Lijoi *et al.* (*Ann. Appl. Probab.* **18** (2008) 1519–1547) in Bayesian nonparametrics. Specifically, given some information on the random partition induced by an initial sample from a Gibbs-type random probability measure, we study the conditional distributions of statistics related to the old species, namely those species detected in the initial sample and possibly re-observed in an additional sample. The proposed results contribute to the analysis of conditional properties of Gibbs-type exchangeable random partitions, so far focused mainly on statistics related to those species generated by the additional sample and not already detected in the initial sample.

Keywords: Bayesian nonparametrics; conditional random partitions; Ewens–Pitman sampling model; Gibbs-type exchangeable random partitions; looking-backward probabilities; species diversity; species sampling problems

1. Introduction

Let $\mathbb{X}$ be a complete and separable metric space equipped with the Borel σ-algebra $\mathscr{X}$, and let $(X_i)_{i\geq 1}$ be an exchangeable sequence of $\mathbb{X}$-valued random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. According to the celebrated de Finetti's representation theorem there exists a random probability measure $\tilde{P}$ on $\mathbb{X}$ such that, conditionally on $\tilde{P}$, the random variables $(X_i)_{i\geq 1}$ are independent and identically distributed according to $\tilde{P}$, that is,

$$\begin{aligned} X_i \,|\, \tilde{P} &\overset{\text{i.i.d.}}{\sim} \tilde{P}, \\ \tilde{P} &\sim \Pi. \end{aligned}$$

The distribution Π is commonly known as the de Finetti probability measure of $(X_i)_{i\geq 1}$ and it takes on the interpretation of the prior distribution in Bayesian nonparametrics. In the present

paper, we consider almost surely discrete random probability measures, namely $\tilde{P}$ is such that $\Pi[\tilde{P} \in \mathscr{D}] = 1$, where $\mathscr{D}$ stands for the set of discrete probability measures on $(\mathbb{X}, \mathscr{X})$.

If $\tilde{P}$ is discrete almost surely, we expect ties in a sample $(X_1, \ldots, X_n)$ from $\tilde{P}$; that is, we expect $K_n \leq n$ distinct observations with frequencies $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$ satisfying $\sum_{1 \leq i \leq K_n} N_i = n$. Accordingly, the sample induces a random partition of the set $\{1, \ldots, n\}$, 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 symmetric function corresponding to the probability of any particular partition of $\{1, \ldots, n\}$ having j distinct blocks with frequencies $(n_1, \ldots, n_j)$. This function is known as the exchangeable partition probability function (EPPF), a concept introduced in [17] as a development of earlier results in [12]. The EPPF can be specified for every $n \geq 1$ and $1 \leq j \leq n$ either via the exchangeable sequence $(X_i)_{i \geq 1}$ or by defining a random partition of $\mathbb{N}$. In the latter case, the distribution of the random partition must satisfy certain consistency conditions and a symmetry property that guarantees exchangeability. See [19] and references therein for a comprehensive account on EPPFs.

Exchangeable random partitions play an important role in a variety of research areas. In population genetics, models for exchangeable random partitions are useful for describing the configurations of a sample of genes into a number of distinct allelic types. See [6] and references therein. In machine learning, probabilistic models for linguistic applications are often based on clustering structures for collections of words in documents. See, for example, [22] and [21] for a review. In Bayesian nonparametrics, exchangeable random partitions are commonly employed at the latent level of complex hierarchical mixture models. See [15] and references therein for a review. Other areas of application include storage problems, excursion theory, combinatorics, number theory and statistical physics. Broadly speaking, exchangeable random partitions and their associated EPPFs provide a flexible probabilistic framework for a wide range of theoretical and applied problems that are typically referred to as species sampling problems, namely problems concerning a population composed of individuals belonging to different species. Indeed, the number of partition blocks K_n take on the interpretation of the number of distinct species in the sample $(X_1, \ldots, X_n)$ and the N_i's are the corresponding species frequencies. Given the relevance and intuitiveness of such a framework, throughout the paper we will resort to the species metaphor.

The main object of our investigation is the class of Gibbs-type exchangeable random partitions. These are random partitions which arise by sampling from a random probability measure, say of Gibbs-type, here denoted by $\tilde{P}_G$. See [18] for details. Introduced in [10] these exchangeable random partitions represent the subject of a rich and active literature. A recent development, first proposed in [16], is the study of their conditional properties. This study consists in evaluating, conditional on some information about the random partition induced by an initial sample $(X_1, \ldots, X_n)$ from $\tilde{P}_G$, the distribution of certain statistics of an additional sample $(X_{n+1}, \ldots, X_{n+m})$. In particular, in [16] the main focus is on the conditional distributions of statistics related to the *new* species, namely those species generated by the additional sample and not coinciding with species already detected in the initial sample. A representative example is given by the distribution of the number of new distinct species generated by $(X_{n+1}, \ldots, X_{n+m})$, conditional on the information of both the number of distinct species in $(X_1, \ldots, X_n)$ and their corresponding frequencies. See [8] for a generalization to the number of new distinct species with a certain frequency of interest. As shown in [8,13] and [16] these

conditional distributions have direct applications in Bayesian nonparametric analysis of species sampling problems arising in ecology and genomics. We refer to [3,4,7] and [11] for other contributions at the interface between Bayesian nonparametrics and Gibbs-type exchangeable random partitions.

Many problems in the conditional analysis of Gibbs-type exchangeable random partitions remain unresolved. For instance, [16] pointed out the practical interest in the conditional distributions of statistics related to the *old* species, namely those species detected in the initial sample and possibly re-observed in the additional sample. Two illustrative examples are given in Proposition 4 of [16] and in Theorem 3 of [8]. In general the class of species sampling problems concerning old species has been referred to as *looking-backward* and it represent the focus of the present paper. We study two novel, and practically applicable, looking-backward species sampling problems. In particular, we derive

(i) the conditional distribution of the number of old distinct species re-observed in $(X_{n+1}, \ldots, X_{n+m})$, given complete or incomplete information on the random partition induced by $(X_1, \ldots, X_n)$;
(ii) the conditional distribution of the number of old distinct species re-observed with a specific frequency of interest in $(X_{n+1}, \ldots, X_{n+m})$, given complete or incomplete information on the random partition induced by $(X_1, \ldots, X_n)$.

Specifically, by complete information we refer jointly to the number of distinct species in $(X_1, \ldots, X_n)$ and their frequencies, whereas by incomplete information we refer solely to the number of distinct species in $(X_1, \ldots, X_n)$. Besides the sets of complete and incomplete information, we also consider almost-complete information. This information refers jointly to the number of distinct species in $(X_1, \ldots, X_n)$ and a subset of their corresponding frequencies.

The present paper broadens the scope of previous literature on conditional distributions for Gibbs-type exchangeable random partitions, by investigating in depth some statistics related to old species. In the framework of Gibbs-type exchangeable random partitions, looking-backward problems create a distinction between conditioning on complete, incomplete and almost complete information, which to the best of our knowledge has not been dealt with explicitly in previous studies. We expect the results introduced here to have an impact in the analysis of Bayesian nonparametric models for species sampling problems, which have acquired increasingly complex forms in recent years to meet the demands of scientific applications. The paper is structured as follows. Section 2 recalls the definition of Gibbs-type exchangeable random partition and introduces preliminary results relevant to the analysis of their conditional structure. Section 3 deals with the looking-backward species sampling problems (i) and (ii) in the general case of Gibbs-type exchangeable random partitions and in the special case of the celebrated Ewens–Pitman sampling model. The context of almost-complete information is also dealt with in Section 3. Section 4 contains some numerical illustrations of the present results. Proofs are deferred to the Appendix.

2. Preliminaries and main definitions

Gibbs-type exchangeable random partitions were introduced in [10] and further investigated in [18]. This class of exchangeable random partitions is characterized by an EPPF with a product

form, a feature which is crucial for mathematical tractability and, in particular, facilitates intuition. Let $\mathcal{D}_{n,j} = \{(n_1, \ldots, n_j) : n_i \geq 1 \text{ and } \sum_{i=1}^{j} n_i = n\}$ be the set of the partitions of $n \geq 1$ into $j \leq n$ positive integers. Moreover, for any $x > 0$ and any positive integer n, we denote by $(x)_{n\uparrow 1}$ and $(x)_{n\downarrow 1}$ the rising factorials and falling factorials, respectively.

Definition 2.1. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by $\tilde{P}_G$. Then, the exchangeable random partition induced by $(X_i)_{i\geq 1}$ is said of Gibbs-type and it is characterized by an EPPF of the form*

$$p_j^{(n)}(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1-\sigma)_{(n_i-1)\uparrow 1}, \tag{2.1}$$

for $\sigma < 1$ and nonnegative weights $(V_{n,j})_{j\leq n, n\geq 1}$ satisfying the recursion $V_{n,j} = V_{n+1,j+1} + (n - j\sigma)V_{n+1,j}$, with $V_{1,1} = 1$.

Gibbs-type exchangeable random partitions are completely specified by the parameter $\sigma < 1$ and the collection of weights $(V_{n,j})_{j\leq n,n\geq 1}$ satisfying a backward recursion. Note that Definition 2.1 also provides the distribution of the number K_n of distinct species in a sample of size n from $\tilde{P}_G$, that is,

$$\mathbb{P}[K_n = j] = V_{n,j} \frac{\mathscr{C}(n, j; \sigma)}{\sigma^j}, \tag{2.2}$$

with $\mathscr{C}(n, j; \sigma)$ being the so-called generalized factorial coefficient. We refer to [2] for details. The next example recalls the Ewens–Pitman sampling model, a noteworthy example of Gibbs-type exchangeable random partition introduced in [17] and generalizing the celebrated Ewens sampling model in [5]. See [1] and references therein for a comprehensive account on the Ewens sampling model. Another notable Gibbs-type exchangeable random partition, still related to the Ewens–Pitman sampling model, has been recently introduced and investigated in [9].

Example 2.1. For any $\sigma \in (0, 1)$ and $\theta > -\sigma$, the Ewens–Pitman sampling model is a Gibbs-type exchangeable random partition with weights $(V_{n,j})_{j\leq n,n\geq 1}$ of the following form

$$V_{n,j} = \frac{\prod_{i=0}^{j-1}(\theta + i\sigma)}{(\theta)_{n\uparrow 1}}. \tag{2.3}$$

The Ewens sampling model with parameter $\theta > 0$ is recovered from the Ewens–Pitman sampling model by letting $\sigma \to 0$. See, for example, [17] and [20] for details and further developments.

The recursion in Definition 2.1, for a fixed σ, 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 [10] describes the extreme points of such a convex set. For any $n \geq 1$ let

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

For every Gibbs-type exchangeable random partition there exists a positive and almost surely finite random variable S_σ such that

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

as $n \to +\infty$, and such that a Gibbs-type exchangeable random partition is a unique mixture over $\varkappa$ of extreme exchangeable random partitions for which $S_\sigma = \varkappa$ almost surely. For $\sigma \in (-\infty, 0)$ the extremes are Ewens–Pitman sampling models with parameter $(\sigma, -\sigma\varkappa)$; for $\sigma = 0$ the extremes are Ewens sampling models with parameter $\varkappa \geq 0$; for $\sigma \in (0, 1)$ the Ewens–Piman sampling models are not extremes. See Section 6.1 in [18] for details on S_σ.

A generalization of Definition 2.1 has been recently introduced in [16] to study conditional properties of Gibbs-type exchangeable random partitions. To recall this generalization a few quantities, analogous to those describing the random partition induced by an initial sample $(X_1, \ldots, X_n)$ from $\tilde{P}_G$, need to be introduced. Let $X_1^*, \ldots, X_{K_n}^*$ be the labels identifying the K_n distinct species detected in the initial sample and, for any $m > 1$, define

$$L_m^{(n)} = \sum_{i=1}^{m} \prod_{j=1}^{K_n} \mathbb{1}_{\{X_j^*\}^C}(X_{n+i}) \tag{2.4}$$

as the number of observations in an additional sample $(X_{n+1}, \ldots, X_{n+m})$ not coinciding with any of the K_n distinct species. Denote by $K_m^{(n)}$ the number of new distinct species generated by these $L_m^{(n)}$ observations and by $X_{K_n+1}^*, \ldots, X_{K_n+K_m^{(n)}}^*$ their corresponding identifying labels. Therefore,

$$\mathbf{M}_{L_m^{(n)}} = (M_1, \ldots, M_{K_m^{(n)}}),$$

with

$$M_i = \sum_{j=1}^{m} \mathbb{1}_{\{X_{K_n+i}^*\}}(X_{n+j}) \tag{2.5}$$

for $i = 1, \ldots, K_m^{(n)}$, are the frequencies of the new $K_m^{(n)}$ distinct species detected among the $L_m^{(n)}$ observations of the additional sample. Analogously,

$$\mathbf{S}_{m-L_m^{(n)}} = (S_1, \ldots, S_{K_n}),$$

with

$$S_i = \sum_{j=1}^{m} \mathbb{1}_{\{X_i^*\}}(X_{n+j}), \tag{2.6}$$

corresponds to number of observations, among the $m - L_m^{(n)}$ observations of the additional sample, coinciding with the ith distinct old species detected in the initial sample, for $i = 1, \ldots, K_n$. As pointed out in [13], from a Bayesian nonparametric perspective the joint conditional distribution of the random variables (2.4), (2.5), (2.6) and $K_m^{(n)}$, given $(X_1, \ldots, X_n)$, can be interpreted as the posterior counterpart of the EPPF (2.1). This then provides a natural framework for Bayesian nonparametric analysis of species sampling problem.

In [16], the main focus is on conditional distributions of statistics related to the new species generated by $(X_{n+1}, \ldots, X_{n+m})$. For instance, by suitably marginalizing the joint conditional distribution of the random variables (2.4), (2.5), (2.6) and $K_m^{(n)}$, given $(X_1, \ldots, X_n)$, one obtains the conditional distribution of the number of new distinct species, namely

$$\mathbb{P}\big[K_m^{(n)} = k | K_n = j, \mathbf{N}_n = \mathbf{n}\big] = \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m, k; \sigma, -n + j\sigma)}{\sigma^k} \tag{2.7}$$

with $\mathscr{C}(n, j; \sigma, \rho)$ being the so-called noncentral generalized factorial coefficient. We refer to [2] for details. Accordingly, the Bayesian nonparametric estimator, under quadratic loss function, of the number of new distinct species generated by the additional sample coincides with

$$\mathcal{K}_m^{(n)} = \mathbb{E}\big[K_m^{(n)} | K_n = j, \mathbf{N}_n = \mathbf{n}\big] = \mathbb{E}\big[K_m^{(n)} | K_n = j\big]. \tag{2.8}$$

We refer to [3,13,14] and [16] for applications of (2.7) and (2.8), under the choice of $V_{n,j}$ in (2.3), to Bayesian nonparametric inference for species variety in genetic experiments. As a generalization of (2.7), Theorem 3 in [8] provides the conditional distribution, given $(X_1, \ldots, X_n)$, of

$$\sum_{i=1}^{K_n} \mathbb{1}_{\{N_i + S_i = l\}} + \sum_{i=1}^{K_m^{(n)}} \mathbb{1}_{\{M_i = l\}}, \tag{2.9}$$

for any $l = 1, \ldots, n + m$. In words, (2.9) corresponds to the number of distinct species with frequency l generated by $(X_{n+1}, \ldots, X_{n+m})$. The conditional expected value of (2.9), given $(X_1, \ldots, X_n)$, provides the Bayesian nonparametric estimator, under quadratic loss function, of the number of distinct species with frequency l generated by the additional sample.

3. Two looking-backward probabilities

Before presenting our results, it is worth stating the fundamental difference between looking-backward species sampling problems and the species sampling problems investigated in [16].

A common feature of the conditional distributions introduced in [16] is their independence from the information on the frequencies $\mathbf{N}_n$ induced by the initial sample $(X_1, \ldots, X_n)$. As a representative example, note that the distribution (2.7) satisfies the following identity

$$\mathbb{P}\big[K_m^{(n)} = k | K_n = j, \mathbf{N}_n = \mathbf{n}\big] = \mathbb{P}\big[K_m^{(n)} = k | K_n = j\big].$$

Such a property of independence characterizes all the statistics concerning the new species in the additional sample $(X_{n+1}, \ldots, X_{n+m})$. Indeed, since (2.6) does not contain any information on new species, the conditional distributions of these statistics can be obtained from the joint conditional distribution of the random variables (2.4), (2.5) and $K_m^{(n)}$, given $(X_1, \ldots, X_n)$. In Proposition 1 of [13], this joint conditional distribution is shown to be independent of $\mathbf{N}_n$. Hence, K_n is a sufficient statistic for the species sampling problems discussed in [16].

Differently, the conditional distributions of statistics concerning old species depend on the information of both the number K_n of distinct species and the corresponding frequencies $\mathbf{N}_n$. This is to say that, letting $T_m^{(n)}$ be a statistic related to old species, in most cases, one obtains

$$\mathbb{P}\big[T_m^{(n)} \in \cdot | K_n = j, \mathbf{N}_n = \mathbf{n}\big] \neq \mathbb{P}\big[T_m^{(n)} \in \cdot | K_n = j\big]. \tag{3.1}$$

As an example, the distribution of (2.9) satisfies (3.1). See Theorem 3 of [8] for details. See also Proposition 4 in [16] for another example. According to (3.1), the analysis of the looking-backward species sampling problems naturally leads to consider at least two sets of information on the random partition induced by $(X_1, \ldots, X_n)$: (i) a complete information, namely K_n and $\mathbf{N}_n$; (ii) an incomplete information, namely K_n. We also consider almost-complete information, namely K_n and a subset of $\mathbf{N}_n$. In the next subsections, we present and discuss the results of our paper. We focus on deriving the conditional distributions of two looking-backward statistics, given complete or incomplete information. This will be the subject of Section 3.1 and Section 3.2. The conditional distributions of these two statistics given almost-complete information can be derived through similar arguments applied when conditioning on incomplete information. This will be discussed in Section 3.3.

3.1. Probabilities of re-observing old species

In this section, we consider the distribution of the number of old distinct species that are re-observed in $(X_{n+1}, \ldots, X_{n+m})$, conditional on complete and incomplete information on the random partition induced by $(X_1, \ldots, X_n)$. Formally, in the context of complete information, we are interested in the random variable $R_m^{(n,j,\mathbf{n})}$ which is defined in distribution as

$$\mathbb{P}\big[R_m^{(n,j,\mathbf{n})} = x\big] = \mathbb{P}\left[\sum_{i=1}^{K_n} \mathbb{1}_{\{S_i>0\}} = x \,\middle|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right]. \tag{3.2}$$

In the context of incomplete information, we are interested in the random variable $\tilde{R}_m^{(n,j)}$ which is defined in distribution as

$$\mathbb{P}\big[R_m^{(n,j)} = x\big] = \mathbb{P}\left[\sum_{i=1}^{K_n} \mathbb{1}_{\{S_i>0\}} = x \,\middle|\, K_n = j\right]. \tag{3.3}$$

In the next theorem, we derive the factorial moments of the random variables in (3.2) and (3.3). By means of Theorem 1 in [8], we obtain (3.4). Accordingly, (3.5) follows from (3.4) by suitably marginalizing the frequencies $\mathbf{N}_n$. These moments then lead to the corresponding distributions by means of standard arguments involving probability generating functions.

Theorem 1. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by $\tilde{P}_G$. Then, for any integer $r \geq 1$ one has*

$$\begin{aligned}
&\mathbb{E}\big[\big(R_m^{(n,j,\mathbf{n})}\big)_{r\downarrow 1}\big] \\
&\quad = r! \sum_{v=0}^{r} \binom{j-v}{r-v} (-1)^v \\
&\qquad \times \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m,k;\sigma,-n+\sum_{i=1}^{v} n_{c_i}+(j-v)\sigma)}{\sigma^k}
\end{aligned} \tag{3.4}$$

and

$$\begin{aligned}
&\mathbb{E}\big[\big(R_m^{(n,j)}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{r!}{\mathscr{C}(n,j;\sigma)} \sum_{v=0}^{r} \binom{j-v}{r-v} (-1)^v \sum_{s=v}^{n-(j-v)} \binom{n}{s} \mathscr{C}(s,v;\sigma)\mathscr{C}(n-s,j-v;\sigma) \\
&\qquad \times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m,k;\sigma,-n+s+(j-v)\sigma)}{\sigma^k},
\end{aligned} \tag{3.5}$$

where $\mathcal{C}_{j,v}$ denotes the set of the v-combinations (without repetitions) of the elements $\{1,\ldots,j\}$.

The distributions of $R_m^{(n,j,\mathbf{n})}$ and $R_m^{(n,j)}$ are interpretable as the posterior distributions of the number of old distinct species that are re-observed in $(X_{n+1},\ldots,X_{n+m})$ given, respectively, complete and incomplete information on the random partition induced by $(X_1,\ldots,X_n)$. Accordingly, the Bayesian nonparametric estimators, under a quadratic loss function, coincide with the expected values of the random variables $R_m^{(n,j,\mathbf{n})}$ and $R_m^{(n,j)}$. An expression for these Bayesian nonparametric estimators, denoted by $\mathcal{R}_m^{(n,j,\mathbf{n})} = \mathbb{E}[R_m^{(n,j,\mathbf{n})}]$ and $\mathcal{R}_m^{(n,j)} = \mathbb{E}[R_m^{(n,j)}]$, is presented in the next corollary. See Proposition 1 and Proposition 2 for an expression of these estimators under the Ewens–Pitman sampling model.

Corollary 3.1. *The Bayesian nonparametric estimator of the number of old distinct species that are re-observed in an additional sample of size m, given complete information on $(X_1, \ldots, X_n)$, coincides with*

$$\mathcal{R}_m^{(n,j,\mathbf{n})} = j - \sum_{i=1}^{n} m_i \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m,k;\sigma,-n+i+(j-1)\sigma)}{\sigma^k}.$$

Moreover, given incomplete information on $(X_1, \ldots, X_n)$, the Bayesian nonparametric estimator coincides with

$$\mathcal{R}_m^{(n,j)} = j - \frac{1}{\mathscr{C}(n,j;\sigma)} \sum_{s=1}^{n-(j-1)} \binom{n}{s} \mathscr{C}(s,1;\sigma)\mathscr{C}(n-s,j-1;\sigma) \times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m,k;\sigma,-n+s+(j-1)\sigma)}{\sigma^k}.$$

Here $m_i \geq 0$ denotes the number of distinct species observed in the initial sample with frequency i.

The distributions of $R_m^{(n,j,\mathbf{n})}$ and $R_m^{(n,j)}$, under the Ewens–Pitman sampling model, are specified in the next propositions. We devote special attention to the Ewens–Pitman sampling model because it has proven suitable for inference in species sampling problems, particularly in genomics. See, for example, [13] and [8] for details. The corresponding results for the Ewens sampling model can be recovered by letting $\sigma \to 0$ and applying equation 2.63 in [2].

Proposition 1. *Under the Ewens–Pitman sampling model, the distribution of $R_m^{(n,j,\mathbf{n})}$ coincides with*

$$\begin{aligned} &\mathbb{P}\big[R_m^{(n,j,\mathbf{n})} = x\big] \\ &= \frac{1}{(\theta+n)_{m\uparrow 1}} (-1)^j \sum_{v=j-x}^{j} \binom{v}{j-x} (-1)^{v+x} \\ &\quad \times \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \left(\theta + n - \sum_{i=1}^{v} n_{c_i} + \sigma v\right)_{m\uparrow 1} \end{aligned} \tag{3.6}$$

and

$$\mathcal{R}_m^{(n,j,\mathbf{n})} = j - \frac{1}{(\theta+n)_{m\uparrow 1}} \sum_{i=1}^{n} m_i (\theta + n - i + \sigma)_{m\uparrow 1}. \tag{3.7}$$

The random variable $R_m^{(n,j,\mathbf{n})}$ assigns positive probability to any integer value x such that $0 \leq x \leq \min(j,m)$.

Proposition 2. *Under the Ewens–Pitman sampling model, the distribution of* $R_m^{(n,j)}$ *coincides with*

$$\begin{aligned}\mathbb{P}\big[R_m^{(n,j)} = x\big] &= \frac{1}{\mathscr{C}(n,\, j;\sigma)(\theta+n)_{m\uparrow 1}}(-1)^j \sum_{v=j-x}^{j} \binom{v}{j-x}(-1)^{v+x} \\ &\quad \times \sum_{s=v}^{n-(j-v)} \binom{n}{s}(\theta+n-s+v\sigma)_{m\uparrow 1}\mathscr{C}(s,\, v;\sigma)\mathscr{C}(n-s,\, j-v;\sigma)\end{aligned} \tag{3.8}$$

and

$$\begin{aligned}\mathcal{R}_m^{(n,j)} &= j - \frac{1}{\mathscr{C}(n,\, j;\sigma)(\theta+n)_{m\uparrow 1}} \\ &\quad \times \sum_{s=1}^{n-(j-1)} \binom{n}{s}(\theta+n-s+\sigma)_{m\uparrow 1}\mathscr{C}(s,\, 1;\sigma)\mathscr{C}(n-s,\, j-1;\sigma).\end{aligned} \tag{3.9}$$

The random variable $R_m^{(n,j)}$ *assigns positive probability to any integer value* x *such that* $0 \le x \le \min(j, m)$.

3.2. Probabilities of re-observing old species with a certain frequency

In this section, we consider the distribution of the number of old distinct species that are re-observed in $(X_{n+1}, \ldots, X_{n+m})$ with frequency $0 \le l \le m$, conditional on complete and incomplete information on the random partition induced by the initial observed sample $(X_1, \ldots, X_n)$. Note that the case $l = 0$ is of particular interest, representing the number of old distinct species that are not re-observed in the additional sample. Formally, in the context of complete information, we are interested in the random variable $R_{l,m}^{(n,j,\mathbf{n})}$ which is defined in distribution as

$$\mathbb{P}\big[R_{l,m}^{(n,j,\mathbf{n})} = x\big] = \mathbb{P}\left[\sum_{i=1}^{K_n} \mathbb{1}_{\{S_i=l\}} = x \,\middle|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right]. \tag{3.10}$$

In the context of incomplete information, we are interested in the random variable $R_{l,m}^{(n,j)}$ which is defined in distribution as

$$\mathbb{P}\big[R_{l,m}^{(n,j)} = x\big] = \mathbb{P}\left[\sum_{i=1}^{K_n} \mathbb{1}_{\{S_i=l\}} = x \,\middle|\, K_n = j\right]. \tag{3.11}$$

In the next theorem, we derive the factorial moments of the random variables in (3.10) and (3.11). The factorial moment (3.12) is obtained by a direct application of Theorem 1

in [8]. With regards to the factorial moment (3.13), this is obtained from (3.12) by suitably marginalizing the frequencies $\mathbf{N}_n$. Again, these factorial moments lead to the corresponding distributions by means of standard arguments involving probability generating functions.

Theorem 2. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by $\tilde{P}_G$. Then, for any $0 \leq l \leq m$ and any integer $r \geq 1$ one has*

$$\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j,\mathbf{n})}\big)_{r\downarrow 1}\big] \\
&\quad = r!\binom{m}{l,\ldots,l,m-rl} \sum_{\{c_1,\ldots,c_r\}\in\mathcal{C}_{j,r}} \prod_{i=1}^{r} (n_{c_i}-\sigma)_{l\uparrow 1} \\
&\qquad \times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m-rl,k;\sigma,-n+\sum_{i=1}^{r} n_{c_i}+(j-r)\sigma)}{\sigma^k}
\end{aligned} \tag{3.12}$$

and

$$\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j)}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{r!}{\mathscr{C}(n,j;\sigma)}\binom{m}{l,\ldots,l,m-rl}\big(-\sigma(1-\sigma)_{(l-1)\uparrow 1}\big)^r \\
&\qquad \times \sum_{s=r}^{n-(j-r)} \binom{n}{s} \mathscr{C}(s,r;\sigma-l)\mathscr{C}(n-s,j-r;\sigma) \\
&\qquad \times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m-rl,k;\sigma,-n+s+(j-r)\sigma)}{\sigma^k},
\end{aligned} \tag{3.13}$$

where $\mathcal{C}_{j,r}$ denotes the set of the r-combinations (without repetitions) of the elements $\{1,\ldots,j\}$.

Again, the distributions of $R_{l,m}^{(n,j,\mathbf{n})}$ and $R_{l,m}^{(n,j)}$ are interpretable as the posterior distributions of the number of old distinct species that are re-observed in $(X_{n+1},\ldots,X_{n+m})$ with frequency $0 \leq l \leq m$ given, respectively, complete and incomplete information on the random partition induced by $(X_1,\ldots,X_n)$. The corresponding Bayesian nonparametric estimators, denoted by $\mathcal{R}_{l,m}^{(n,j,\mathbf{n})} = \mathbb{E}[R_{l,m}^{(n,j,\mathbf{n})}]$ and $\mathcal{R}_{l,m}^{(n,j)} = \mathbb{E}[R_{l,m}^{(n,j)}]$, are specified in the next corollary. See Proposition 3 and Proposition 4 for an expression for these estimators under the Ewens–Pitman sampling model.

Corollary 3.2. *The Bayesian nonparametric estimator of the number of old distinct species that are re-observed, with frequency $0 \leq l \leq m$, in an additional sample of size m, given complete*

information on $(X_1, \ldots, X_n)$, *coincides with*

$$\mathcal{R}_{l,m}^{(n,j,\mathbf{n})} = \binom{m}{l} \sum_{i=1}^{n} m_i (i-\sigma)_{l\uparrow 1} \times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m-l,k;\sigma,-n+i+(j-1)\sigma)}{\sigma^k}.$$

Moreover, given incomplete information on $(X_1, \ldots, X_n)$ *the Bayesian nonparametric estimator coincides with*

$$\mathcal{R}_{l,m}^{(n,j)} = \frac{1}{\mathscr{C}(n,j;\sigma)} \binom{m}{l} \left(-\sigma(1-\sigma)_{(l-1)\uparrow 1}\right) \times \sum_{s=1}^{n-(j-1)} \binom{n}{s} \mathscr{C}(s,1;\sigma-l)\mathscr{C}(n-s,j-1;\sigma) \times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m-l,k;\sigma,-n+s+(j-1)\sigma)}{\sigma^k}.$$

Here $m_i \geq 0$ *denotes the number of distinct species observed in the initial sample with frequency* i.

Finally, the distributions of $R_m^{(n,j,\mathbf{n})}$ and $R_m^{(n,j)}$, under the Ewens–Pitman sampling model, are specified in the next propositions.

Proposition 3. *Under the Ewens–Pitman sampling model, for any* $0 \leq l \leq m$, *the distribution of* $R_{l,m}^{(n,j,\mathbf{n})}$ *coincides with*

$$\mathbb{P}\left[R_{l,m}^{(n,j,\mathbf{n})} = x\right] = \frac{1}{(\theta+n)_{m\uparrow 1}} \sum_{y=x}^{m} \binom{y}{y-x} (-1)^{y-x} \times \binom{m}{l,\ldots,l,m-yl} \sum_{\{c_1,\ldots,c_y\}\in\mathcal{C}_{j,y}} \prod_{i=1}^{y} (n_{c_i}-\sigma)_{l\uparrow 1} \left(\theta+n-\sum_{i=1}^{y} n_{c_i}+\sigma y\right)_{(m-yl)\uparrow 1} \tag{3.14}$$

and

$$\mathcal{R}_{l,m}^{(n,j,\mathbf{n})} = \frac{1}{(\theta+n)_{m\uparrow 1}} \binom{m}{l} \sum_{i=1}^{n} m_i (i-\sigma)_{l\uparrow 1} (\theta+n-i+\sigma)_{(m-l)\uparrow 1}. \tag{3.15}$$

The random variable $R_{l,m}^{(n,j,\mathbf{n})}$ *assigns positive probability to any integer value* x *such that* $0 \leq x \leq \min(j,m)$.

Proposition 4. *Under the Ewens–Pitman sampling model, for any* $0 \le l \le m$, *the distribution of* $R_{l,m}^{(n,j)}$ *coincides with*

$$\begin{aligned}
&\mathbb{P}\big[R_{l,m}^{(n,j)} = x\big] \\
&\quad = \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \sum_{y=x}^{m} \binom{y}{y-x} (-1)^{y-x} \\
&\quad \times \binom{m}{l,\ldots,l,m-yl} \big(-\sigma(1-\sigma)_{(l-1)\uparrow 1}\big)^{y} \\
&\quad \times \sum_{s=y}^{n-(j-y)} \binom{n}{s} (\theta+n-s+\sigma y)_{(m-yl)\uparrow 1} \mathscr{C}(s,y;\sigma-l)\mathscr{C}(n-s,j-y;\sigma)
\end{aligned} \tag{3.16}$$

and

$$\begin{aligned}
\mathcal{R}_{l,m}^{(n,j)} &= \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \binom{m}{l} \big(-\sigma(1-\sigma)_{(l-1)\uparrow 1}\big) \\
&\quad \times \sum_{s=1}^{n-(j-1)} \binom{n}{s} (\theta+n-s+\sigma)_{(m-l)\uparrow 1} \mathscr{C}(s,1;\sigma-l)\mathscr{C}(n-s,j-1;\sigma).
\end{aligned} \tag{3.17}$$

The random variable $R_{l,m}^{(n,j)}$ *assigns positive probability to any integer value* x *such that* $0 \le x \le \min(j,m)$.

3.3. Conditioning on almost-complete information

We now consider the distribution of the number of old distinct species that are re-observed in the additional sample $(X_{n+1},\ldots,X_{n+m})$, conditional on almost-complete information. This looking-backward species sampling problem can be seen as a generalization of the problems discussed above. For any integer $p \in \{1,\ldots,K_n\}$ let $\tau = \{\tau_1,\ldots,\tau_p\}$ be a collection of integers such that $1 \le \tau_1 < \cdots < \tau_p \le K_n$ and define the subset of p frequencies $\mathbf{N}_{\tau,n} = (N_{\tau_1},\ldots,N_{\tau_p})$. In the context of almost-complete information, we are interested in the random variables $R_m^{(n,j,\mathbf{n}_\tau)}$ and $R_{l,m}^{(n,j,\mathbf{n}_\tau)}$ which are defined in distribution as

$$\mathbb{P}\big[R_m^{(n,j,\mathbf{n}_\tau)} = x\big] = \mathbb{P}\left[\sum_{i=1}^{K_n} \mathbb{1}_{\{S_i>0\}} = x \,\middle|\, K_n = j, \mathbf{N}_{\tau,n} = \mathbf{n}_\tau\right] \tag{3.18}$$

and

$$\mathbb{P}\big[R_{l,m}^{(n,j,\mathbf{n}_\tau)} = x\big] = \mathbb{P}\left[\sum_{i=1}^{K_n} \mathbb{1}_{\{S_i=l\}} = x \,\middle|\, K_n = j, \mathbf{N}_{\tau,n} = \mathbf{n}_\tau\right]. \tag{3.19}$$

The following lemma is fundamental in determining the factorial moments of the random variables introduced in (3.18) and (3.19) and, accordingly, to derive the corresponding distributions.

Lemma 3.1. *Let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a Gibbs-type random probability measure $\tilde{P}_G$. For any integer $p \in \{1, \ldots, K_n\}$, denote by $\nu = \{\nu_1, \ldots, \nu_{K_n-p}\}$ the complement set of τ with $1 \leq \nu_1 < \cdots < \nu_{K_n-p} \leq K_n$ and define the subset of frequencies $\mathbf{N}_{\nu,n} := (N_{\nu_1}, \ldots, N_{\nu_{K_n-p}})$. Then*

$$\begin{aligned}&\mathbb{P}[\mathbf{N}_{\nu,n} = \mathbf{n}_\nu | K_n = j, \mathbf{N}_{\tau,n} = \mathbf{n}_\tau] \\ &= \frac{\sigma^{j-p}}{\mathscr{C}(n - \sum_{i=1}^p n_{\tau_i}, j-p; \sigma)} \frac{1}{(j-p)!} \binom{n - \sum_{i=1}^p n_{\tau_i}}{n_{\nu_1}, \ldots, n_{\nu_{j-p}}} \prod_{i=1}^{j-p} (1-\sigma)_{(n_{\nu_i}-1)\uparrow 1}.\end{aligned} \tag{3.20}$$

The random variable $\mathbf{N}_{\nu,n} = \mathbf{n}_\nu | (K_n = j, \mathbf{N}_{\tau,n} = \mathbf{n}_\tau)$ assigns positive probability to the set $\mathcal{D}_{n-\sum_{i=1}^p n_{\tau_i}, j-p}$.

The factorial moments of $R_m^{(n,j,\mathbf{n}_\tau)}$ and $R_{l,m}^{(n,j,\mathbf{n}_\tau)}$ are derived by means of Lemma 3.1 and along lines similar to the proof of Theorem 1 and Theorem 2, respectively. In particular, with regard to the factorial moments of the random variables in (3.18), one has

$$\begin{aligned}&\mathbb{E}\left[\left(R_m^{(n)}\right)_{r\downarrow 1} | K_n = j, \mathbf{N}_{\tau,n} = \mathbf{n}_\tau\right] \\ &= \frac{r!}{\mathscr{C}(n - \sum_{i=1}^p n_{\tau_i}, j-p; \sigma)} \sum_{\nu_1=0}^{r} \sum_{\nu_2=0}^{r} (-1)^{\nu_1+\nu_2} \binom{j-\nu_1-\nu_2}{r-\nu_1-\nu_2} \\ &\quad \times \sum_{\{d_1,\ldots,d_{\nu_1}\} \in \mathcal{C}_{p,\nu_1}} \sum_{s=\nu_2}^{n-\sum_{i=1}^p n_{\tau_i} - (j-p-\nu_2)} \binom{n - \sum_{i=1}^p n_{\tau_i}}{s} \\ &\quad \times \mathscr{C}(s, \nu_2; \sigma) \mathscr{C}\left(n - \sum_{i=1}^p n_{\tau_i} - s, j-p-\nu_2; \sigma\right) \\ &\quad \times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m, k; \sigma, -n + \sum_{i=1}^{\nu_1} n_{\tau_{d_i}} + (j-\nu_1-\nu_2)\sigma + s)}{\sigma^k}.\end{aligned} \tag{3.21}$$

We point out that (3.21) is a generalization of both the results stated in Theorem 1. Indeed, by setting $\tau = j$ in (3.21) one obtains (3.5), whereas by setting $p = j$ in (3.21) one obtains (3.4). With regard to the factorial moments of the random variables in (3.19), one has

$$\begin{aligned}&\mathbb{E}\left[\left(R_{l,m}^{(n)}\right)_{r\downarrow 1} | K_n = j, \mathbf{N}_{\tau,n} = \mathbf{n}_\tau\right] \\ &= \frac{r!}{\mathscr{C}(n - \sum_{i=1}^p n_{\tau_i}, j-p; \sigma)} \binom{m}{l, \ldots, l, m-rl} \sum_{v=0}^{r} \left(-\sigma(1-\sigma)_{(l-1)\uparrow 1}\right)^{r-v}\end{aligned}$$

$$\begin{aligned}
&\times \sum_{\{d_1,\dots,d_v\}\in\mathcal{C}_{p,v}} \prod_{i=1}^{v} (n_{\tau_{d_i}} - \sigma)_{l\uparrow 1} \\
&\times \sum_{s=r-v}^{n-\sum_{i=1}^{p} n_{\tau_i} - (j-p-(r-v))} \binom{n - \sum_{i=1}^{p} n_{\tau_i}}{s} \\
&\times \mathscr{C}(s, r-v; \sigma - l)\mathscr{C}\left(n - \sum_{i=1}^{p} n_{\tau_i} - s, j-p-(r-v); \sigma\right) \\
&\times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m - rl, k; \sigma, -n + \sum_{i=1}^{v} n_{\tau_{d_i}} + s + (j-r)\sigma)}{\sigma^k}.
\end{aligned} \tag{3.22}$$

Note that (3.22) includes as special cases both the results stated in Theorem 2. Indeed, by setting $\tau = j$ in (3.22) one obtains (3.13), whereas by setting $p = j$ in (3.22) one obtains (3.12).

4. Numerical illustrations

We can now apply the derived conditional results which are interpretable, from a Bayesian nonparametric standpoint, as estimators or predictions. The range of problems to be addressed can be delineated using the following hypothetical setting. A nineteenth century naturalist samples a number of marine species in an expedition to a remote island, reporting in his notebook the number of distinct species sampled and their frequencies. We are interested in estimating the abundance of a particular species observed at that point in time. If all the data in the notebook are available, the looking-backward estimators of Theorems 1 and 2 which condition on complete information can be applied to solve this problem. Now suppose that certain critical pages of the notebook are missing, and the only datum available is the number of distinct species in a sample of known size. This corresponds to the setting of incomplete information.

In a general application, the species could be words in a text, mutations of a gene in a population, or the names of newborns in a year. The availability of complete or incomplete information could be determined by constraints of the experimental method used or, in the case of a meta-analysis, restrictions of access to data. For example, techniques routinely used in biology provide indications about presence or absence of a particular species, say a particular bacterium or a genetic mutation of interest, but are not suitable for measuring the relative species abundance. The experimental techniques, in these cases, produce datasets with partial information.

We illustrate an application of the derived looking-backward estimators in a simulation study. Two thousand samples were simulated from the Ewens–Pitman sampling model with $\theta = 100$ and $\sigma = 0.5$. The top row of panels in Figure 1 show the conditional expectations of the number of re-observed species in an additional experiment with sample sizes ranging from 0 to

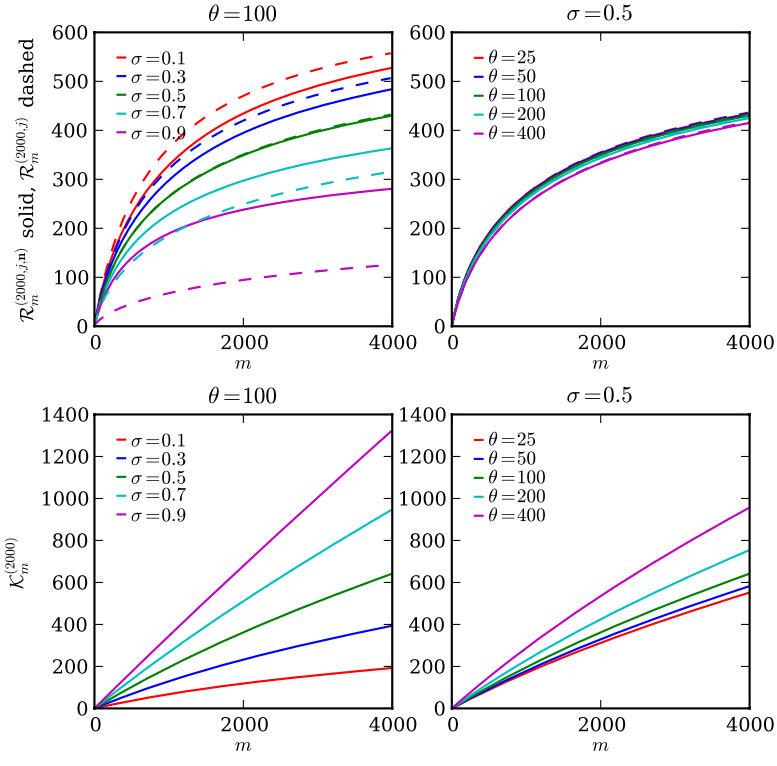

Figure 1. Estimators for the number of old and new distinct species observed as a function of the size m of the additional sample. An initial sample of $n = 2000$ steps was drawn from the Ewens–Pitman sampling model with $\theta = 100$ and $\sigma = 0.5$. The top panels show estimators for the number of old species under complete information, $\mathcal{R}_m^{(2000,j,\mathbf{n})}$, and incomplete information, $\tilde{\mathcal{R}}_m^{(2000,j)}$. The bottom panels show the estimator $\mathcal{K}_m^{(2000)}$ for the number of new species. The panels on the left show estimators computed under $\theta = 100$ and allowing σ to vary. The panels on the right show estimators computed under $\sigma = 0.5$ and allowing θ to vary.

4000. These two panels display discrepancies of the estimates under complete versus partial information and illustrate sensitivity to the choice of the parameters θ and σ. The estimates were computed across a range of possible prior parameters, including the true data distribution. Interestingly, the divergence between the two estimators depends more heavily on σ and is minimized when the parameter match those of the true data distribution. We refer to [13] for detailed arguments on practical selection of the prior parameters in this model. The second row of panels, in contrast, displays estimates for the number of new species in the additional sample. In this case the estimates are identical under complete and partial information.

Figure 2 considers simulated data that have not been sampled from the Ewens–Pitman sampling model. Here, the sample was generated from a Zeta distribution, whose power law behavior

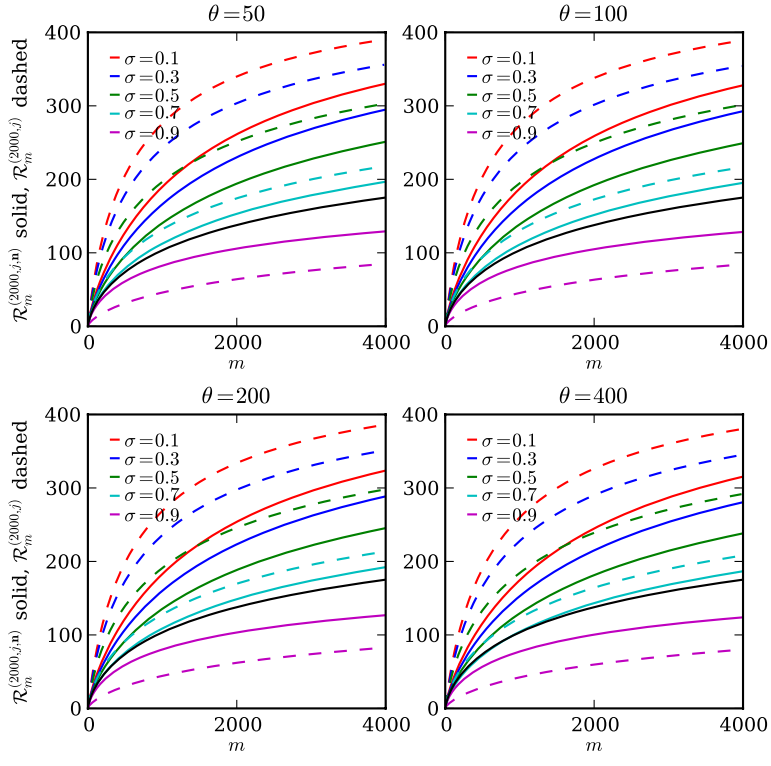

Figure 2. Estimators for the number of old distinct species observed as a function of the size m of the additional sample. An initial sample of $n = 2000$ steps was drawn from a zeta distribution with scale parameter 1.3. Each panel shows estimators computed with a fixed θ and allowing σ to vary. The black line in each figure shows the expected number of old distinct species in the sampling model.

is common in applications, and analyses were still performed using the Ewens–Pitman sampling model. Looking-backward estimators under complete and incomplete information are displayed for several prior parameters values. These are consistent with the relationship between the choice of the model parameters and the resulting conditional expectations shown in Figure 1. Figure 2 also displays (black line) the conditional expectations under the true zeta sampling model, assumed unknown to the investigator.

The simulations in Figure 1 were iterated, generating 1000 independent datasets of size $n = 2000$ from the Ewens–Pitman sampling model with $\theta = 100$ and $\sigma = 0.5$. Figure 3 shows the distribution of the estimator for the number of distinct old species re-observed in an additional sample of size 500. The blue and red histograms correspond to the estimator under complete and incomplete information, respectively. As expected, the estimators have the same mean but the estimator fit to complete information has slightly higher variance.

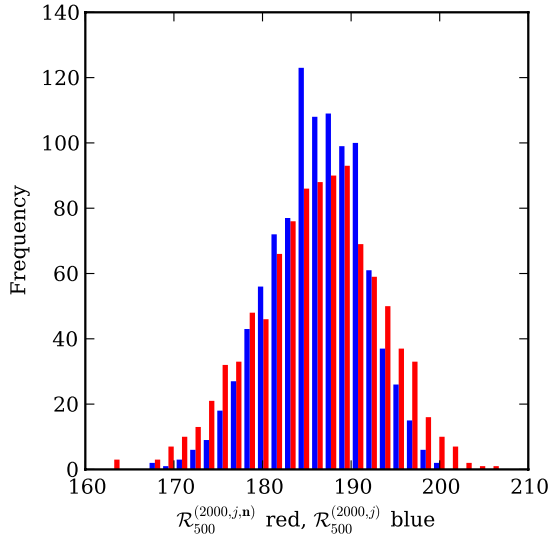

Figure 3. Histograms of the estimators for the number of old species under complete information, $\mathcal{R}_{500}^{(2000,j,\mathbf{n})}$, and incomplete information, $\tilde{\mathcal{R}}_{500}^{(2000,j)}$. To construct the histograms, these estimators were computed conditional on 1000 independent initial samples of length $n = 2000$ each, which were drawn from the Ewens–Pitman sampling model with $\theta = 100$ and $\sigma = 0.5$.

Appendix

A.1. Proofs of the results in Section 3.1

Proof of Theorem 1. With regard to the rth factorial moment of $R_m^{(n,j,\mathbf{n})}$, this is obtained by a direct application of Theorem 1 in [8]. Indeed, by means of the Vandermonde's identity one has

$$\left(R_m^{(n,j,\mathbf{n})}\right)_{r\downarrow 1} = \sum_{v=0}^{r} \binom{r}{v} (-1)^v (j-v)_{(r-v)\downarrow 1} \left(R_{0,m}^{(n,j,\mathbf{n})}\right)_{v\downarrow 1}. \tag{A.1}$$

Theorem 1 in [8] then leads to (3.4) by taking the expected value of both sides of (A.1). This completes the first part of the proof. With regard to rth factorial moment of the random variable $R_m^{(n,j)}$, by combining (3.4) with the distributions displayed in (2.1) and (2.2), we write

$$\begin{aligned}&\mathbb{E}\left[\left(R_m^{(n,j)}\right)_{r\downarrow 1}\right] \\ &\quad = \frac{\sigma^j}{\mathscr{C}(n,j;\sigma)} \sum_{v=0}^{r} \binom{r}{v} (-1)^v (j-v)_{(r-v)\downarrow 1}\end{aligned} \tag{A.2}$$

$$\times \frac{1}{j!} \sum_{(n_1,\ldots,n_j)\in\mathcal{D}_{n,j}} \binom{n}{n_1,\ldots,n_j} \prod_{i=1}^{j} (1-\sigma)_{(n_i-1)\uparrow 1}$$

$$\times v! \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m,k;\sigma,-n+\sum_{i=1}^{v} n_{c_i}+(j-v)\sigma)}{\sigma^k}$$

and prove that it coincides with (3.5). The proof is mainly devoted to solve the sums over the indexes $n_1,\ldots,n_j$ and $c_1,\ldots,c_v$. Once these sums are solved, then (3.5) follows by some algebra involving factorial numbers and noncentral generalized factorial coefficients. By means of equation 2.61 in [2], and using the fact that $\mathcal{C}_{j,v}$ has cardinality $\binom{j}{v}$, from (A.2) one has

$$\begin{aligned}
&\mathbb{E}\left[\left(R_m^{(n,j)}\right)_{r\downarrow 1}\right] \\
&= \frac{\sigma^j}{\mathscr{C}(n,j;\sigma)} \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{1}{\sigma^k} \sum_{v=0}^{r} \binom{r}{v} (-1)^v (j-v)_{(r-v)\downarrow 1} \\
&\quad \times \sum_{s_1=1}^{n-j+1} \sum_{s_2=1}^{n-j+1-(s_1-1)} \cdots \sum_{s_v=1}^{n-j+1-\sum_{i=1}^{v-1}(s_i-1)} \binom{n}{s_1,\ldots,s_v,n-\sum_{i=1}^{v} s_i} \qquad \text{(A.3)} \\
&\quad \times \prod_{i=1}^{v} (1-\sigma)_{(s_i-1)\uparrow 1} \\
&\quad \times \frac{1}{\sigma^{j-v}} \mathscr{C}\left(m,k;\sigma,-n+\sum_{i=1}^{v} s_i+(j-v)\sigma\right) \mathscr{C}\left(n-\sum_{i=1}^{v} s_i, j-v;\sigma\right).
\end{aligned}$$

In order to solve the nested sums over the indexes $s_1,\ldots,s_v$ in (A.3), we first deal with the sum over the index s_v and then we introduce a suitable recursive argument for solving the remaining sums over the indexes $s_1,\ldots,s_{v-1}$. First, recall that for any $x \geq 0$ and $0 \leq y \leq x$, for any $a > 0$, $b > 0$, $c > 0$ and for any real number d one has the following identity

$$\binom{y+c}{y} \mathscr{C}(x,y+c;d,a+b) = \sum_{j=y}^{x-c} \binom{x}{j} \mathscr{C}(j,y;d,a) \mathscr{C}(x-j,c;d,b). \qquad \text{(A.4)}$$

See Chapter 2 of [2] for details. Then, let us consider the sum over the index s_v in (A.3), that is,

$$\begin{aligned}
&\sum_{s_v=1}^{n-j+1-\sum_{i=1}^{v-1}(s_i-1)} \binom{n}{s_1,\ldots,s_v,n-\sum_{i=1}^{v} s_i} \prod_{i=1}^{v} (1-\sigma)_{(s_i-1)\uparrow 1} \\
&\quad \times \frac{1}{\sigma^{j-v}} \mathscr{C}\left(m,k;\sigma,-n+\sum_{i=1}^{v} s_i+(j-v)\sigma\right) \mathscr{C}\left(n-\sum_{i=1}^{v} s_i, j-v;\sigma\right)
\end{aligned}$$

$$= \frac{1}{\sigma^{j-v}} \binom{n}{s_1, \ldots, s_{v-1}, n - \sum_{i=1}^{v-1} s_i} \prod_{i=1}^{v-1} (1-\sigma)_{(s_i-1)\uparrow 1}$$

$$\times \sum_{s_v=1}^{n-j+1-\sum_{i=1}^{v-1}(s_i-1)} \binom{n - \sum_{i=1}^{v-1} s_i}{s_v} (1-\sigma)_{(s_v-1)\uparrow 1}$$

$$\times \mathscr{C}\left(m, k; \sigma, -n + \sum_{i=1}^{v-1} s_i + s_v + (j-v)\sigma\right) \mathscr{C}\left(n - \sum_{i=1}^{v-1} s_i - s_v, j-v; \sigma\right).$$

By a direct application of (A.4) to the coefficients $\mathscr{C}(m, k; \sigma, -n + \sum_{i=1}^{v} s_i + (j-v)\sigma)$ and $\mathscr{C}(n - \sum_{i=1}^{v} s_i, j-v; \sigma)$ we can write the last expression in the following expanded form

$$\frac{1}{\sigma^{j-v}} \binom{n}{s_1, \ldots, s_{v-1}, n - \sum_{i=1}^{v-1} s_i} \prod_{i=1}^{v-1} (1-\sigma)_{(s_i-1)\uparrow 1}$$

$$\times \sum_{t=k}^{m} \binom{m}{t} \mathscr{C}(t, k; \sigma) \sum_{l=j-v}^{n-\sum_{i=1}^{v-1} s_i - 1} \binom{n - \sum_{i=1}^{v-1} s_i}{l} \mathscr{C}\big(l, j-v; \sigma, -(j-v)\sigma\big)$$

$$\times \sum_{s_v=1}^{n-l-\sum_{i=1}^{v-1} s_i} \binom{n - l - \sum_{i=1}^{v-1} s_i}{s_v} (1-\sigma)_{(s_v-1)\uparrow 1}$$

$$\times \left(n - \sum_{i=1}^{v-1} s_i - s_v - (j-v)\sigma\right)_{(m-t)\uparrow 1} \big(-(j-v)\sigma\big)_{(n-l-\sum_{i=1}^{v-1} s_i - s_v)\uparrow 1}$$

(by the Vandermonde's identity to expand $(n - \sum_{i=1}^{v-1} s_i - s_v - (j-v)\sigma)_{(m-t)\uparrow 1}$)

$$= \frac{1}{\sigma^{j-v}} \binom{n}{s_1, \ldots, s_{v-1}, n - \sum_{i=1}^{v-1} s_i} \prod_{i=1}^{v-1} (1-\sigma)_{(s_i-1)\uparrow 1}$$

$$\times \sum_{t=k}^{m} \binom{m}{t} \mathscr{C}(t, k; \sigma) \sum_{h=0}^{m-t} \binom{m-t}{h} \big(-(j-v)\sigma\big)_{h\uparrow 1}$$

$$\times \sum_{l=j-v}^{n-\sum_{i=1}^{v-1} s_i - 1} \binom{n - \sum_{i=1}^{v-1} s_i}{l} (l)_{(m-t-h)\uparrow 1} \mathscr{C}\big(l, j-v; \sigma, -(j-v)\sigma\big)$$

$$\times \sum_{s_v=1}^{n-l-\sum_{i=1}^{v-1} s_i} \binom{n - l - \sum_{i=1}^{v-1} s_i}{s_v} (1-\sigma)_{(s_v-1)\uparrow 1} \big(-(j-v)\sigma + h\big)_{(n-l-\sum_{i=1}^{v-1} s_i - s_v)\uparrow 1}$$

(by Equation 2.56 in [2] to solve the sum over the index s_v)

$$
\begin{aligned}
&= \frac{1}{\sigma^{j-v}} \binom{n}{s_1, \ldots, s_{v-1}, n - \sum_{i=1}^{v-1} s_i} \prod_{i=1}^{v-1} (1-\sigma)_{(s_i-1)\uparrow 1} \\
&\quad \times \sum_{t=k}^{m} \binom{m}{t} \mathscr{C}(t, k; \sigma) \sum_{h=0}^{m-t} \binom{m-t}{h} \big(-(j-v)\sigma\big)_{h\uparrow 1} \\
&\quad \times \sum_{l=j-v}^{n-\sum_{i=1}^{v-1} s_i - 1} \binom{n - \sum_{i=1}^{v-1} s_i}{l} (l)_{(m-t-h)\uparrow 1} \mathscr{C}\big(l, j-v; \sigma, -(j-v)\sigma\big) \\
&\quad \times \frac{1}{\sigma} \mathscr{C}\left(n - l - \sum_{i=1}^{v-1} s_i, 1; \sigma, (j-v)\sigma - h\right)
\end{aligned}
$$

providing the solution for the innermost nested sum over the index s_v. Therefore, according to the last identity, the rth factorial moment of $R_m^{(n,j)}$ has the following reduced expression

$$
\begin{aligned}
&\mathbb{E}\big[\big(R_m^{(n,j)}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{\sigma^j}{\mathscr{C}(n, j; \sigma)} \sum_{k=0}^{m} \frac{V_{n+m, j+k}}{V_{n,j}} \frac{1}{\sigma^k} \sum_{v=0}^{r} \binom{r}{v} (-1)^v (j-v)_{(r-v)\downarrow 1} \\
&\qquad \times \sum_{s_1=1}^{n-j+1} \sum_{s_2=1}^{n-j+1-(s_1-1)} \cdots \sum_{s_{v-1}=1}^{n-j+1-\sum_{i=1}^{v-2}(s_i-1)} \binom{n}{s_1, \ldots, s_{v-1}, n - \sum_{i=1}^{v-1} s_i} \\
&\qquad \times \prod_{i=1}^{v-1} (1-\sigma)_{(s_i-1)\uparrow 1} \qquad \text{(A.5)} \\
&\qquad \times \sum_{t=k}^{m} \binom{m}{t} \mathscr{C}(t, k; \sigma) \sum_{h=0}^{m-t} \binom{m-t}{h} \big(-(j-v)\sigma\big)_{h\uparrow 1} \\
&\qquad \times \sum_{l=j-v}^{n-\sum_{i=1}^{v-1} s_i - 1} \binom{n - \sum_{i=1}^{v-1} s_i}{l} (l)_{(m-t-h)\uparrow 1} \mathscr{C}\big(l, j-v; \sigma, -(j-v)\sigma\big) \\
&\qquad \times \frac{1}{\sigma^{j-v+1}} \mathscr{C}\left(n - l - \sum_{i=1}^{v-1} s_i, 1; \sigma, (j-v)\sigma - h\right).
\end{aligned}
$$

Starting from (A.5) we can now introduce a recursive argument to solve the remaining nested sums over the indexes $s_1, \ldots, s_{v-1}$. In particular, consider the sum over the index s_{v-1},

that is,

$$\begin{aligned}
&\sum_{s_{v-1}=1}^{n-j+1-\sum_{i=1}^{v-2}(s_i-1)} \binom{n}{s_1,\ldots,s_{v-1},n-\sum_{i=1}^{v-1}s_i} \prod_{i=1}^{v-1}(1-\sigma)_{(s_i-1)\uparrow 1} \\
&\times \sum_{t=k}^{m}\binom{m}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-t}\binom{m-t}{h}\big(-(j-v)\sigma\big)_{h\uparrow 1} \\
&\times \sum_{l=j-v}^{n-\sum_{i=1}^{v-1}s_i-1}\binom{n-\sum_{i=1}^{v-1}s_i}{l}(l)_{(m-t-h)\uparrow 1}\mathscr{C}\big(l,j-v;\sigma,-(j-v)\sigma\big) \\
&\times \frac{1}{\sigma^{j-v+1}}\mathscr{C}\left(n-l-\sum_{i=1}^{v-1}s_i,1;\sigma,(j-v)\sigma-h\right)
\end{aligned} \tag{A.6}$$

which can be written as

$$\begin{aligned}
&\frac{1}{\sigma^{j-v+1}}\binom{n}{s_1,\ldots,s_{v-2},n-\sum_{i=1}^{2}s_i}\prod_{i=1}^{v-2}(1-\sigma)_{(s_i-1)\uparrow 1} \\
&\times \sum_{t=k}^{m}\binom{m}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-t}\binom{m-t}{h}\big(-(j-v)\sigma\big)_{h\uparrow 1} \\
&\times \sum_{l=j-v+1}^{n-\sum_{i=1}^{v-2}s_i-1}(l-1)_{(m-t-h)\uparrow 1}\mathscr{C}\big(l-1,j-v;\sigma,-(j-v)\sigma\big) \\
&\times \sum_{s_{v-1}=1}^{n-l-\sum_{i=1}^{v-2}s_i}\binom{n-\sum_{i=1}^{v-2}s_i}{s_{v-1}}\binom{n-\sum_{i=1}^{v-2}s_i-s_{v-1}}{l-1} \\
&\times (1-\sigma)_{(s_{v-1}-1)\uparrow 1}\mathscr{C}\left(n-l+1-\sum_{i=1}^{v-2}s_i-s_{v-1},1;\sigma,(j-v)\sigma-h\right)
\end{aligned}$$

(by (A.4) to expand $\mathscr{C}(n-l-\sum_{i=1}^{v-2}s_i-s_{v-1}+1,1;\sigma,(j-v)\sigma-h)$)

$$\begin{aligned}
&= \frac{1}{\sigma^{j-v+1}}\binom{n}{s_1,\ldots,s_{v-2},n-\sum_{i=1}^{2}s_i}\prod_{i=1}^{v-2}(1-\sigma)_{(s_i-1)\uparrow 1} \\
&\quad\times \sum_{t=k}^{m}\binom{m}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-t}\binom{m-t}{h}\big(-(j-v)\sigma\big)_{h\uparrow 1}
\end{aligned}$$

$$\times \sum_{l=j-v+1}^{n-\sum_{i=1}^{v-2} s_i - 1} (l-1)_{(m-t-h)\uparrow 1} \mathscr{C}\big(l-1, j-v; \sigma, -(j-v)\sigma\big)$$

$$\times \sum_{z=1}^{n-l-\sum_{i=1}^{v-2} s_i} \binom{n-\sum_{i=1}^{v-2} s_i}{l-1, z, n-l+1-z-\sum_{i=1}^{v-2} s_i} \mathscr{C}(z, 1; \sigma)$$

$$\times \sum_{s_{v-1}=1}^{n-l+1-z-\sum_{i=1}^{v-2} s_i} \binom{n-l+1-z-\sum_{i=1}^{v-2} s_i}{s_{v-1}}$$

$$\times (1-\sigma)_{(s_{v-1}-1)\uparrow 1} \big(-(j-v)\sigma + h\big)_{(n-l+1-z-\sum_{i=1}^{v-2} s_i - s_{v-1})\uparrow 1}$$

(by equation 2.56 in [2] to solve the sum over the index s_{v-1})

$$= \frac{1}{\sigma^{j-v+1}} \binom{n}{s_1, \dots, s_{v-2}, n-\sum_{i=1}^{2} s_i} \prod_{i=1}^{v-2} (1-\sigma)_{(s_i-1)\uparrow 1}$$

$$\times \sum_{t=k}^{m} \binom{m}{t} \mathscr{C}(t, k; \sigma) \sum_{h=0}^{m-t} \binom{m-t}{h} \big(-(j-v)\sigma\big)_{h\uparrow 1}$$

$$\times \sum_{l=j-v+1}^{n-\sum_{i=1}^{v-2} s_i - 1} (l-1)_{(m-t-h)\uparrow 1} \mathscr{C}\big(l-1, j-v; \sigma, -(j-v)\sigma\big)$$

$$\times \sum_{z=1}^{n-l-\sum_{i=1}^{v-2} s_i} \binom{n-\sum_{i=1}^{v-2} s_i}{l-1, z, n-l+1-z-\sum_{i=1}^{v-2} s_i} \mathscr{C}(z, 1; \sigma)$$

$$\times \frac{1}{\sigma} \mathscr{C}\left(n-l+1-z-\sum_{i=1}^{v-2} s_i, 1; \sigma, (j-v)\sigma - h\right)$$

(by (A.4) to solve the sum over the index z)

$$= \binom{2}{1} \binom{n}{s_1, \dots, s_{v-2}, n-\sum_{i=1}^{2} s_i} \prod_{i=1}^{v-2} (1-\sigma)_{(s_i-1)\uparrow 1}$$

$$\times \sum_{t=k}^{m} \binom{m}{t} \mathscr{C}(t, k; \sigma) \sum_{h=0}^{m-t} \binom{m-t}{h} \big(-(j-v)\sigma\big)_{h\uparrow 1}$$

$$\times \sum_{l=j-v}^{n-\sum_{i=1}^{v-2} s_i} \binom{n-\sum_{i=1}^{v-2} s_i}{l} (l)_{(m-t-h)\uparrow 1} \mathscr{C}\big(l, j-v; \sigma, -(j-v)\sigma\big)$$

$$\times \frac{1}{\sigma^{j-v+2}} \mathscr{C}\left(n-l-\sum_{i=1}^{v-2} s_i, 2; \sigma, (j-v)\sigma - h\right).$$

Note that the resulting expression has the same structure of the summand in (A.6). This fact suggests the possibility of repeating the above arguments to each of the remaining nested sums over the indexes $s_{v-2}, \ldots, s_1$, respectively. In particular, after a repeated application of these arguments we can write the rth factorial moment of $R_m^{(n,j)}$ as follows

$$\begin{aligned}
&\mathbb{E}\big[\big(R_m^{(n,j)}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{\sigma^j}{\mathscr{C}(n, j; \sigma)} \sum_{k=0}^{m} \frac{V_{n+m, j+k}}{V_{n,j}} \frac{1}{\sigma^k} r! \sum_{v=0}^{r} \binom{j-v}{r-v} (-1)^v \\
&\qquad \times \sum_{t=k}^{m} \binom{m}{t} \mathscr{C}(t, k; \sigma) \sum_{h=0}^{m-t} \binom{m-t}{h} \big(-(j-v)\sigma\big)_{h\uparrow 1} \\
&\qquad \times \frac{1}{\sigma^j} \sum_{l=j-v}^{n-v} \binom{n}{l} (l)_{(m-t-h)\uparrow 1} \mathscr{C}\big(l, j-v; \sigma, -(j-v)\sigma\big) \mathscr{C}\big(n-l, v; \sigma, (j-v)\sigma - h\big).
\end{aligned} \tag{A.7}$$

Finally, a direct application of (A.4) to expand $\mathscr{C}(n-l, v; \sigma, (j-v)\sigma - h)$ we can write (A.7) as

$$\begin{aligned}
&\mathbb{E}\big[\big(R_m^{(n,j)}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{\sigma^j}{\mathscr{C}(n, j; \sigma)} \sum_{k=0}^{m} \frac{V_{n+m, j+k}}{V_{n,j}} \frac{1}{\sigma^k} r! \sum_{v=0}^{r} \binom{j-v}{r-v} (-1)^v \frac{1}{\sigma^j} \sum_{s=v}^{n-(j-v)} \binom{n}{s} \mathscr{C}(s, v; \sigma) \\
&\qquad \times \sum_{t=k}^{m} \binom{m}{t} \mathscr{C}(t, k; \sigma) \sum_{h=0}^{m-t} \binom{m-t}{h} \big(-(j-v)\sigma\big)_{h\uparrow 1} \\
&\qquad \times \sum_{l=j-v}^{n-s} \binom{n-s}{s} (l)_{(m-t-h)\uparrow 1} \big(-(j-v)\sigma + h\big)_{(n-s-l)\uparrow 1} \mathscr{C}\big(l, j-v; \sigma, -(j-v)\sigma\big)
\end{aligned}$$

which leads to (3.5) by means of (A.4) and some standard algebra involving factorial numbers and noncentral generalized factorial coefficients. This completes the second part of the proof. □

Proof of Proposition 1. By combining the rth factorial moment of $R_m^{(n,j,\mathbf{n})}$ in Theorem 1 with $V_{n,j}$ displayed in (2.3) one has

$$\begin{aligned}
&\mathbb{E}\big[\big(R_m^{(n,j,\mathbf{n})}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{r!}{(\theta+n)_{m\uparrow 1}} \sum_{v=0}^{r} \binom{j-v}{r-v} (-1)^v \\
&\qquad \times \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \sum_{k=0}^{m} \left(\frac{\theta}{\sigma}+j\right)_{k\uparrow 1} \mathscr{C}\left(m,k;\sigma,-n+\sum_{i=1}^{v} n_{c_i}+(j-v)\sigma\right) \qquad \text{(A.8)} \\
&\quad = \frac{r!}{(\theta+n)_{m\uparrow 1}} \sum_{v=0}^{r} \binom{j-v}{r-v} (-1)^v \\
&\qquad \times \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \left(\theta+n-\sum_{i=1}^{v} n_{c_i}+\sigma v\right)_{m\uparrow 1},
\end{aligned}$$

where the last identity follows from equation 2.49 in [2]. Accordingly, (3.7) follows from (A.8) by setting $r=1$. Regarding (3.6), an inversion of the generating function for the rth factorial moment in (A.8) leads to

$$\begin{aligned}
&\mathbb{P}\big[R_m^{(n,j,\mathbf{n})}=x\big] \\
&\quad = \frac{1}{(\theta+n)_{m\uparrow 1}} \sum_{y\geq 0} \frac{1}{x!} \frac{\mathrm{d}^x}{\mathrm{d}t^x} (t-1)^{x+y}\Big|_{t=0} \qquad \text{(A.9)} \\
&\qquad \times \sum_{v=0}^{x+y} \binom{j-v}{x+y-v} (-1)^v \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \left(\theta+n-\sum_{i=1}^{v} n_{c_i}+\sigma v\right)_{m\uparrow 1},
\end{aligned}$$

where

$$\frac{\mathrm{d}^x}{\mathrm{d}t^x}(t-1)^{x+y}\Big|_{t=0} = (-1)^y (x+y)_{x\downarrow 1}.$$

The proof is then completed by means of standard algebra involving factorial numbers and binomial coefficients. Specifically, since $\binom{j-v}{x+y-v}=0$ for any $y>j-x$ then (A.9) can be written as

$$\begin{aligned}
&\mathbb{P}\big[R_m^{(n,j,\mathbf{n})}=x\big] \\
&\quad = \frac{1}{(\theta+n)_{m\uparrow 1}} \sum_{y=0}^{j} (-1)^{y-x} \binom{y}{y-x} \sum_{v=0}^{y} (-1)^v \binom{j-v}{y-v}
\end{aligned}$$

$$
\begin{aligned}
&\times \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \left(\theta+n-\sum_{i=1}^{v} n_{c_i}+\sigma v\right)_{m\uparrow 1} \\
&= \frac{1}{(\theta+n)_{m\uparrow 1}}(-1)^x \sum_{v=0}^{j}\sum_{y=0}^{j-v}(-1)^y \binom{j-v}{y}\binom{y+v}{x} \\
&\times \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \left(\theta+n-\sum_{i=1}^{v} n_{c_i}+\sigma v\right)_{m\uparrow 1} \\
&= \frac{1}{(\theta+n)_{m\uparrow 1}}(-1)^x \sum_{v=0}^{j}(-1)^{j-v} \binom{v}{x-j+v} \\
&\times \sum_{\{c_1,\ldots,c_v\}\in\mathcal{C}_{j,v}} \left(\theta+n-\sum_{i=1}^{v} n_{c_i}+\sigma v\right)_{m\uparrow 1}
\end{aligned}
$$

which leads to (3.6) by means of standard algebraic manipulations involving factorial numbers. □

Proof of Proposition 2. A combination of the rth factorial moment of $R_m^{(n,j)}$ in Theorem 1 with $V_{n,j}$ displayed in (2.3) leads to

$$
\begin{aligned}
&\mathbb{E}\left[\left(R_m^{(n,j)}\right)_{r\downarrow 1}\right] \\
&= \frac{r!}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \sum_{v=0}^{r}\binom{j-v}{r-v}(-1)^v \\
&\quad \times \sum_{s=v}^{n-(j-v)} \binom{n}{s}\mathscr{C}(s,v;\sigma)\mathscr{C}(n-s,j-v;\sigma) \\
&\quad \times \sum_{k=0}^{m}\left(\frac{\theta}{\sigma}+j\right)_{k\uparrow 1}\mathscr{C}\left(m,k;\sigma,-n+s+(j-v)\sigma\right) \\
&= \frac{r!}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \sum_{v=0}^{r}\binom{j-v}{r-v}(-1)^v \\
&\quad \times \sum_{s=v}^{n-(j-v)} \binom{n}{s}(\theta+n-s+v\sigma)_{m\uparrow 1}\mathscr{C}(s,v;\sigma)\mathscr{C}(n-s,j-v;\sigma),
\end{aligned}
\tag{A.10}
$$

where the last identity follows from equation 2.49 in [2]. Accordingly, (3.9) follows from (A.10) by setting $r=1$. Regarding (3.8), an inversion of the generating function for the rth factorial

moment in (A.10) leads to

$$\begin{aligned}
&\mathbb{P}\big[R_m^{(n,j)} = x\big] \\
&\quad = \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \sum_{y\geq 0} \frac{1}{x!} \frac{\mathrm{d}^x}{\mathrm{d}t^x} (t-1)^{x+y} \Big|_{t=0} \\
&\qquad \times \sum_{v=0}^{x+y} \binom{j-v}{x+y-v} (-1)^v \\
&\qquad \times \sum_{s=v}^{n-(j-v)} \binom{n}{s} (\theta+n-s+v\sigma)_{m\uparrow 1} \mathscr{C}(s,v;\sigma) \mathscr{C}(n-s,j-v;\sigma),
\end{aligned} \tag{A.11}$$

where

$$\frac{\mathrm{d}^x}{\mathrm{d}t^x} (t-1)^{x+y} \Big|_{t=0} = (-1)^y (x+y)_{x\downarrow 1}.$$

The proof is then completed by means of standard algebra involving factorial numbers and binomial coefficients. Specifically, since $\binom{j-v}{x+y-v} = 0$ for any $y > j-x$ then (A.11) can be written as

$$\begin{aligned}
&\mathbb{P}\big[R_m^{(n,j)} = x\big] \\
&\quad = \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \sum_{y=0}^{j} (-1)^{y-x} \binom{y}{y-x} \sum_{v=0}^{y} \binom{j-v}{y-v} (-1)^v \\
&\qquad \times \sum_{s=v}^{n-(j-v)} \binom{n}{s} (\theta+n-s+v\sigma)_{m\uparrow 1} \mathscr{C}(s,v;\sigma) \mathscr{C}(n-s,j-v;\sigma) \\
&\quad = \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} (-1)^x \sum_{v=0}^{j} \sum_{y=0}^{j-v} (-1)^y \binom{j-v}{y} \binom{y+v}{x} \\
&\qquad \times \sum_{s=v}^{n-(j-v)} \binom{n}{s} (\theta+n-s+v\sigma)_{m\uparrow 1} \mathscr{C}(s,v;\sigma) \mathscr{C}(n-s,j-v;\sigma) \\
&\quad = \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} (-1)^x \sum_{v=0}^{j} (-1)^{j-v} \binom{v}{x-j+v} \\
&\qquad \times \sum_{s=v}^{n-(j-v)} \binom{n}{s} (\theta+n-s+v\sigma)_{m\uparrow 1} \mathscr{C}(s,v;\sigma) \mathscr{C}(n-s,j-v;\sigma)
\end{aligned}$$

which leads to (3.8) by means of standard algebraic manipulations involving factorial numbers. □

A.2. Proofs of the results in Section 3.2

Proof of Theorem 2. With regard to the rth factorial moment of $R_{l,m}^{(n,j,\mathbf{n})}$, this is obtained by a direct application of Theorem 1 in [8]. This completes the first part of the proof. With regard the rth factorial moment of $R_{l,m}^{(n,j)}$, this is obtained by combining (3.12) with the distributions displayed in (2.1) and (2.2). Specifically, we can write the following expression

$$\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j)}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{\sigma^j}{\mathscr{C}(n,j;\sigma)} r! \binom{m}{l,\ldots,l,m-rl} \\
&\qquad \times \frac{1}{j!} \sum_{(n_1,\ldots,n_j)\in\mathcal{D}_{n,j}} \binom{n}{n_1,\ldots,n_j} \prod_{i=1}^{j} (1-\sigma)_{(n_i-1)\uparrow 1} \\
&\qquad \times \sum_{\{c_1,\ldots,c_r\}\in\mathcal{C}_{j,r}} \prod_{i=1}^{r} (n_{c_i}-\sigma)_{l\uparrow 1} \\
&\qquad \times \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{\mathscr{C}(m-rl,k;\sigma,-n+\sum_{i=1}^{r} n_{c_i}+(j-r)\sigma)}{\sigma^k}
\end{aligned} \tag{A.12}$$

and prove that it coincides with (3.13). As in Theorem 1 the main issue consists in solving the sums over the collection of indexes $n_1,\ldots,n_j$ and $c_1,\ldots,c_r$. First, by means of equation 2.61 in [2] and using the fact that $\mathcal{C}_{j,r}$ has cardinality $\binom{j}{r}$, from (A.12) one has

$$\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j)}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{\sigma^j}{\mathscr{C}(n,j;\sigma)} \binom{m}{l,\ldots,l,m-rl} \sum_{k=0}^{m} \frac{V_{n+m,j+k}}{V_{n,j}} \frac{1}{\sigma^k} \\
&\qquad \times \sum_{s_1=1}^{n-j+1} \sum_{s_2=1}^{n-j+1-(s_1-1)} \cdots \sum_{s_r=1}^{n-j+1-\sum_{i=1}^{r-1}(s_i-1)} \binom{n}{s_1,\ldots,s_r,n-\sum_{i=1}^{r} s_i} \\
&\qquad \times \prod_{i=1}^{r} (1-\sigma)_{(s_i-1)\uparrow 1} (s_i-\sigma)_{l\uparrow 1} \\
&\qquad \times \frac{1}{\sigma^{j-r}} \mathscr{C}\Big(m-rl,k;\sigma,-n+\sum_{i=1}^{r} s_i+(j-r)\sigma\Big) \mathscr{C}\Big(n-\sum_{i=1}^{r} s_i, j-r;\sigma\Big).
\end{aligned} \tag{A.13}$$

As in the proof of Theorem 1, in order to solve the nested sums over the indexes $s_1,\ldots,s_r$ in (A.3) we first deal with the sum over the index s_r. Recall that for any $x \geq 0$ and $0 \leq y \leq x$, for

any $a > 0$, $b > 0$, $c > 0$ and for any real number d one has the following identity

$$\binom{y+c}{y}\mathscr{C}(x, y+c; d, a+b) = \sum_{j=y}^{x-c}\binom{x}{j}\mathscr{C}(j, y; d, a)\mathscr{C}(x-j, c; d, b). \tag{A.14}$$

See Chapter 2 of [2] for details. Then, let us consider the sum over the index s_r in (A.13), that is,

$$\begin{aligned}
&\sum_{s_r=1}^{n-j+1-\sum_{r=1}^{r-1}(s_i-1)} \binom{n}{s_1, \ldots, s_r, n-\sum_{i=1}^{r} s_i} \prod_{i=1}^{r}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1} \\
&\quad \times \frac{1}{\sigma^{j-r}}\mathscr{C}\left(m-rl, k; \sigma, -n+\sum_{i=1}^{r} s_i + (j-r)\sigma\right)\mathscr{C}\left(n-\sum_{i=1}^{r} s_i, j-r; \sigma\right) \\
&= \binom{n}{s_1, \ldots, s_{r-1}, n-\sum_{i=1}^{r-1} s_i} \prod_{i=1}^{r-1}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1} \\
&\quad \times \sum_{s_r=1}^{n-j+1-\sum_{i=1}^{r-1}(s_i-1)} \binom{n-\sum_{i=1}^{r-1} s_i}{s_r}(1-\sigma)_{(s_r-1)\uparrow 1}(s_r-\sigma)_{l\uparrow 1} \\
&\quad \times \frac{1}{\sigma^{j-r}}\mathscr{C}\left(m-rl, k; \sigma, -n+\sum_{i=1}^{r-1} s_i + s_r + (j-r)\sigma\right)\mathscr{C}\left(n-\sum_{i=1}^{r-1} s_i + s_r, j-r; \sigma\right).
\end{aligned}$$

By a direct application of (A.14) to the coefficients $\mathscr{C}(m-rl, k; \sigma, -n+\sum_{i=1}^{r} s_i + (j-r)\sigma)$ and $\mathscr{C}(n-\sum_{i=1}^{r} s_i, j-r; \sigma)$ we can write the last expression in the following expanded form

$$\begin{aligned}
&\frac{1}{\sigma^{j-r}}\binom{n}{s_1, \ldots, s_{r-1}, n-\sum_{i=1}^{r-1} s_i} \prod_{i=1}^{r-1}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1} \\
&\quad \times \sum_{t=k}^{m-rl}\binom{m-rl}{t}\mathscr{C}(t, k; \sigma) \sum_{z=j-r}^{n-\sum_{i=1}^{r-1} s_i - 1}\binom{n-\sum_{i=1}^{r-1} s_i}{z}\mathscr{C}\big(z, j-r; \sigma, -(j-r)\sigma\big) \\
&\quad \times \sum_{s_r=1}^{n-z-\sum_{i=1}^{r-1} s_i}\binom{n-z-\sum_{i=1}^{r-1} s_i}{s_r}(1-\sigma)_{(s_r-1)\uparrow 1}(s_r-\sigma)_{l\uparrow 1} \\
&\quad \times \left(n-\sum_{i=1}^{r-1} s_i - s_r - (j-r)\sigma\right)_{(m-rl-t)\uparrow 1} \big(-(j-r)\sigma\big)_{(n-z-\sum_{i=1}^{r-1} s_i - s_r)\uparrow 1}
\end{aligned}$$

(by the Vandermonde's identity to expand $(n-\sum_{i=1}^{r-1}s_i-s_r-(j-r)\sigma)_{(m-rl-t)\uparrow 1}$)

$$
\begin{aligned}
&= \frac{1}{\sigma^{j-r}}\binom{n}{s_1,\ldots,s_{r-1},n-\sum_{i=1}^{r-1}s_i}\prod_{i=1}^{r-1}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1}\\
&\quad\times\sum_{t=k}^{m-rl}\binom{m-rl}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-rl-t}\binom{m-rl-t}{h}\big(-(j-r)\sigma\big)_{h\uparrow 1}\\
&\quad\times\sum_{z=j-r}^{n-\sum_{i=1}^{r-1}s_i-1}\binom{n-\sum_{i=1}^{r-1}s_i}{z}(z)_{(m-rl-t-h)\uparrow 1}\mathscr{C}\big(z,j-r;\sigma,-(j-r)\sigma\big)\\
&\quad\times(1-\sigma)_{(l-1)\uparrow 1}\sum_{s_r=1}^{n-z-\sum_{i=1}^{r-1}s_i}\binom{n-z-\sum_{i=1}^{r-1}s_i}{s_r}\\
&\quad\times(l-\sigma)_{s_r\uparrow 1}\big(-(j-r)\sigma+h\big)_{(n-z-\sum_{i=1}^{r-1}s_i-s_r)\uparrow 1}
\end{aligned}
$$

(by equation 2.60 in [2] to solve the sum over the index s_r)

$$
\begin{aligned}
&= \frac{1}{\sigma^{j-r}}\binom{n}{s_1,\ldots,s_{r-1},n-\sum_{i=1}^{r-1}s_i}\prod_{i=1}^{r-1}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1}\\
&\quad\times\sum_{t=k}^{m-rl}\binom{m-rl}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-rl-t}\binom{m-rl-t}{h}\big(-(j-r)\sigma\big)_{h\uparrow 1}\\
&\quad\times\sum_{z=j-r}^{n-\sum_{i=1}^{r-1}s_i-1}\binom{n-\sum_{i=1}^{r-1}s_i}{z}(z)_{(m-rl-t-h)\uparrow 1}\mathscr{C}\big(z,j-r;\sigma,-(j-r)\sigma\big)\\
&\quad\times(1-\sigma)_{(l-1)\uparrow 1}(-1)\mathscr{C}\left(n-z-\sum_{i=1}^{r-1}s_i;1;\sigma-l,(j-r)\sigma-h\right)
\end{aligned}
$$

providing the solution for the innermost nested sum over the index s_r. Therefore, according to the last identity, the rth factorial moment of $R_{l,m}^{(n,j)}$ in (A.13) has the following reduced expression

$$
\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j)}\big)_{r\downarrow 1}\big]\\
&\quad=\frac{\sigma^j}{\mathscr{C}(n,j;\sigma)}\binom{m}{l,\ldots,l,m-rl}\sum_{k=0}^{m}\frac{V_{n+m,j+k}}{V_{n,j}}\frac{1}{\sigma^k}
\end{aligned}
$$

$$
\begin{aligned}
&\times \sum_{s_1=1}^{n-j+1} \sum_{s_2=1}^{n-j+1-(s_1-1)} \cdots \sum_{s_{r-1}=1}^{n-j+1-\sum_{i=1}^{r-2}(s_i-1)} \binom{n}{s_1,\ldots,s_{r-1},n-\sum_{i=1}^{r-1}s_i} \\
&\times \prod_{i=1}^{r-1}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1} \\
&\times \sum_{t=k}^{m-rl}\binom{m-rl}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-rl-t}\binom{m-rl-t}{h}\big(-(j-r)\sigma\big)_{h\uparrow 1} \\
&\times \sum_{z=j-r}^{n-\sum_{i=1}^{r-1}s_i-1}\binom{n-\sum_{i=1}^{r-1}s_i}{z}(z)_{m-rl-t-h}\mathscr{C}\big(z,j-r;\sigma,-(j-r)\sigma\big) \\
&\times \frac{1}{\sigma^{j-r}}(1-\sigma)_{(l-1)\uparrow 1}(-1)\mathscr{C}\left(n-z-\sum_{i=1}^{r-1}s_i;1;\sigma-l,(j-r)\sigma-h\right).
\end{aligned}
\tag{A.15}
$$

Starting from (A.15) we can now repeatedly apply equation 2.60 in [2] to solve the remaining sums over the indexes $s_1,\ldots,s_{r-1}$, respectively, starting from the index s_{r-1} and proceeding backward to the index s_1. As an example, consider the sum over the index s_{r-1}, that is,

$$
\begin{aligned}
&\sum_{s_{r-1}=1}^{n-j+1-\sum_{i=1}^{r-2}(s_i-1)} \binom{n}{s_1,\ldots,s_{r-1},n-\sum_{i=1}^{r-1}s_i}\prod_{i=1}^{r-s}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1} \\
&\times \sum_{t=k}^{m-rl}\binom{m-rl}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-rl-t}\binom{m-rl-t}{h}\big(-(j-r)\sigma\big)_{h\uparrow 1} \\
&\times \sum_{z=j-r}^{n-\sum_{i=1}^{r-1}s_i-1}\binom{n-\sum_{i=1}^{r-1}s_i}{z}(z)_{(m-rl-t-h)\uparrow 1}\mathscr{C}\big(z,j-r;\sigma,-(j-r)\sigma\big) \\
&\times \frac{1}{\sigma^{j-r}}(1-\sigma)_{(l-1)\uparrow 1}(-1)\mathscr{C}\left(n-z-\sum_{i=1}^{r-1}s_i;1;\sigma-l,(j-r)\sigma-h\right)
\end{aligned}
\tag{A.16}
$$

which can be written as

$$
\begin{aligned}
&\frac{1}{\sigma^{j-r}}\binom{n}{s_1,\ldots,s_{r-2},n-\sum_{i=1}^{r-2}s_i}\prod_{i=1}^{r-2}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1} \\
&\quad\times \sum_{t=k}^{m-rl}\binom{m-rl}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-rl-t}\binom{m-rl-t}{h}\big(-(j-r)\sigma\big)_{h\uparrow 1}
\end{aligned}
$$

$$\times \sum_{z=j-r+1}^{n-\sum_{i=1}^{r-2}s_i-1} \binom{n-\sum_{i=1}^{r-2}s_i}{z-1}(z-1)_{(m-rl-t-h)\uparrow 1}\mathscr{C}\big(z-1, j-r; \sigma, -(j-r)\sigma\big)$$

$$\times \big((1-\sigma)_{(l-1)\uparrow 1}\big)^2(-1)\sum_{s_{r-1}=1}^{n-z-\sum_{i=1}^{r-2}s_i} \binom{n-z+1-\sum_{i=1}^{r-2}s_i}{s_{r-1}}$$

$$\times (l-\sigma)_{s_{r-1}\uparrow 1}\mathscr{C}\left(n-z+1-\sum_{i=1}^{r-2}s_i-s_{r-1}, 1; \sigma-l, (j-r)\sigma-h\right)$$

(by equation 2.60 in [2] to solve the sum over the index s_{r-1})

$$= \binom{n}{s_1, \ldots, s_{r-2}, n-\sum_{i=1}^{r-2}s_i}\prod_{i=1}^{r-2}(1-\sigma)_{(s_i-1)\uparrow 1}(s_i-\sigma)_{l\uparrow 1}$$

$$\times \sum_{t=k}^{m-rl}\binom{m-rl}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-rl-t}\binom{m-rl-t}{h}\big(-(j-r)\sigma\big)_{h\uparrow 1}$$

$$\times \sum_{z=j-r+1}^{n-\sum_{i=1}^{r-2}s_i-1}\binom{n-\sum_{i=1}^{r-2}s_i}{z-1}(z-1)_{(m-rl-t-h)\uparrow 1}\mathscr{C}\big(z-1, j-r; \sigma, -(j-r)\sigma\big)$$

$$\times \frac{1}{\sigma^{j-r}}\big((1-\sigma)_{(l-1)\uparrow 1}\big)^2 2!(-1)^2\mathscr{C}\left(n-z+1-\sum_{i=1}^{r-2}s_i, 2; \sigma-l, (j-r)\sigma-h\right).$$

The resulting expression has the same structure of the summand in (A.16). This fact suggests the possibility of repeating exactly the above arguments to each of the remaining nested sum over the indexes $s_{r-2}, \ldots, s_1$, respectively. In particular, after a repeated application of these arguments we can write the rth factorial moment of $R_{l,m}^{(n,j)}$ in (A.15) as

$$\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j)}\big)_{r\downarrow 1}\big] \\
&= \frac{\sigma^j}{\mathscr{C}(n,j;\sigma)}\binom{m}{l, \ldots, l, m-rl}r!\frac{(-(1-\sigma)_{l-1})^r}{\sigma^{j-r}}\sum_{k=0}^{m}\frac{V_{n+m,j+k}}{V_{n,j}}\frac{1}{\sigma^k} \\
&\quad \times \sum_{t=k}^{m-rl}\binom{m-rl}{t}\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-rl-t}\binom{m-rl-t}{h}\big(-(j-r)\sigma\big)_{h\uparrow 1} \\
&\quad \times \sum_{z=j-r}^{n-r}\binom{n}{z}(z)_{(m-rl-t-h)\uparrow 1}\mathscr{C}\big(z, j-r; \sigma, -(j-r)\sigma\big)\mathscr{C}\big(n-z, r; \sigma-l, (j-r)\sigma-h\big).
\end{aligned} \tag{A.17}$$

Finally, by applying (A.14) to expand $\mathscr{C}(n-z,r;\sigma-l,(j-r)\sigma-h)$ in (A.17), we can write (A.17) as

$$\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j)}\big)_{r\downarrow 1}\big] \\
&= \frac{\sigma^j}{\mathscr{C}(n,j;\sigma)}\binom{m}{l,\ldots,l,m-rl} r!\frac{(-(1-\sigma)_{l-1})^r}{\sigma^{j-r}}\sum_{k=0}^{m}\frac{V_{n+m,j+k}}{V_{n,j}}\frac{1}{\sigma^k} \\
&\quad\times\sum_{s=r}^{n-(j-r)}\binom{n}{s}\mathscr{C}(s,r;\sigma-l)\sum_{t=k}^{m-rl}\binom{m-rl}{t} \\
&\quad\times\mathscr{C}(t,k;\sigma)\sum_{h=0}^{m-rl-t}\binom{m-rl-t}{h}\big(-(j-r)\sigma\big)_{h\uparrow 1} \\
&\quad\times\sum_{z=j-r}^{n-s}\binom{n-s}{z}(z)_{(m-rl-t-h)\uparrow 1}\big(-(j-r)\sigma+h\big)_{(n-s-z)\uparrow 1} \\
&\quad\times\mathscr{C}\big(z,j-r;\sigma,-(j-r)\sigma\big)
\end{aligned}$$

which leads to (3.13) by means on (A.14) and some standard algebra involving factorial numbers and noncentral generalized factorial coefficients. This completes the second part of the proof. □

Proof of Proposition 3. By combining the rth factorial moment of $R_{l,m}^{(n,j,\mathbf{n})}$ in Theorem 2 with $V_{n,j}$ displayed in (2.3) one has

$$\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j,\mathbf{n})}\big)_{r\downarrow 1}\big] \\
&= \frac{r!}{(\theta+n)_{m\uparrow 1}}\binom{m}{l,\ldots,l,m-rl} \\
&\quad\times\sum_{\{c_1,\ldots,c_r\}\in\mathcal{C}_{j,r}}\prod_{i=1}^{r}(n_{c_i}-\sigma)_{l\uparrow 1} \\
&\quad\times\sum_{k=0}^{m}\left(\frac{\theta}{\sigma}+j\right)_{k\uparrow 1}\mathscr{C}\left(m-rl,k;\sigma,-n+\sum_{i=1}^{r}n_{c_i}+(j-r)\sigma\right) \\
&= \frac{r!}{(\theta+n)_{m\uparrow 1}}\binom{m}{l,\ldots,l,m-rl} \\
&\quad\times\sum_{\{c_1,\ldots,c_r\}\in\mathcal{C}_{j,r}}\prod_{i=1}^{r}(n_{c_i}-\sigma)_{l\uparrow 1}\left(\theta+n-\sum_{i=1}^{r}n_{c_i}+\sigma r\right)_{(m-rl)\uparrow 1},
\end{aligned} \tag{A.18}$$

where the last identity follows equation 2.49 in [2]. Accordingly, (3.15) follows from (A.18) by setting $r = 1$. With regard to (3.15), an inversion of the generating function for the rth factorial moment in (A.18) leads to

$$\begin{aligned}
&\mathbb{P}\big[R_{l,m}^{(n,j,\mathbf{n})} = x\big] \\
&\quad = \frac{1}{(\theta+n)_{m\uparrow 1}} \sum_{y\geq 0} \frac{1}{x!} \frac{\mathrm{d}^x}{\mathrm{d}t^x}(t-1)^{x+y}\Big|_{t=0} \binom{m}{l,\ldots,l,m-(x+y)l} \\
&\qquad \times \sum_{\{c_1,\ldots,c_{x+y}\}\in\mathcal{C}_{j,x+y}} \prod_{i=1}^{x+y} (n_{c_i}-\sigma)_{l\uparrow 1} \left(\theta+n-\sum_{i=1}^{x+y} n_{c_i}+\sigma(x+y)\right)_{(m-(x+y)l)\uparrow 1},
\end{aligned} \tag{A.19}$$

where

$$\frac{\mathrm{d}^x}{\mathrm{d}t^x}(t-1)^{x+y}\Big|_{t=0} = (-1)^y (x+y)_{x\downarrow 1}.$$

Then (3.14) follows from (A.19) by means of standard algebra involving factorial numbers and binomial coefficients. □

Proof of Proposition 4. A combination of the rth factorial moment of $R_{l,m}^{(n,j)}$ in Theorem 2 with $V_{n,j}$ displayed in (2.3) leads to

$$\begin{aligned}
&\mathbb{E}\big[\big(R_{l,m}^{(n,j)}\big)_{r\downarrow 1}\big] \\
&\quad = \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \binom{m}{l,\ldots,l,m-rl} r!\big(-\sigma(1-\sigma)_{(l-1)\uparrow 1}\big)^r \\
&\qquad \times \sum_{s=r}^{n-(j-r)} \binom{n}{s} \mathscr{C}(s,r;\sigma-l)\mathscr{C}(n-s,j-r;\sigma) \\
&\qquad \times \sum_{k=0}^{m} \left(\frac{\theta}{\sigma}+j\right)_{k\uparrow 1} \mathscr{C}\big(m-rl,k;\sigma,-n+s+(j-r)\sigma\big) \\
&\quad = \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \binom{m}{l,\ldots,l,m-rl} r!\big(-\sigma(1-\sigma)_{(l-1)\uparrow 1}\big)^r \\
&\qquad \times \sum_{s=r}^{n-(j-r)} \binom{n}{s} (\theta+n-s+\sigma r)_{(m-rl)\uparrow 1} \\
&\qquad \mathscr{C}(s,r;\sigma-l)\mathscr{C}(n-s,j-r;\sigma),
\end{aligned} \tag{A.20}$$

where the last identity follows from equation 2.49 in [2]. Accordingly, (3.17) follows from (A.20) by setting $r = 1$. With regard to (3.16), an inversion of the generating function for the rth factorial

moment in (A.20) leads to

$$\begin{aligned}
&\mathbb{P}\big[R_{l,m}^{(n,j)} = x\big] \\
&= \frac{1}{\mathscr{C}(n,j;\sigma)(\theta+n)_{m\uparrow 1}} \sum_{y\geq 0} \frac{1}{x!} \frac{\mathrm{d}^r}{\mathrm{d}t^x}(t-1)^{x+y}\Big|_{t=0} \\
&\quad \times \binom{m}{l,\ldots,l,m-(x+y)l} \big(-\sigma(1-\sigma)_{(l-1)\uparrow 1}\big)^{(x+y)} \\
&\quad \times \sum_{s=x+y}^{n-(j-x-y)} \binom{n}{s} \big(\theta+n-s+\sigma(x+y)\big)_{(m-(x+y)l)\uparrow 1} \\
&\quad \times \mathscr{C}(s,x+y;\sigma-l)\mathscr{C}\big(n-s,j-(x+y);\sigma\big),
\end{aligned} \tag{A.21}$$

where

$$\frac{\mathrm{d}^x}{\mathrm{d}t^x}(t-1)^{x+y}\Big|_{t=0} = (-1)^y (x+y)_{x\downarrow 1}.$$

Then (3.16) follows from (A.21) by means of standard algebra involving factorial numbers and binomial coefficients. □

A.3. Proofs of the results in Section 3.3

Proof of Lemma 3.1. By suitably marginalizing the EPPF in (2.1) one obtains the distribution of $(K_n, \mathbf{N}_\tau)$, that is the main ingredient for determining (3.20). Specifically, one has

$$\begin{aligned}
&\mathbb{P}[K_n = j, \mathbf{N}_{\tau,n} = \mathbf{n}_\tau] \\
&= V_{n,j} \frac{(j-p)!}{j!} \binom{n}{n_{\tau_1},\ldots,n_{\tau_p}, n-\sum_{i=1}^p n_{\tau_i}} \prod_{i=1}^p (1-\sigma)_{(n_{\tau_i}-1)\uparrow 1} \\
&\quad \times \frac{1}{(j-p)!} \sum_{(n_{\nu_1},\ldots,n_{\nu_{j-p}})\in\mathcal{D}_{n-\sum_{i=1}^p n_{\tau_i},j-p}} \binom{n-\sum_{i=1}^p n_{\tau_i}}{n_{\nu_1},\ldots,n_{\nu_{j-p}}} \\
&\quad \times \prod_{i=1}^{(j-p)} (1-\sigma)_{(n_{\nu_i}-1)\uparrow 1} \\
&= V_{n,j} \frac{(j-p)!}{j!} \binom{n}{n_{\tau_1},\ldots,n_{\tau_p}, n-\sum_{i=1}^p n_{\tau_i}} \prod_{i=1}^p (1-\sigma)_{(n_{\tau_i}-1)\uparrow 1} \\
&\quad \times \frac{\mathscr{C}(n-\sum_{i=1}^p n_{\tau_i}, j-p;\sigma)}{\sigma^{j-p}}
\end{aligned} \tag{A.22}$$

where the last identity is obtained by a direct application of equation 2.61 in [2]. The proof is completed by taking the ratio between the distributions displayed in (2.1) and (A.22). □

Acknowledgements

The authors are grateful to an Associate Editor and a Referee for valuable remarks and suggestions that have lead to a substantial improvement in the presentation. Stefano Favaro is supported by the European Research Council (ERC) through StG "N-BNP" 306406.

References

[1] Arratia, R., Barbour, A.D. and Tavaré, S. (2003). *Logarithmic Combinatorial Structures*: *A Probabilistic Approach*. *EMS Monographs in Mathematics*. Zürich: European Mathematical Society (EMS). MR2032426

[2] Charalambides, C.A. (2005). *Combinatorial Methods in Discrete Distributions*. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR2131068

[3] De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I. and Ruggiero, M. (2014). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. To appear.

[4] De Blasi, P., Lijoi, A. and Prünster, I. (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statist*. *Sinica* **23** 1299–1321. MR3114715

[5] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoret*. *Population Biology* **3** 87–112; erratum, ibid. **3** (1972), 240, 376. MR0325177

[6] Ewens, W.J. and Tavaré, S. (1998). The Ewens sampling formula. In *Encyclopedia of Statistical Sciences*. *A Wiley-Interscience Publication* **2** update (S. Kotz, C.B. Read and L.D. Banks, eds.) 230–234. New York: Wiley. MR1605063

[7] Favaro, S., Lijoi, A. and Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics* **68** 1188–1196. MR3040025

[8] Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *Ann*. *Appl*. *Probab*. **23** 1721–2160.

[9] Gnedin, A. (2010). A species sampling model with finitely many types. *Electron*. *Commun*. *Probab*. **15** 79–88. MR2606505

[10] Gnedin, A. and Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap*. *Nauchn*. *Sem*. *S*.*-Peterburg*. *Otdel*. *Mat*. *Inst*. *Steklov*. (*POMI*) **325** 83–102, 244–245. MR2160320

[11] Griffiths, R.C. and Spanò, D. (2007). Record indices and age-ordered frequencies in exchangeable Gibbs partitions. *Electron*. *J*. *Probab*. **12** 1101–1130. MR2336601

[12] Kingman, J.F.C. (1978). The representation of partition structures. *J*. *London Math*. *Soc*. (2) **18** 374–380. MR0509954

[13] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. MR2416792

[14] Lijoi, A., Mena, R.H. and Prünster, I. (2007). A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinformatics* **8** 339.

[15] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N.L. Hjort, C.C. Holmes, P. Müller and S.G. Walker, eds.). *Camb*. *Ser*. *Stat*. *Probab*. *Math*. 80–136. Cambridge: Cambridge Univ. Press. MR2730661

[16] Lijoi, A., Prünster, I. and Walker, S.G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18** 1519–1547. MR2434179

[17] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158. MR1337249

[18] Pitman, J. (2003). Poisson–Kingman partitions. In *Statistics and Science: A Festschrift for Terry Speed* (D.R. Goldstein, ed.). *Institute of Mathematical Statistics Lecture Notes – Monograph Series* **40** 1–34. Beachwood, OH: IMS. MR2004330

[19] Pitman, J. (2006). *Combinatorial Stochastic Processes. Lectures from the* 32*nd Summer School on Probability Theory held in Saint-Flour, July* 7–24, 2002. *Lecture Notes in Math.* **1875**. Berlin: Springer. MR2245368

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

[21] Teh, Y.W. and Jordan, M.I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics* (N.L. Hjort, C.C. Holmes, P. Müller and S.G. Walker, eds.). *Camb. Ser. Stat. Probab. Math.* 158–207. Cambridge: Cambridge Univ. Press. MR2730663

[22] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480

Received October 2012 and revised June 2013