A new estimator of the discovery probability

Stefano Favaro[1], Antonio Lijoi[2] and Igor Prünster[3]

[1] Università degli Studi di Torino and Collegio Carlo Alberto.
E-mail: stefano.favaro@unito.it

[2] Università degli Studi di Pavia and Collegio Carlo Alberto.
E-mail: lijoi@unipv.it

[3] Università degli Studi di Torino and Collegio Carlo Alberto.
E-mail: igor@econ.unito.it

New version[*]

Abstract

Species sampling problems have a long history in ecological and biological studies and a number of issues, including the evaluation of species richness, the design of sampling experiments, the estimation of rare species variety, are to be addressed. Such inferential problems have recently emerged also in genomic applications, however exhibiting some peculiar features that make them more challenging: specifically, one has to deal with very large populations (genomic libraries) containing a huge number of distinct species (genes) and only a small portion of the library has been sampled (sequenced). These aspects motivate the Bayesian nonparametric approach we undertake, since it allows to achieve the degree of flexibility typically needed in this framework. Basing on an observed sample of size n, focus will be on prediction of a key aspect of the outcome from an additional sample of size m, namely the so–called *discovery probability*. In particular, conditionally on an observed basic sample of size n, we derive a novel estimator of the probability of detecting, at the $(n+m+1)$–th observation, species that have been observed with any given frequency in the enlarged sample of size $n+m$. Such an estimator admits a closed form expression that can be exactly evaluated. The result we obtain allows us to quantify both the rate at which rare species are detected and the achieved sample coverage of abundant species, as m increases. Natural applications are represented by the estimation of the probability of discovering rare genes within genomic libraries and the results are illustrated by means of two Expressed Sequence Tags datasets.

Key words and phrases: Bayesian nonparametrics; Gibbs–type priors; Rare species discovery; Species sampling models; Two–parameter Poisson–Dirichlet process.

[*]A minor inaccuracy appearing in the previous version has been corrected.

1 Introduction

Species sampling problems have a long history in ecological and biological studies. Suppose data are recorded from a population whose statistical units belong to different species. Therefore, data will consist of species labels and their corresponding frequencies. Based on a sample, various interesting predictive issues concerning the composition of the population arise. Such problems have regained popularity in recent years due to their frequent appearance in genomic applications, which are characterized by very large populations (genomic libraries) containing a huge number of distinct species (genes) and only a small portion of the library has been sampled (sequenced). Recently a Bayesian nonparametric method, particularly suited to such genomic contexts, has been proposed and implemented in Lijoi et al. (2007a,b). This approach is based on the randomization of the unknown proportions p_i of the individuals belonging to the species i in the whole population, for $i \geq 1$. It is further assumed that the recorded data $X_1, \ldots, X_n$ are part of an exchangeable sequence $(X_n)_{n\geq 1}$. In this case, it is well–known that by de Finetti's representation theorem $(X_n)_{n\geq 1}$ can be characterized by a hierarchical model, namely the X_n's as a random sample from some distribution $\tilde{P}$ and a prior Π on $\tilde{P}$, that is

$$\begin{aligned} X_i|\tilde{P} &\overset{\text{iid}}{\sim} \tilde{P} \\ \tilde{P} &\sim \Pi. \end{aligned} \tag{1}$$

Moreover, $\tilde{P}$ is a discrete random probability measure $\tilde{P} = \sum_{i\geq 1} w_i\, \delta_{Y_i}$, with $\sum_{i\geq 1} w_i = 1$ almost surely, belonging to the class of *Gibbs–type priors* (Gnedin and Pitman, 2006). Note that δ_c is used to denote the unit mass at point c, and the Y_i's are independent and identically distributed from some non–atomic distribution P_0: hence any two Y_i and Y_j, with $i \neq j$, are different with probability 1 and can be used to represent labels identifying different species in the population whose random proportions are p_i and p_j. The n observations $X_1, \ldots, X_n$ forming a *basic sample* are, then, observed species labels and, due to the discrete nature of $\tilde{P}$, may feature ties. If $K_n = j$ is the number of different species (labels) detected in the basic sample with respective frequencies $n_1, \ldots, n_j$, an additional potential and unobserved sample of size m, $X_{n+1}, \ldots, X_{n+m}$, is considered. Focus will be on prediction of a key aspect of the outcome of the additional sample, namely the so–called *discovery probability*, which coincides with the probability of sampling certain types of genes in further sequencing. The discovery probability represents a natural tool for a quantitative assessment of qualitative concepts such as species richness, rare species variety, sample coverage and optimal design of sampling experiments. A first partial result in this direction is obtained in Lijoi et al. (2007a), where an estimator for the probability that the $(n+m+1)$–th observation leads to discover a new species without actually observing the additional m–size sample is provided.

More specifically, an important inferential goal in ecological and biological studies is the evaluation of the probability that further sampling reveals: (i) new species; (ii) species that appear only once in the observed sample, the so–called unique species; (iii) rare species, where by rare species one refers to species whose frequency is below a certain threshold τ. In ecology, for instance, conservation of biodiversity is a fundamental theme and it can be formalized in terms of the proportion of species whose frequency is greater than a specified threshold of abundance. Indeed, any form of management on a sustained basis requires a certain proportion of sufficiently abundant species, the so–called breeding stock. See Magurran (2003), and references therein, for a comprehensive and stimulating survey on measurement of biodiversity, conservation of populations, commonness and rarity of species. On the other hand, in genomics rare genes are a fundamental issue in several problems, the most evident being that they are often associated with deleterious diseases. See, e.g., Laird and Lange (2010).

In this paper we provide an estimator for detecting a species of any given frequency at the $(n+m+1)$–th observation, therefore addressing the issues (ii) and (iii) above and improving remarkably, both in terms of theory and applications potential, on the results in Lijoi et al. (2007a) concerning problem (i) only. Specifically, based on a sample of size n, we will provide, for any frequency $k = 0, \ldots, n+m$ and additional unobserved sample size $m \geq 0$, an explicit estimator for the probability that the $(n+m+1)$–th observation coincides with a species whose frequency, within the sample of size $n+m$, is exactly k. In the sequel, we refer to such a probability as *m–step k–discovery probability* or, in short, *$[m:k]$–discovery*. Expressed in terms of the unknown species proportions in the whole population, p_i for $i \geq 1$, the determination of the $[m:k]$–discovery corresponds to estimating

$$U_{n+m}(k) := \sum_{i \geq 1} p_i \mathbb{1}_{\{k\}}(N_{i,n+m}), \tag{2}$$

where $N_{i,n+m}$ is the frequency with which the i–th species is recorded in the enlarged sample of size $n+m$ and clearly $U_{n+m}(0)$ represents the proportion of yet unobserved species or, equivalently, the probability of observing a new species.

For $m = 0$ such a problem has already been tackled in the literature. In particular, Good (1953) proposed a popular estimator for (2) of the form

$$\check{U}_{n+0}(k) = (k+1)\frac{\ell_{k+1}}{n} \tag{3}$$

where ℓ_r is the number of species in the sample with frequency r. Such an estimator is sometimes referred to as *Turing estimator*. In Mao and Lindsay (2002) an interesting new moment based derivation of (3) is provided. Note that $1 - \check{U}_{n+0}(0)$ estimates the *sample coverage* that is the proportion of distinct species present in the observed sample. On the other hand, Good and

Toulmin (1956) and Mao (2004) faced the general case of $m \geq 0$, however with the restriction $k = 0$, which corresponds only to the discovery of a new species. The resulting estimator is known as *Good–Toulmin estimator*. Further references addressing related issues are Chao (1981); Starr (1979); Clayton and Frees (1987); Guindani and Müller (2010); Sepulveda et al. (2010); Barger and Bunge (2010).

Our contribution therefore fills in an important gap since it allows to deal with the cases of both the additional sample size m and the frequency k to be any desired integer. As a by–product also the corresponding coverage estimators are obtained. Moreover, the coherent probabilistic structure of Bayesian nonparametric modeling allows to avoid some possible drawbacks related to the estimator (3), which will be detailed in the sequel by comparing the two estimators.

The structure of the paper is as follows. In Section 2 we describe the Bayesian nonparametric model and provide the main result. A completely explicit expression of the new estimator is then provided in the two–parameter Poisson–Dirichlet process case. Section 3 contains an illustration dealing with genomic EST datasets. The proofs are deferred to the Appendix.

2 Estimation of the m–step k–discovery

As outlined in the Introduction, the main aim of the present contribution is the estimation of the $[m : k]$–discovery. The deduced Bayesian estimator is based on a nonparametric model for the unknown distribution of species labels in (1) and, more importantly, can be exactly evaluated. The latter seems to be an attractive feature since it makes our approach readily implementable by practitioners. Let us first provide a quick introduction of the nonparametric process prior we resort to, along with some known results. We will, then, state the main result.

2.1 Preliminaries

It is apparent, from (1), that the prior Π can be defined as a distribution on the space of all probability measures for the observables $(X_n)_{n\geq 1}$. We now consider the case where the support of Π does not degenerate on a finite–dimensional space, so that the model is nonparametric, and Π is concentrated on the set of all discrete probability distributions for the X_i's. Such a specification is consistent with the nature of our data, since the process of observation leads to detecting multiple records for the same species. In other terms, if we let Π select discrete probability measures with probability one, then $\mathbb{P}[X_i = X_j] > 0$ for $i \neq j$ which is what we aim at. In this setup, a sample $X_1, \ldots, X_n$ is partitioned into K_n distinct values, i.e. species, $X_1^*, \ldots, X_{K_n}^*$ with vector of respective frequencies $\boldsymbol{N}_n = (N_{1,n}, \ldots, N_{K_n,n})$ being such that $\sum_{i=1}^{K_n} N_{i,n} = n$. In the sequel we consider a broad class of discrete random probability measures

defined as $\tilde{P} = \sum_{i \geq 1} w_i \, \delta_{Y_i}$, where the non–negative weights w_i's sum up to 1, i.e. $\sum_i w_i = 1$ (almost surely), and are independent from the locations Y_i's. The Y_i's, in turn, are i.i.d. from some non–atomic distribution P_0. It is further supposed that $\tilde{P}$ is such that the probability distribution of the partition of a sample into K_n distinct values with respective frequencies $\boldsymbol{N}_n$ is of the form

$$\mathbb{P}[K_n = j, \boldsymbol{N}_n = (n_1, \ldots, n_j)] = V_{n,j} \prod_{i=1}^{j} (1-\sigma)_{n_i - 1} \tag{4}$$

where $(a)_n = a(a+1) \cdots (a+n-1)$ stands for the n–th ascending factorial of a, with $(a)_0 \equiv 1$, σ is some parameter in $(0, 1)$ and the nonnegative weights $\{V_{n,j} : n \geq 1, 1 \leq j \leq n\}$ satisfy the following forward recursion $V_{n,j} = V_{n+1,j+1} + (n - j\sigma) V_{n+1,j}$. Such random probability measures and their corresponding distribution Π, which represents our prior distribution, are said of *Gibbs–type* and were introduced by Gnedin and Pitman (2006). The parameter σ that characterizes Π influences the partition structure of the data: indeed, as noted in Lijoi et al. (2007c), it regulates a reinforcement mechanism that determines the concentration of the observations in the different groups forming the partition. In summary, large values of σ tend to favour partitions with a larger number of clusters, i.e. of different observed species: most of the clusters (species) display very low frequencies whereas just a few clusters have very large abundances. Hence, a large value of σ reflects a prior opinion according to which a large number of species will be detected though only a few of them with high frequencies, which is a common situation in genomic datasets like the ones we are going to consider in Section 3.

In various applications, it is useful to describe the probability distribution of the random partition induced by Π in terms of the vector $\boldsymbol{\mathcal{L}}^{(n)} := (\mathcal{L}_{1,n}, \ldots, \mathcal{L}_{n,n})$ where $\mathcal{L}_{i,n}$ stands for the number of distinct species in the basic sample, of size n, with frequency i. Hence, (4) can be rewritten as

$$\mathbb{P}[\boldsymbol{\mathcal{L}}^{(n)} = (\ell_1, \ldots, \ell_n)] = V_{n,j} \, n! \prod_{i=1}^{n} \left\{ \frac{(1-\sigma)_{i-1}}{i!} \right\}^{\ell_i} \frac{1}{\ell_i!}, \tag{5}$$

for any vector of nonnegative integers $(\ell_1, \ldots, \ell_n)$ such that $\ell_i \geq 0$ for any i, $\sum_{i=1}^{n} i\ell_i = n$ and $\sum_{i=1}^{n} \ell_i = j$. The probability distribution in (5) is termed Gibbs–type sampling formula and it includes as special case the well–known Ewens sampling formula (Ewens, 1972) and Ewens–Pitman sampling formula (Pitman, 1995).

Gibbs–type priors lead to predictive distributions, given a sample of size n, $(X_1, \ldots, X_n)$, featuring $K_n = j$ distinct species $X_1^*, \ldots, X_j^*$ with frequencies $n_1, \ldots, n_j$, having the following simple structural form, which can be deduced from (4),

$$\mathbb{P}[X_{n+1} \in A \,|\, X_1, \ldots, X_n] = \frac{V_{n+1,j+1}}{V_{n,j}} P_0(A) + \frac{V_{n+1,j}}{V_{n,j}} \sum_{i=1}^{j} (n_i - \sigma) \delta_{X_i^*}(A). \tag{6}$$

Various noteworthy priors fall within the class of Gibbs–type priors, according as to the specific form of the weights $V_{n,j}$. For example, the Dirichlet process (Ferguson, 1973) with parameter measure θP_0 is recovered when $V_{n,j} = \theta^j/(\theta)_n$, whereas the two–parameter Poisson–Dirichlet process (Pitman, 1995) is obtained with $V_{n,j} = \prod_{i=1}^{j-1}(\theta + i\sigma)/(\theta+1)_{n-1}$. Also the normalized inverse Gaussian process (Lijoi et al., 2005) belongs to this class, with the $V_{n,j}$'s expressed in terms of a linear combination of incomplete gamma functions. Another instance of Gibbs–type prior is the normalized generalized gamma process that has been used in Lijoi et al. (2007c) for hierarchical mixture modelling and, more recently, in Kolossiatis et al. (2011) for modelling overdispersion in count data.

In Lijoi et al. (2007a) one can find a few results concerning Gibbs–type priors in view of their application to species sampling problems. For example, an estimator of the sample coverage can be easily deduced from the predictive distributions (6) so that

$$1 - \widehat{U}_n(0) := \widehat{C}_n = \frac{V_{n+1,j}}{V_{n,j}}(n - j\sigma). \tag{7}$$

Moreover, they determine an estimator for the $[m : 0]$–discovery given by

$$\widehat{U}_{n+m}(0) = \sum_{k=0}^{m} \frac{V_{n+m+1,j+k+1}}{V_{n,j}} \frac{\mathscr{C}(m,k;\sigma,-n+j\sigma)}{\sigma^k} \tag{8}$$

with $\mathscr{C}(m,k;\sigma,-n+j\sigma) = (k!)^{-1}\sum_{r=0}^{k}(-1)^r\binom{k}{r}(n-\sigma(r+j))_m$ being the non–central generalized factorial coefficient; see Charalambides (2005) for details and properties. The corresponding estimator of the sample coverage, given an observed sample of size n and an additional unobserved sample of size m, $\widehat{C}_{n+m}$ is then obtained as $1-\widehat{U}_{n+m}(0)$. The estimators (8) and, consequently, $\widehat{C}_{n+m}$ can then be specialized to various particular cases by plugging in the corresponding $V_{n,j}$'s.

2.2 Main results

As anticipated, our main goal is the estimation of $U_{n+m}(k)$, for any value of m and k, conditional on an observed sample of size n featuring j distinct species. We first state a simple result yielding a nonparametric Bayes estimator of $U_{n+0}(k)$ for any integer k. Such an estimation involves the one–step predictive distribution since it amounts to evaluating

$$\mathbb{P}[X_{n+1} \in \mathcal{G}_{k,n} \,|\, X_1, \ldots, X_n] \tag{9}$$

where $\mathcal{G}_{k,n} = \{X_i^* : N_{i,n} = k\}$ is the set of all distinct species that have appeared k times in the sample $X_1, \ldots, X_n$. For $k = 0$ it is apparent that (9) coincides with the probability that the $(n+1)$–th observation reveals a new species and is readily available from the predictive distribution (6), i.e. $V_{n+1,j+1}/V_{n,j}$. See also (7). Hence, we can focus on $k \geq 1$ and provide an explicit expression of the predictive probability (9).

THEOREM 1. *If* $(X_n)_{n\geq 1}$ *is an exchangeable sequence with Gibbs–type prior* Π *in* (1)

$$\widehat{U}_{n+0}(k) = \frac{V_{n+1,j}(k-\sigma)}{V_{n,j}}\,\ell_k, \tag{10}$$

where $\ell_k \geq 0$ *is the number of distinct species with frequency* k *observed within the sample* $(X_1,\ldots,X_n)$.

It is worth noting that (10) implies that $(K_n, \mathcal{L}_{k,n})$ is sufficient for estimating the $[0:k]$–discovery $U_{n+0}(k)$.

REMARK 1. The estimator $\widehat{U}_{n+0}(k)$ represents a Bayesian nonparametric counterpart to the popular Turing estimator $\check{U}_{n+0}(k)$ recalled in (3) and used in a frequentist approach. The most notable difference between the two estimators can be traced back to the different frequency count used. Indeed $\check{U}_{n+0}(k)$ in (3) depends only on the frequency ℓ_{k+1} of species that appeared $k+1$ times in the basic sample. This seems, in our opinion, counterintuitive since ℓ_k should be the main ingredient for the estimation of the discovery probability of species appearing k times. And the Bayesian estimator $\widehat{U}_{n+0}(k)$ in (10) perfectly adheres to such an intuition. Note that if there are no species appearing $k+1$ times (i.e. $\ell_{k+1}=0$) in the original sample of size n, then one would estimate as zero the probability of detecting a species with frequency k at the $(n+1)$–th observation; and this regardless of how many species with frequency k one has observed in the basic sample. This is not the case for the Bayesian nonparametric estimator (10). □

In many applications one is interested, rather than in identifying the $[0:k]$–discovery, in evaluating the chances of observing species that can be considered as *rare*, namely the sum of those not yet observed and of those having frequency below a certain threshold τ. In this case, an estimator of $\sum_{k=0}^{\tau} U_{n+0}(k)$ is

$$\widehat{U}_{n+0,\tau} = \sum_{k=0}^{\tau} \widehat{U}_{n+0}(k)$$

and it depends on the vector of summary statistics $(K_n, \mathcal{L}_{1,n},\ldots,\mathcal{L}_{\tau,n})$.

REMARK 2. It is worth pointing out that the Bayesian nonparametric estimator of the sample coverage (7) can also be recovered from (10). Indeed, by definition one has $\widehat{C}_n = 1-\widehat{U}_{n+0}(0) = \sum_{k=1}^{n}\widehat{U}_{n+0}(k)$ and the expression in (7) can be determined by resorting to Theorem 1 and by recalling that $\sum_{k=1}^{n} k\ell_k = n$ and $\sum_{k=1}^{n}\ell_k = j$. Another related quantity of interest can be determined from Theorem 1, namely an estimated *abundant species coverage*, where by *abundant*

we mean species which are not rare. This is given by the proportion of species with frequency larger than τ represented in the sample, i.e. $\widehat{C}_n^{abund} = 1 - \sum_{k=0}^{\tau} \widehat{U}_{n+0}(k)$. □

Let us now consider the case where $m \geq 1$: we are, then, going to evaluate the conditional probability, given the sample $X_1, \ldots, X_n$, that the $(n+m+1)$–th observation displays a species observed k times within $X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}$. This implies that we wish to estimate

$$\mathbb{P}[X_{n+m+1} \in \mathcal{G}_{k,n+m} | X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}]. \tag{11}$$

Assuming a squared loss function, we estimate (11) by evaluating the expectation with respect to the conditional distribution of the unobserved portion of the sample, $X_{n+1}, \ldots, X_{n+m}$, given the sample that has been actually observed, $X_1, \ldots, X_n$, i.e. the m–step ahead predictive distribution. In other terms the desired estimator $\widehat{U}_{n+m}(k)$ is obtained by minimizing

$$\mathbb{E}\left\{ \left(\mathbb{P}[X_{n+m+1} \in \mathcal{G}_{k,n+m} \,|\, X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}] - U\right)^2 \,|\, X_1, \ldots, X_n \right\}$$

with respect to U and coincides with $\mathbb{P}[X_{n+m+1} \in \mathcal{G}_{k,n+m} \,|\, X_1, \ldots, X_n]$. For $k = 0$, the $[m : 0]$–discovery coincides with the probability of discovering a new species at the $(n+m+1)$–th step given the basic sample of size n, which has been derived in Lijoi et al. (2007a) and is recalled in (7). The case of $k \geq 1$ is dealt with by the following result.

THEOREM 2. *If $(X_n)_{n\geq 1}$ is an exchangeable sequence with Gibbs–type prior Π in (1), a Bayesian nonparametric estimator of the $[m : k]$–discovery coincides with*

$$\begin{aligned} \widehat{U}_{n+m}(k) = \sum_{i=1}^{k} \ell_i \, (i-\sigma)_{k+1-i} \binom{m}{k-i} Q_{m,k}^{(n,j)}(i, 0, i-\sigma) \\ + \sigma(1-\sigma)_k \binom{m}{k} Q_{m,k}^{(n,j)}(1, 1, 0) \end{aligned} \tag{12}$$

where

$$Q_{m,l}^{(n,j)}(\alpha, \beta, \gamma) := \sum_{k=\beta}^{m-l+\alpha} \frac{V_{n+m+1,j+k}}{V_{n,j}} \frac{\mathscr{C}(m-l+\alpha-\beta, k-\beta; \sigma, -n+j\sigma+\gamma)}{\sigma^k}$$

and $\ell_i \geq 0$ is the number of species observed with frequency i in the sample $(X_1, \ldots, X_n)$.

The analytic form of the estimator in (12) implies that the vector of the number of species and frequency counts $(K_n, \mathcal{L}_{1,n}, \ldots, \mathcal{L}_{k,n})$ is sufficient for estimating the $[m : k]$–discovery. It is worth remarking that the novel estimator in Theorem 2 does not have any counterpart both in the frequentist and in the Bayesian frameworks. As recalled in the introduction, frequentist estimators for m–step k–discovery exist only when either $m = 0$ and $k \geq 0$ or $m \geq 1$ and $k = 0$:

the former corresponds to the Turing estimator displayed in (3), whereas the latter corresponds to the so–called Good–Toulmin estimator (Good and Toulmin, 1956), which however becomes unstable for $m \geq 2n$ (see e.g. Lijoi et al. (2007b) for an illustration).

Obviously, from (12) one can also deduce estimators for cumulative $[m : k]$–discoveries, which allow to evaluate the probability of observing a rare species at the $(n + m + 1)$–th draw from the population. Hence, for an abundance threshold τ one just needs to evaluate

$$\widehat{U}_{n+m,\tau} = \sum_{k=0}^{\tau} \widehat{U}_{n+m}(k). \tag{13}$$

This implies that $(K_n, \mathcal{L}_{1,n}, \ldots, \mathcal{L}_{\tau,n})$ is sufficient for estimating the $[m : k]$-discovery probability for species with frequencies less than or equal to τ, conditionally on the basic sample. An important application of the estimator in (13) is related to sample size determination. Indeed, one is often willing to determine the size of the additional sample m as the smallest integer m^* for which $\widehat{U}_{n+m^*,\tau}$ is not smaller than a certain threshold κ. See, e.g., Christen and Nakamura (2003). The rationale is that the smaller the probability of sampling rare species the less informative is the sampling procedure since, in this case, it mainly consists of re–observing species with large abundances. Hence, a further enlargement of the sample size beyond m^* would not yield enough relevant information compared the cost of further sampling. For instance, in applications one often sets such a threshold to $\kappa \in \{0.1, 0.2, 0.5\}$.

The result in Theorem 2 allows to obtain a decomposition of the sample coverage estimator, given an observed sample of size n and an additional unobserved sample of size m, $\widehat{C}_{n+m}$. Indeed, one has that

$$\widehat{C}_{n+m} = \sum_{k=1}^{n+m} \widehat{U}_{n+m}(k) \tag{14}$$

which provides an alternative derivation of $\widehat{C}_{n+m}$ w.r.t. the derivation based on (8). Similarly to the case $m = 0$, based on Theorem 2, one can deduce an estimator for *abundant species coverage* after sampling $n + m$ species, which is given by the proportion of species with frequency larger than τ represented in both the observed basic and unobserved additional sample, i.e.

$$\widehat{C}^{abund}_{n+m} = 1 - \sum_{k=0}^{\tau} \widehat{U}_{n+m}(k).$$

2.3 The two–parameter Poisson–Dirichlet process

As already mentioned, a prominent special case of Gibbs type prior is the two–parameter Poisson–Dirichlet process. The system of predictive distributions it induces for the data in (1) can be deduced from (6) by recalling the simple form of its weights $V_{n,j}$'s, namely $V_{n,j} =$

$\prod_{i=1}^{j-1}(\theta+i\sigma)/(\theta+1)_{n-1}$, which leads to

$$\mathbb{P}[X_{n+1}\in A\,|\,X_1,\ldots,X_n]=\frac{\theta+k\sigma}{\theta+n}\,P_0(A)+\frac{1}{\theta+n}\sum_{i=1}^{j}(n_i-\sigma)\,\delta_{X_i^*}(A) \tag{15}$$

for any set A. The vector of parameters is such that $\sigma\in(0,1)$ and $\theta>-\sigma$ and we shall, henceforth, adopt the concise notation $\mathrm{PD}(\sigma,\theta;P_0)$ for denoting such a prior. Its specification is particularly appealing: besides being analytically tractable for addressing various inferential issues in a Bayesian setting, it also represents a very flexible model, as detailed, e.g., in Lijoi et al. (2007a). For these reasons, it has become an increasingly popular prior in several applications such as, e.g., mixture models (Ishwaran and James, 2001), linguistics (Teh, 2006), species sampling models (Navarrete et al., 2008), information retrieval in document modeling (Teh and Jordan, 2010) and survival analysis (Jara et al., 2010). And while its use can be considered as routine in Bayesian nonparametrics modeling, its effective implementation has become easily accessible also to practitioners due to the availability of an efficient R package that embodies various MCMC algorithms with the two–parameter Poisson–Dirichlet process. See Jara et al. (2011) for details.

Within the present framework, Lijoi et al. (2007b); Favaro et al. (2009) provided closed form expressions for $\mathrm{PD}(\sigma,\theta;P_0)$–estimators of: (i) the expected number of new species $K_m^{(n)}$ that will be observed from a further sampling of size m and (ii) the $[m:0]$–discovery as described in (8). However, as mentioned before, it is important to extend the analysis to rare species, whose quantification is often a major goal not only in EST analysis but also in different genomic applications such as Cap Analysis Gene Expression (CAGE) (see, e.g., Valen (2009)). Based on the general results of Section 2.2, here we provide completely explicit estimators of the $[m:k]$–discovery in this noteworthy particular case.

Proposition 3. *If $(X_n)_{n\geq1}$ is an exchangeable sequence with Π in (1) being a Poisson–Dirichlet process with parameters (σ,θ), then*

$$\widehat{U}_{n+0}(k)=\frac{k-\sigma}{\theta+n}\ell_k \tag{16}$$

where $\ell_k\geq0$ is the number of distinct species with frequency k observed within the sample $(X_1,\ldots,X_n)$. Moreover,

$$\begin{aligned}\widehat{U}_{n+m}(k)=&\sum_{i=1}^{k}\ell_i(i-\sigma)_{k+1-i}\binom{m}{k-i}\frac{(\theta+n-i+\sigma)_{m-k+i}}{(\theta+n)_{m+1}}\\&+(1-\sigma)_k\binom{m}{k}(\theta+j\sigma)\frac{(\theta+n+\sigma)_{m-k}}{(\theta+n)_{m+1}}.\end{aligned} \tag{17}$$

From the previous result one then immediately deduces the cumulative discovery estimator $\widehat{U}_{n+m,\tau}$. Finally, for completeness we also consider the special case of the Dirichlet process with parameter $\theta > 0$ and baseline probability measure P_0, which coincides, in distribution, with the PD$(0, \theta; P_0)$ process. Therefore, from (16)–(17) it is straightforward to see that the corresponding estimators reduce to, respectively, $\widehat{U}_{n+0}(k) = k\,\ell_k/(\theta + n)$ and

$$\widehat{U}_{n+m}(k) = k! \sum_{i=1}^{k} \frac{\ell_i}{(i-1)!} \binom{m}{k-i} \frac{(\theta + n - i)_{m-k+i}}{(\theta + n)_{m+1}} + \binom{m}{k} \frac{k!}{(\theta + n)_{m+1}} \theta\,(\theta + n)_{m-k}$$

It is interesting to note that in the Dirichlet case $\widehat{U}_{n+0}(k+1) = (\theta + n)\check{U}_{n+0}(k)/n$ where $\check{U}_{n+0}$ is the Turing estimator in (3).

3 An application to genomic data

The genomic datasets we analyze consist of expressed sequence tags (ESTs) samples, which play an important role in the identification, discovery and characterization of organisms as they provide an attractive and efficient alternative to full genome sequencing. ESTs are single–read sequences of *c*omplementary DNA (cDNA) fragments obtained by sequencing randomly selected cDNA clones from a cDNA library. Since a cDNA library consists of millions of cDNA clones, only a small fraction is usually sequenced because of cost constraints. This is a natural setting in which prediction and, in particular, the estimation of discovery probabilities is of great relevance.

We focus on EST data obtained from *Naegleria gruberi* cDNA libraries. Naegleria gruberi is a widespread free–living soil and freshwater *amoeboflagellate* widely studied in the biological literature. The two datasets we consider are sequenced from two cDNA libraries prepared from cells grown under different culture conditions, aerobic and anaerobic. These data sets have been previously analyzed in Susko and Roger (2004), where a full account of their preparation is detailed. The datasets, which will constitute our basic samples, can be effectively summarized in terms of their frequency counts ℓ_k, for $k = 1, \ldots, n$ and are reported in Table 1 below. One observes that the frequency counts ℓ_k are large for small values of k and, for initial values of k, ℓ_k decreases monotonically with a few isolated points of increase for larger values of k. These are features common to most EST datasets and to samples drawn from genomic libraries in general. Moreover, the *Naegleria Aerobic* dataset exhibits several genes with high frequency and, in particular, two genes appearing 27 and 55 times, respectively. On the other hand, the *Naegleria Anaerobic* dataset yields $\ell_k = 0$ for any $k > 14$.

Table 1: *ESTs from two Nageleria gruberi libraries. Reported data include: frequency counts ℓ_k, for different values of k, total number of distinct genes j and sample size n. Source: Susko and Roger (2004).*

Library	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Naeglaria Aerobic	346	57	19	12	9	5	4	2	4	5	4
Naeglaria Anaerobic	491	72	30	9	13	5	3	1	2	0	1

Library	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*27*	*55*	j	n
Naeglaria Aerobic	1	0	0	0	1	1	1	1	1	473	959
Naeglaria Anaerobic	0	1	3	0	0	0	0	0	0	631	969

The nonparametric prior in (1) we adopt for analyzing these datsets is given by the two–parameter Poisson–Dirichlet process, concisely discussed in Section 2.3. We will focus on quantifying the rare species variety and provide estimates of both the $[m : k]$–discovery and of the cumulative discovery $U_{n+m,\tau}$ for different values of the additional sample size m, frequency k and threshold τ. It is worth noting that such an analysis, still within the two–parameter Poisson–Dirichlet framework, has been carried out in Lijoi et al. (2007b): there the analysis is focused on overall species variety and a comparison of the heterogeneity of the two cDNA libraries, which in this context is typically quantified in terms of their redundancy (the counterpart of species variety), is carried out. According to such a prior specification, one is left to eliciting the actual values of σ and θ that will be used in the inferential process. A natural way to fix these parameters is through an empirical Bayes argument. In particular, $(\widehat{\sigma}, \widehat{\theta})$ are fixed so to maximize the distribution of $\boldsymbol{\mathcal{L}}^{(n)}$ in correspondence to the observed frequency counts $(\ell_1, \ldots, \ell_n)$, i.e. $(\widehat{\sigma}, \widehat{\theta})$ result from

$$\arg\max_{(\sigma,\theta)} \mathbb{P}\left[\boldsymbol{\mathcal{L}}^{(n)} = (\ell_1, \ldots, \ell_n)\right] = \arg\max_{(\sigma,\theta)} \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{(\theta+1)_{n-1}}\, n! \prod_{i=1}^{n} \left\{\frac{(1-\sigma)_{i-1}}{i!}\right\}^{\ell_i} \frac{1}{\ell_i!}. \tag{18}$$

For the two considered libraries we then obtain $(\widehat{\sigma}, \widehat{\theta}) = (0.67, 46.3)$ in the aerobic case and $(\widehat{\sigma}, \widehat{\theta}) = (0.66, 155.5)$ in the anaerobic case. Alternatively, one could specify a prior for (σ, θ) and devise a suitable Markov Chain Monte Carlo scheme for sampling from their posteriors as in Lijoi et al. (2008a). However, with such relatively large basic samples, the posteriors are highly peaked and the results essentially coincide.

In Table 2 we show the estimates of the $[0 : k]$–discoveries and of the cumulative $[0 : k]$–discoveries for different values of k and of τ resulting from the Bayesian nonparametric and the Turing estimators. The estimates are basically the same for $k = 0$ and $k = 1$, which

Table 2: *Bayesian nonparametric and Turing estimates, of the* $[0:k]$*-discoveries (for* $k=0,\dots,4$*) and of the cumulative* $[0:k]$*-discoveries (for* $\tau=3,4,5$*) for the two EST datasets.*

	Naegleria aerobic		*Naegleria anaerobic*	
	BNP	Turing	BNP	Turing
$U_{n+0}(0)$	0.3613	0.3608	0.5086	0.5067
$U_{n+0}(1)$	0.1136	0.1189	0.1485	0.1486
$U_{n+0}(2)$	0.0754	0.0594	0.0858	0.0929
$U_{n+0}(3)$	0.0440	0.0501	0.0624	0.0372
$U_{n+0}(4)$	0.0397	0.0469	0.0267	0.0671
$U_{n+0,3}$	0.5943	0.5892	0.8053	0.7854
$U_{n+0,4}$	0.6341	0.6361	0.8320	0.8524
$U_{n+0,5}$	0.6728	0.6674	0.8822	0.8834

correspond to the probabilities of sequencing a new gene and a unique gene, respectively. Larger discrepancies are detectable as k gets bigger. This is also reflected in the estimation of the cumulative discoveries. With reference to the discussion in Remark 1, it is worth noting that the *Naegleria anaerobic* basic sample displays two species with frequency 9 meaning that $\ell_9=2$ and no species with frequency equal to 10, i.e. $\ell_{10}=0$. In such a case the Turing estimator (3) that depends on $\ell_{k+1}=\ell_{10}$ is 0: one would then conclude that the probability that the $(n+1)$–th observation coincides with one of the two species appearing 9 times in the basic sample is zero. Moreover, if we were to estimate the $[0:10]$ discovery, the frequentist estimator would lead to a positive value, since $\ell_{11}>0$, a conclusion that seems again counterintuitive since there are no species with frequency equal to 10 in the basic sample. This puzzling feature is due to the fact the estimator of the $[0:k]$–discovery in (3) depends on ℓ_{k+1}. In contrast, the Bayesian nonparametric estimator (10) is positive and actually equal to 0.014833 for the $[0:9]$–discovery and 0 for the $[0:10]$–discovery.

In addition to one–step prediction estimates, we are also able to provide estimates of $U_{n+m}(k)$, and for the corresponding cumulative values, in the case where both m and k are strictly positive. These represent the main quantities of interest in such genomic experiments and as noted in Section 2 no other estimators are available in the literature. The results are summarized in Table 3. It is worth remarking that the displayed numbers are exact, in the sense that (12) and (13) can be computed without the need to resorting to any approximation scheme.

The behaviour of the cumulative $[m:k]$–discoveries, that provide evidence of the possibility of detecting rare genes as the additional sample size increases, is consistent with intuition: the

Table 3: *Estimates of the $[m:k]$-discoveries (for $k = 0, \ldots, 4$) and of the cumulative $[m:k]$–discoveries (for $\tau = 3, 4, 5$) for different sizes of the additional sample m.*

	Naegleria Aerobic					
	$m = 250$	$m = 500$	$m = 750$	$m = 1000$	$m = 1250$	$m = 1500$
$U_{n+m}(0)$	0.3358	0.3162	0.3006	0.2877	0.2768	0.2673
$U_{n+m}(1)$	0.1066	0.1011	0.0965	0.0927	0.0894	0.0865
$U_{n+m}(2)$	0.0703	0.0664	0.0634	0.0609	0.0587	0.0569
$U_{n+m}(3)$	0.0475	0.0476	0.0467	0.0455	0.0443	0.0432
$U_{n+m}(4)$	0.0373	0.0370	0.0366	0.0361	0.0355	0.0348
$U_{n+m,3}$	0.5602	0.5313	0.5072	0.4867	0.4692	0.4539
$U_{n+m,4}$	0.5974	0.5683	0.5438	0.5228	0.5046	0.4887
$U_{n+m,5}$	0.6307	0.5996	0.5743	0.5528	0.5342	0.5178
	Naegleria Anaerobic					
	$m = 250$	$m = 500$	$m = 750$	$m = 1000$	$m = 1250$	$m = 1500$
$U_{n+m}(0)$	0.4751	0.4489	0.4275	0.4097	0.3945	0.3813
$U_{n+m}(1)$	0.1428	0.1377	0.1330	0.1289	0.1251	0.1218
$U_{n+m}(2)$	0.0849	0.0834	0.0817	0.0800	0.0783	0.0767
$U_{n+m}(3)$	0.0612	0.0602	0.0593	0.0584	0.0575	0.0565
$U_{n+m}(4)$	0.0388	0.0429	0.0443	0.0447	0.0446	0.0444
$U_{n+m,3}$	0.7639	0.7301	0.7015	0.6769	0.6554	0.6363
$U_{n+m,4}$	0.8027	0.7729	0.7458	0.7216	0.7000	0.6807
$U_{n+m,5}$	0.8384	0.8074	0.7809	0.7572	0.7360	0.7167

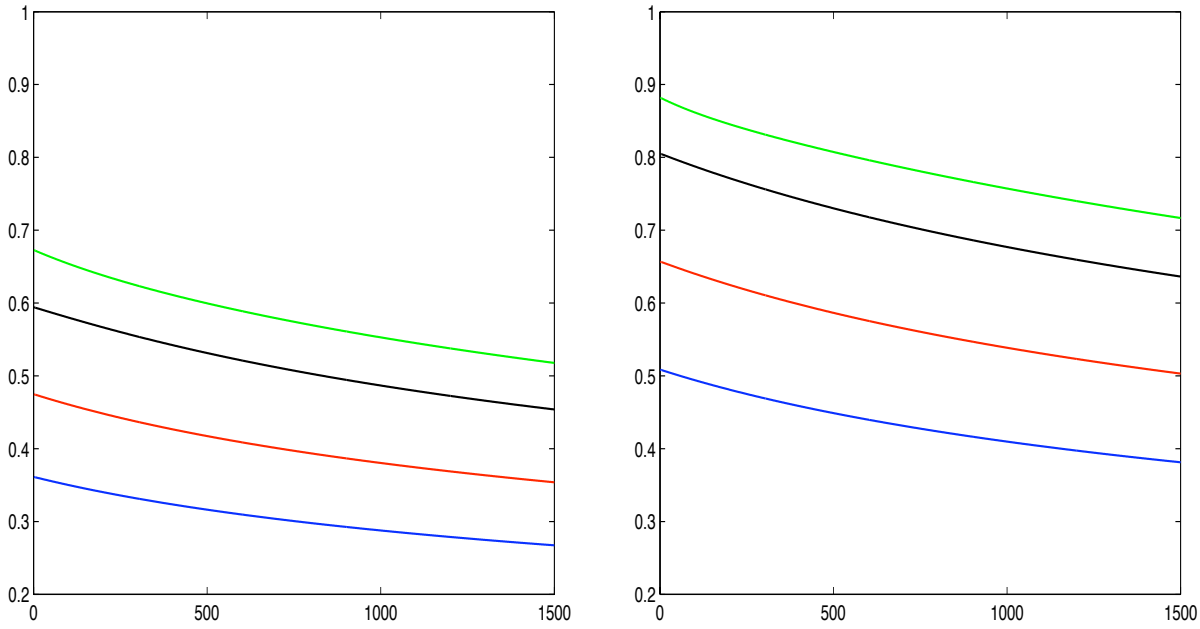

Figure 1: *Cumulative* $[m : k]$*–discovery estimates for* $\tau = 0$ *(blue line),* 1 *(red line),* 3 *(black line),* 5 *(green line) and additional sample sizes* $m = 0, \ldots, 1500$*. Left and right panels refer, respectively, to Naegleria Aerobic and Anaerobic datasets.*

larger the additional sample size m, the lower the probability that the $(n+m+1)$–th observation identifies a rare gene i.e. a gene that has never been observed before or that has been observed with frequency not greater than τ. Moreover, the larger is the considered frequency level k the slower is the decay rate. The two datasets display quite different results, which become even more apparent from Figure 1. In particular, the *Naegleria Anaerobic* is a richer (or, equivalently, in biological terminology, less redundant) library in the sense that the probability of detecting rare species is significantly larger if compared with the *Naegleria Aerobic* library. This is consistent with the fact that the basic sample of the *Naegleria Aerobic* library exhibits several genes with high frequency, which in fact hints towards redundancy. As anticipated in Section 2.2, one can also address the issue of sample size determination. If we agree that $\tau = 3$ defines the upper bound for the frequency of species considered as rare, one can consider $m \mapsto \widehat{U}_{n+m,3}$ and identify the largest m that makes $\widehat{U}_{n+m}$ not smaller than a certain threshold κ. For example, still referring to the *Naegleria Aerobic* dataset, if we were to choose $\kappa = 0.5$ then one should set a value of the additional sample size $m = 833$.

4 Appendix

Resorting to the notation set forth in Sections 2 and 3 we introduce some useful quantities induced by the basic sample $\boldsymbol{X}_1^{(n)} = (X_1, \ldots, X_n)$ and the additional sample $\boldsymbol{X}_2^{(m)} = (X_{n+1}, \ldots, X_{n+m})$. If $X_1^*, \ldots, X_{K_n}^*$ are the labels identifying the K_n distinct species detected in the basic sample, define

$$L_m^{(n)} = \sum_{i=1}^{m} \prod_{j=1}^{K_n} \mathbb{1}_{\{X_{n+i} \neq X_j^*\}} = \text{card}\left(\{X_{n+1}, \ldots, X_{n+m}\} \cap \{X_1^*, \ldots, X_{K_n}^*\}^c\right) \tag{19}$$

to be the number of observations from the additional sample that generate the new $K_m^{(n)}$ distinct species not coinciding with any of those already observed in the basic sample. Moreover, if $X_{K_n+1}^*, \ldots, X_{K_n+K_m^{(n)}}^*$ are the labels identifying the $K_m^{(n)}$ new distinct species detected in the additional sample, define

$$S_i^* = \sum_{j=1}^{m} \mathbb{1}_{\{X_{n+j} = X_{K_n+i}^*\}} \qquad i = 1, \ldots, K_m^{(n)} \tag{20}$$

to be the frequencies of the $K_m^{(n)}$ new distinct species detected among the $L_m^{(n)}$ observations in the additional sample. Similarly, define

$$V_i = \sum_{j=1}^{m} \mathbb{1}_{\{X_{n+j} = X_i^*\}} \qquad i = 1, \ldots, K_n \tag{21}$$

as the number of observations from the additional sample coinciding with the i–th distinct species detected in the basic sample. Introduce, now, the partition sets

$$\mathcal{D}(L_m^{(n)}, K_m^{(n)}) := \{(S_1^*, \ldots, S_{K_m^{(n)}}^*) : S_i^* > 0 \text{ and } \sum_{i=1}^{K_m^{(n)}} S_i^* = L_m^{(n)}\}.$$

$$\mathcal{D}_0(m - L_m^{(n)}, K_n) := \{(V_1, \ldots, V_{K_n}) : V_i \geq 0 \text{ and } \sum_{i=1}^{K_n} V_i = m - L_m^{(n)}\}.$$

According to these definitions the random vectors $(S_1^*, \ldots, S_{K_m^{(n)}}^*)$ and $(V_1, \ldots, V_{K_n})$ take values in $\mathcal{D}(L_m^{(n)}, K_m^{(n)})$ and in $\mathcal{D}_0(m - L_m^{(n)}, K_n)$, respectively. The quantities defined in (19)–(21) can obviously be extended to include a further additional sample $\boldsymbol{X}_3^{(m')} = (X_{n+m+1}, \ldots, X_{n+m+m'})$, for some $m' \geq 1$. Indeed

$$L_{m'}^{(n+m)} = \sum_{i=1}^{m'} \prod_{j=1}^{K_n + K_m^{(n)}} \mathbb{1}_{\{X_{n+m+i} \neq X_j^*\}} \tag{22}$$

to be the number of observations from a further additional sample of size m' that generate the new $K_{m'}^{(n+m)}$ distinct species not coinciding with any of those already observed in the basic

sample and in the additional sample. Moreover, we can define

$$S'_i = \sum_{j=1}^{m'} \mathbb{1}_{\{X_{n+m+j}=X^*_{K_n+K_m^{(n)}+i}\}} \qquad i = 1, \ldots, K_{m'}^{(n+m)} \tag{23}$$

to be the frequencies of the new distinct species detected among the $L_{m'}^{(n+m)}$ observations and

$$V'_i = \sum_{j=1}^{m'} \mathbb{1}_{\{X_{n+m+j}=X^*_i\}} \qquad i = 1, \ldots, K_n + K_m^{(n)} \tag{24}$$

to be the number of observations coinciding with the i–th distinct species detected in the basic sample and the additional sample.

Finally, we set $A_n(j, \boldsymbol{n}_n) = \{K_n = j,\, \boldsymbol{N}_n = \boldsymbol{n}_n\}$ as the event identifying a partition of $[n] = \{1, \ldots, n\}$ into j clusters with vector of respective frequencies $\boldsymbol{n}_n = (n_{1,n}, \ldots, n_{j,n})$. Correspondingly

$$\widetilde{\Pi}^{(l)}_{s,m}(\boldsymbol{v}, \boldsymbol{s} \,|\, A_n(j, \boldsymbol{n}_n)) = \mathbb{P}[L_m^{(n)} = s,\, K_m^{(n)} = l,\, \boldsymbol{V} = \boldsymbol{v},\, \boldsymbol{S}^* = \boldsymbol{s} \,|\, A_n(j, \boldsymbol{n}_n)] \tag{25}$$

is the conditional probability that the additional sample of size m, $X_{n+1}, \ldots, X_{n+m}$, has exactly s observations generating l new species with positive frequencies $\boldsymbol{s} = (s_1, \ldots, s_l)$ and $m-s$ replicating any of the j previously observed species with non–negative frequencies $\boldsymbol{v} = (v_1, \ldots, v_j)$, given the basic sample of size n.

Proof of Theorem 1. Note that, since Π is a Gibbs–type prior, we can easily derive the following conditional probability distribution

$$\widetilde{\Pi}^{(l)}_{s,m}(\boldsymbol{v}, \boldsymbol{s} \,|\, A_n(j, \boldsymbol{n}_n)) = \frac{V_{n+m,j+l}}{V_{n,j}} \frac{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i-1} \prod_{i=1}^{l}(1-\sigma)_{s_i-1}}{\prod_{i=1}^{j}(1-\sigma)_{n_i-1}} \times \mathbb{1}_{\{1,\ldots,m\}\times\{1,\ldots,s\}\times\mathcal{D}_0(m-s,j)\times\mathcal{D}(s,l)}(s, l, \boldsymbol{v}, \boldsymbol{s}) \tag{26}$$

where $\boldsymbol{v} := (v_1, \ldots, v_j)$ and $\boldsymbol{s} := (s_1, \ldots, s_l)$.

$$\widetilde{\Pi}^{(0)}_{0,m}(\boldsymbol{v} \,|\, A_n(j, \boldsymbol{n}_n)) = \frac{V_{n+m,j}}{V_{n,j}} \frac{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i-1}}{\prod_{i=1}^{j}(1-\sigma)_{n_i-1}} \mathbb{1}_{\mathcal{D}_0(m,j)}(\boldsymbol{v}). \tag{27}$$

Hence, under squared loss we have

$$\widehat{U}_n(k) = \mathbb{P}[X_{n+1} \in \mathcal{G}_{k,n} \,|\, A_n(j, \boldsymbol{n}_n)] = \sum_{i=1}^{j} \widetilde{\Pi}^{(0)}_{0,1}(\boldsymbol{e}_i(k) \,|\, A_n(j, \boldsymbol{n}_n))$$

where $\boldsymbol{e}_i(k) = (0, \ldots, 0, k+1-n_i, 0, \ldots, 0)$ is the j–dimensional vector with all zero entries but the i–th that coincides with $k+1-n_i$ and $\mathcal{G}_{k,n} = \{X_i^* :\, N_{i,n} = k\}$. From (27) one immediately

deduces that

$$\sum_{i=1}^{j} \widetilde{\Pi}^{(0)}_{0,1}(\boldsymbol{e}_i(k) \,|\, A_n(j,\boldsymbol{n}_n)) = \frac{V_{n+1,j}}{V_{n,j}}(1-\sigma)_k \sum_{i=1}^{j} \frac{\prod_{1\le t\neq i\le j}(1-\sigma)_{n_t-1}}{\prod_{i=1}^{j}(1-\sigma)_{n_i-1}}\, \mathbb{1}_{\{k\}}(n_i).$$

Therefore, we have

$$\widehat{U}_n(k) = \frac{V_{n+1,j}}{V_{n,j}}(1-\sigma)_k \sum_{i=1}^{j} \frac{1}{(1-\sigma)_{n_i-1}} \mathbb{1}_{\{k\}}(n_i) = \frac{V_{n+1,j}}{V_{n,j}}\,(k-\sigma)\,\ell_k$$

and the proof is complete. □

Proof of Theorem 2 In order to simplify the notation, we introduce some quantities to be used in the sequel. In particular, using (19)–(24) we define the vectors

$$\boldsymbol{V}' := (V'_1, \ldots, V'_{K_n+K_m^{(n)}}), \qquad \boldsymbol{S}' := (S'_1, \ldots, S'_{K_{m'}^{(n+m)}})$$

and the set

$$A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s}) := \{K_n = j, \boldsymbol{N}_n = \boldsymbol{n}_n, L_m^{(n)} = s, K_m^{(n)} = l, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{S} = \boldsymbol{s}\}.$$

Let, further,

$$\begin{aligned}\widetilde{\Pi}^{(k')}_{s',m'}(\boldsymbol{v}',\boldsymbol{s}' \,|\, A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s}))& \\ = \mathbb{P}[L_{m'}^{(n+m)} = s', K_{m'}^{(n+m)} = k', \boldsymbol{V}' = \boldsymbol{v}', \boldsymbol{S}' = \boldsymbol{s}' \,|\, A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s})]&.\end{aligned}$$

Since Π is a Gibbs–type prior, one can easily derive the following conditional probability distribution

$$\begin{aligned}\widetilde{\Pi}^{(k')}_{s',m'}(\boldsymbol{v}',\boldsymbol{s}' \,|\, A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s})) &= \frac{V_{n+m+m',j+l+k'}}{V_{n+m,j+l}} \\ \times \frac{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i+v'_i-1}\prod_{r=1}^{l}(1-\sigma)_{s_{j+r}+v'_{j+r}-1}\prod_{t=1}^{k'}(1-\sigma)_{s'_t-1}}{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i-1}\prod_{r=1}^{l}(1-\sigma)_{s_{j+r}-1}}& \\ \times \mathbb{1}_{\{1,\ldots,m'\}\times\{1,\ldots,s'\}\times\mathcal{D}_0(m'-s',j+l)\times\mathcal{D}(s',k')}(s',k',\boldsymbol{v}',\boldsymbol{s}'),&\end{aligned} \tag{28}$$

where $\boldsymbol{v}' := (v'_1, \ldots, v'_{j+l})$ and $\boldsymbol{s}' := (s'_1, \ldots, s'_{k'})$. Hence, (28) corresponds to the probability distribution of the random partition induced by the further additional sample of size m', given both the basic sample $\boldsymbol{X}_1^{(n)}$ and the additional sample $\boldsymbol{X}_2^{(m)}$. To deal with the case $s' = k' = 0$, namely none of the additional m' observations generates new species that differ from those already observed in the previous sample of size $n+m$, one has

$$\widetilde{\Pi}^{(0)}_{0,m'}(\boldsymbol{v}' \,|\, A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s})) = \frac{V_{n+m+m',j+l}}{V_{n+m,j+l}} \times \frac{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i+v'_i-1}\prod_{r=1}^{l}(1-\sigma)_{s_{j+r}+v'_{j+r}-1}}{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i-1}\prod_{r=1}^{l}(1-\sigma)_{s_{j+r}-1}}\,\mathbb{1}_{\mathcal{D}_0(m',j+l)}(\boldsymbol{v}'). \quad (29)$$

For $m'=1$, from (29) one deduces

$$\widetilde{\Pi}^{(0)}_{0,1}(\boldsymbol{v}' \,|\, A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s})) = \frac{V_{n+m+1,j+l}}{V_{n+m,j+l}} \times \frac{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i+v'_i-1}\prod_{r=1}^{l}(1-\sigma)_{s_{j+r}+v'_{j+r}-1}}{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i-1}\prod_{r=1}^{l}(1-\sigma)_{s_{j+r}-1}}\,\mathbb{1}_{\mathcal{D}_0(1,j+l)}(\boldsymbol{v}'). \quad (30)$$

Hence

$$\begin{aligned}\widehat{U}_{n+m}(k) = \sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\sum_{\boldsymbol{v}\in\mathcal{D}_0(m-s,j)}\binom{m-s}{v_1,\ldots,v_j} \quad (31)\\ \times\frac{1}{l!}\sum_{\boldsymbol{s}\in\mathcal{D}(s,l)}\binom{s}{s_1,\ldots,s_l}\mathbb{P}[X_{n+m+1}\in\mathcal{G}_{k,n+m}\,|\,A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s})]\\ \times\widetilde{\Pi}^{(l)}_{s,m}(\boldsymbol{v},\boldsymbol{s}\,|\,A_n(j,\boldsymbol{n}_n)).\end{aligned}$$

For any $i=1,\ldots,j$, let us introduce $\boldsymbol{e}_i(k)=(0,\ldots,0,k+1-n_i-v_i,0,\ldots,0)$, a $(j+l)$–dimensional vector whose components are all zero but the i-th which is equal to $k+1-n_i-v_i$. Analogously, for any $r=1,\ldots,l$, let $\boldsymbol{g}_r(k)=(0,\ldots,0,k+1-s_r,0,\ldots,0)$ stand for a $(j+l)$–dimensional vector with all zero components but the $(j+r)$–th being equal to $k+1-s_r$. One can accordingly write

$$\begin{aligned}\mathbb{P}[X_{n+m+1}\in\mathcal{G}_{k,n+m}\,|\,A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s})] = \sum_{i=1}^{j}\widetilde{\Pi}^{(0)}_{0,1}(\boldsymbol{e}_i(k)\,|\,A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s}))\\ +\sum_{r=1}^{l}\widetilde{\Pi}^{(0)}_{0,1}(\boldsymbol{g}_r(k)\,|\,A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s})).\end{aligned}$$

This expression suggests that (31) can be decomposed in the sum of the following two quantities:

$$\begin{aligned}\mathcal{O}(\sigma,n,j,\boldsymbol{n}_n,k) := \sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\sum_{\boldsymbol{v}\in\mathcal{D}_0(m-s,j)}\binom{m-s}{v_1,\ldots,v_j}\\ \times\frac{1}{l!}\sum_{\boldsymbol{s}\in\mathcal{D}(s,l)}\binom{s}{s_1,\ldots,s_l}\widetilde{\Pi}^{(l)}_{s,m}(\boldsymbol{v},\boldsymbol{s}\,|\,A_n(j,\boldsymbol{n}_n))\sum_{i=1}^{j}\widetilde{\Pi}^{(0)}_{0,1}(\boldsymbol{e}_i(k)\,|\,A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s}))\end{aligned}$$

and

$$\mathcal{N}(\sigma,n,j,k) := \sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\sum_{\boldsymbol{v}\in\mathcal{D}_0(m-s,j)}\binom{m-s}{v_1,\ldots,v_j}$$

$$\times \frac{1}{l!} \sum_{\boldsymbol{s} \in \mathcal{D}(s,l)} \binom{s}{s_1, \ldots, s_l} \widetilde{\Pi}_{s,m}^{(l)}(\boldsymbol{v}, \boldsymbol{s} \,|\, A_n(j, \boldsymbol{n}_n)) \sum_{r=1}^{l} \widetilde{\Pi}_{0,1}^{(0)}(\boldsymbol{g}_r(k) \,|\, A_{n,m}(j, \boldsymbol{n}_n, s, l, \boldsymbol{v}, \boldsymbol{s})).$$

We first consider the term $\mathcal{O}(\sigma, n, j, \boldsymbol{n}_n, k)$. According to expression (30) it can be easily verified that

$$\sum_{i=1}^{j} \widetilde{\Pi}_{0,1}^{(0)}(\boldsymbol{e}_i(k) \,|\, A_{n,m}(j, \boldsymbol{n}_n, s, l, \boldsymbol{v}, \boldsymbol{s})) = \frac{V_{n+m+1,j+l}}{V_{n+m,j+l}} (1-\sigma)_k \times \sum_{i=1}^{j} \frac{\prod_{1 \leq t \neq i \leq j} (1-\sigma)_{n_t+v_t-1}}{\prod_{i=1}^{j} (1-\sigma)_{n_i+v_i-1}} \mathbb{1}_{\{k\}}(n_i + v_i).$$

Therefore, we have

$$\begin{aligned}
\mathcal{O}(\sigma, n, j, \boldsymbol{n}_n, k) &= (1-\sigma)_k \sum_{s=0}^{m} \sum_{l=0}^{s} \binom{m}{s} \frac{V_{n+m+1,j+l}}{V_{n,j}} \\
&\quad \times \sum_{i=1}^{j} \sum_{\boldsymbol{v} \in \mathcal{D}_0(m-s,j)} \binom{m-s}{v_1, \ldots, v_j} \frac{\prod_{1 \leq t \neq i \leq j} (1-\sigma)_{n_t+v_t-1}}{\prod_{i=1}^{j} (1-\sigma)_{n_i-1}} \mathbb{1}_{\{k\}}(n_i + v_i) \\
&\quad \times \frac{1}{l!} \sum_{\boldsymbol{s} \in \mathcal{D}(s,l)} \binom{s}{s_1, \ldots, s_l} \prod_{r=1}^{l} (1-\sigma)_{s_r-1} \\
&= (1-\sigma)_k \sum_{s=0}^{m} \sum_{l=0}^{s} \binom{m}{s} \frac{V_{n+m+1,j+l}}{V_{n,j}} \frac{\mathscr{C}(s,l;\sigma)}{\sigma^l} \\
&\quad \times \sum_{i=1}^{j} \sum_{\boldsymbol{v} \in \mathcal{D}_0(m-s,j)} \binom{m-s}{v_1, \ldots, v_j} \frac{\prod_{1 \leq t \neq i \leq j} (1-\sigma)_{n_t+v_t-1}}{\prod_{i=1}^{j} (1-\sigma)_{n_i-1}} \mathbb{1}_{\{k\}}(n_i + v_i),
\end{aligned}$$

where the last identity is obtained by applying Equation 2.6.1 in Charalambides (2005). Observe that we can write the last expression as follows

$$\begin{aligned}
&(1-\sigma)_k \sum_{s=0}^{m} \sum_{l=0}^{s} \binom{m}{s} \frac{V_{n+m+1,j+l}}{V_{n,j}} \frac{\mathscr{C}(s,l;\sigma)}{\sigma^l} \\
&\quad \times \sum_{i=1}^{j} \sum_{v_i=0}^{m-s} \binom{m-s}{v_i} \mathbb{1}_{\{k\}}(n_i + v_i) \\
&\quad \times \sum_{\boldsymbol{v}_{-i} \in \mathcal{D}_0(m-s-v_i, j-1)} \binom{m-s-v_i}{v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_j} \frac{\prod_{1 \leq t \neq i \leq j} (1-\sigma)_{n_t+v_t-1}}{\prod_{i=1}^{j} (1-\sigma)_{n_i-1}} \\
&= (1-\sigma)_k \sum_{s=0}^{m} \sum_{l=0}^{s} \binom{m}{s} \frac{V_{n+m+1,j+l}}{V_{n,j}} \frac{\mathscr{C}(s,l;\sigma)}{\sigma^l} \\
&\quad \times \sum_{i=1}^{j} \sum_{v_i=0}^{m-s} \binom{m-s}{v_i} \frac{(n - n_i - (j-1)\sigma)_{m-s-v_i}}{(1-\sigma)_{n_i-1}} \mathbb{1}_{\{k\}}(n_i + v_i),
\end{aligned}$$

where the last identity is obtained by applying Lemma A.1 in Lijoi et al. (2008b). It can be easily verified that

$$\sum_{i=1}^{j}\sum_{v_i=0}^{m-s}\binom{m-s}{v_i}\frac{(n-n_i-(j-1)\sigma)_{m-s-v_i}}{(1-\sigma)_{n_i-1}}\,\mathbb{1}_{\{k\}}(n_i+v_i)$$
$$=\sum_{i=1}^{k}\binom{m-s}{k-i}\frac{(n-i-(j-1)\sigma)_{m-s-(k-i)}}{(1-\sigma)_{i-1}}\,\ell_i.$$

Therefore, we obtain

$$\begin{aligned}\mathcal{O}(\sigma,n,j,\boldsymbol{n}_n,k) &= (1-\sigma)_k\sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\frac{V_{n+m+1,j+l}}{V_{n,j}}\frac{\mathscr{C}(s,l;\sigma)}{\sigma^l}\\ &\quad\times\sum_{i=1}^{k}\binom{m-s}{k-i}\frac{(n-i-(j-1)\sigma)_{m-s-(k-i)}}{(1-\sigma)_{i-1}}\,\ell_i\\ &= (1-\sigma)_k\sum_{i=1}^{k}\frac{\ell_i}{(1-\sigma)_{i-1}}\sum_{l=0}^{m}\frac{1}{\sigma^l}\frac{V_{n+m+1,j+l}}{V_{n,j}}\\ &\quad\times\sum_{s=l}^{m}\binom{m}{s}\binom{m-s}{k-i}(n-i-(j-1)\sigma)_{m-s-(k-i)}\mathscr{C}(s,l;\sigma)\\ &= (1-\sigma)_k\sum_{i=1}^{k}\binom{m}{k-i}\frac{\ell_i}{(1-\sigma)_{i-1}}\sum_{l=0}^{m-k+i}\frac{1}{\sigma^l}\frac{V_{n+m+1,j+l}}{V_{n,j}}\\ &\quad\times\mathscr{C}(m-k+i,l;\sigma,-n+i+(j-1)\sigma)\end{aligned}$$

where the last identity is obtained by applying Equation 2.56 in Charalambides (2005). Finally, we consider the term $\mathcal{N}(\sigma,n,j,l)$. According to expression (30) it can be easily verified that

$$\sum_{r=1}^{l}\widetilde{\Pi}^{(0)}_{0,1}(\boldsymbol{g}_r(k)\,|\,A_{n,m}(j,\boldsymbol{n}_n,s,l,\boldsymbol{v},\boldsymbol{s})) = \frac{V_{n+m+1,j+l}}{V_{n+m,j+l}}(1-\sigma)_k$$
$$\times\sum_{r=1}^{l}\frac{\prod_{1\le t\ne r\le l}(1-\sigma)_{s_t-1}}{\prod_{r=1}^{l}(1-\sigma)_{s_r-1}}\,\mathbb{1}_{\{k\}}(s_r).$$

Hence, we have

$$\begin{aligned}\mathcal{N}(\sigma,n,j,k) &= (1-\sigma)_k\sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\frac{V_{n+m+1,j+l}}{V_{n,j}}\\ &\quad\times\sum_{r=1}^{l}\frac{1}{l!}\sum_{\boldsymbol{s}\in\mathcal{D}(s,l)}\binom{s}{s_1,\ldots,s_l}\prod_{1\le t\ne r\le l}(1-\sigma)_{s_t-1}\,\mathbb{1}_{\{k\}}(s_r)\\ &\quad\times\sum_{\boldsymbol{v}\in\mathcal{D}_0(m-s,j)}\binom{m-s}{v_1,\ldots,v_j}\frac{\prod_{i=1}^{j}(1-\sigma)_{n_i+v_i-1}}{\prod_{i=1}^{j}(1-\sigma)_{n_i-1}}\end{aligned}$$

$$= (1-\sigma)_k \sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\frac{V_{n+m+1,j+l}}{V_{n,j}}(n-j\sigma)_{m-s}$$
$$\times \sum_{r=1}^{l}\frac{1}{l!}\sum_{\boldsymbol{s}\in\mathcal{D}(s,l)}\binom{s}{s_1,\ldots,s_l}\prod_{1\le t\ne r\le l}(1-\sigma)_{s_t-1}\,\mathbb{1}_{\{k\}}(s_r),$$

where the last identity is obtained by applying Lemma A.1 in Lijoi et al. (2008b). Observe that we can write the last expression as follows

$$(1-\sigma)_k \sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\frac{V_{n+m+1,j+l}}{V_{n,j}}(n-j\sigma)_{m-s}\,\frac{1}{l!}\sum_{r=1}^{l}\sum_{s_r=1}^{s}\binom{s}{s_r}\mathbb{1}_{\{k\}}(s_r)$$
$$\times \sum_{\boldsymbol{s}_{-r}\in\mathcal{D}(s-s_r,l-1)}\binom{s-s_r}{s_1,\ldots,s_{r-1},s_{r+1},\ldots,s_{r+l}}\prod_{1\le t\ne r\le l}(1-\sigma)_{s_{j+t}-1}$$
$$= (1-\sigma)_k \sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\frac{V_{n+m+1,j+l}}{V_{n,j}}(n-j\sigma)_{m-s}$$
$$\times \frac{(l-1)!}{l!}\sum_{r=1}^{l}\sum_{s_r=1}^{s}\binom{s}{s_r}\frac{\mathscr{C}(s-s_r,l-1;\sigma)}{\sigma^{l-1}}\,\mathbb{1}_{\{k\}}(s_r),$$

where the last identity is obtained by applying Equation 2.6.1 in Charalambides (2005). It can be easily verified that

$$\sum_{r=1}^{l}\sum_{s_r=1}^{s}\binom{s}{s_r}\frac{\mathscr{C}(s-s_r,l-1);\sigma}{\sigma^{l-1}}\,\mathbb{1}_{\{k\}}(s_r) = l\binom{s}{k}\frac{\mathscr{C}(s-k,l-1;\sigma)}{\sigma^{l-1}}.$$

Therefore, we obtain

$$\mathcal{N}(\sigma,n,j,k) = (1-\sigma)_k \sum_{s=0}^{m}\sum_{l=0}^{s}\binom{m}{s}\frac{V_{n+m+1,j+l}}{V_{n,j}}(n-j\sigma)_{m-s}\binom{s}{k}\frac{\mathcal{C}(s-k,l-1;\sigma)}{\sigma^{l-1}}$$
$$= (1-\sigma)_k \sum_{l=0}^{m}\frac{V_{n+m+1,j+l}}{V_{n,j}}\frac{1}{\sigma^{l-1}}\sum_{s=l}^{m}\binom{m}{s}\binom{s}{k}(n-j\sigma)_{m-s}\mathcal{C}(s-k,l-1;\sigma)$$
$$= (1-\sigma)_k \sum_{l=1}^{m-k+1}\frac{V_{n+m+1,j+l}}{V_{n,j}}\frac{1}{\sigma^{l-1}}\binom{m}{k}\mathcal{C}(m-k,l-1;\sigma,-n+j\sigma)$$

where the last identity is obtained by applying Equation 2.56 in Charalambides (2005). The proof is, then, complete. □

Proof of Proposition 3. The expression for $\widehat{U}_{n+0}(k)$ follows from Equation (10) in Theorem 1 by simply inserting the corresponding weights $V_{n,j} = \prod_{i=1}^{j-1}(\theta+i\sigma)/(\theta+1)_{n-1}$. As for $\widehat{U}_{n+m}(k)$, by inserting the weights $V_{n,j} = \prod_{i=1}^{j-1}(\theta+i\sigma)/(\theta+1)_{n-1}$ in Equation (12) in Theorem 2 we obtain

$$\widehat{U}_{n+m}(k) = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}}\sum_{i=1}^{k}\ell_i(i-\sigma)_{k+1-i}\binom{m}{k-i} \tag{32}$$

$$\times \sum_{l=0}^{m-k+i} \prod_{t=j}^{j+l-1} (\theta + t\sigma) \frac{1}{\sigma^l} \mathscr{C}(m-k+i, l; \sigma, -n+j\sigma+i-\sigma)$$

$$+ (1-\sigma)_k \binom{m}{k} \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \sum_{l=1}^{m-k+1} \prod_{t=j}^{j+l-1} (\theta + t\sigma) \frac{1}{\sigma^{l-1}} \mathscr{C}(m-k, l-1; \sigma, -n+j\sigma).$$

We first consider the second part of expression (32), i.e.

$$\begin{aligned}
&(1-\sigma)_k \binom{m}{k} \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \\
&\quad \times \sum_{l=1}^{m-k+1} \prod_{t=j}^{j+l-1} (\theta + t\sigma) \frac{1}{\sigma^{l-1}} \mathscr{C}(m-k, l-1; \sigma, -n+j\sigma) \\
&= (1-\sigma)_k \binom{m}{k} \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \\
&\quad \times (\theta + j\sigma) \sum_{l=0}^{m-k} \left(j+1+\frac{\theta}{\sigma}\right)_l \mathscr{C}(m-k, l; \sigma, -n+j\sigma) \\
&= (1-\sigma)_k \binom{m}{k} (\theta + j\sigma) \frac{(\theta+n+\sigma)_{m-k}}{(\theta+n)_{m+1}}.
\end{aligned} \tag{33}$$

Now we consider the first part of expression (32), i.e.

$$\begin{aligned}
&\frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \sum_{i=1}^{k} \ell_i (i-\sigma)_{k+1-i} \binom{m}{k-i} \\
&\quad \times \sum_{l=0}^{m-k+i} \prod_{t=j}^{j+l-1} (\theta + t\sigma) \frac{1}{\sigma^l} \mathscr{C}(m-k+i, l; \sigma, -n+j\sigma+i-\sigma) \\
&= \sum_{i=1}^{k} \ell_i (i-\sigma)_{k+1-i} \binom{m}{k-i} \frac{\frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m} \prod_{t=1}^{j-1}(\theta+t\sigma)} (\theta+1)_{n-i-1}}{\frac{(\theta+1)_{n-i-1}}{(\theta+1)_{n-i+m-k+i} \prod_{t=1}^{j-1-1}(\theta+t\sigma)} (\theta+1)_{n-i+m-k+i} \prod_{t=1}^{j-1-1} (\theta+t\sigma)} \\
&\quad \times \sum_{l=0}^{m-k+i} \prod_{t=1}^{j-1+l} (\theta + t\sigma) \frac{1}{\sigma^l} \mathscr{C}(m-k+i, l; \sigma, -n+i+\sigma(j-1)).
\end{aligned} \tag{34}$$

By combining Equation (3) with Equation (8) in Favaro et al. (2009) we can write the following identity

$$\begin{aligned}
&\frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m} \prod_{t=1}^{j-1}(\theta+t\sigma)} \sum_{l=0}^{m} \prod_{t=1}^{j+l} (\theta + t\sigma) \frac{1}{\sigma^l} \mathscr{C}(m, l; \sigma, -n+j\sigma) \\
&\quad = \frac{(\theta+j\sigma)(\theta+n+\sigma)_m}{(\theta+n)(\theta+n+1)_m}.
\end{aligned} \tag{35}$$

Therefore, if we apply the identity (35) to expression (34) then we can write

$$\sum_{i=1}^{k} \ell_i (i-\sigma)_{k+1-i} \binom{m}{k-i} \frac{\frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m} \prod_{t=1}^{j-1}(\theta+t\sigma)} (\theta+1)_{n-i-1}}{\frac{(\theta+1)_{n-i-1}}{(\theta+1)_{n-i+m-k+i} \prod_{t=1}^{j-1-1}(\theta+t\sigma)} (\theta+1)_{n-i+m-k+i} \prod_{t=1}^{j-1-1} (\theta+t\sigma)}$$

$$\times \sum_{l=0}^{m-k+i} \prod_{t=1}^{j-1+l} (\theta + t\sigma) \frac{1}{\sigma^l} \mathscr{C}(m-k+i, l; \sigma, -n+i+\sigma(j-1))$$

$$= \sum_{i=1}^{k} \ell_i (i-\sigma)_{k+1-i} \binom{m}{k-i} \frac{\frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m} \prod_{t=1}^{j-1} (\theta+t\sigma)}}{\frac{(\theta+1)_{n-i-1}}{(\theta+1)_{n-i+m-k+i} \prod_{t=1}^{j-1-1} (\theta+t\sigma)}}$$

$$\times \frac{(\theta+(j-1)\sigma)(\theta+n-i+\sigma)_{m-k+i}}{(\theta+n-i)(\theta+n-i+1)_{m-k+i}}$$

$$= \sum_{i=1}^{k} \ell_i (i-\sigma)_{k+1-i} \binom{m}{k-i} \frac{(\theta+n-i+\sigma)_{m-k+i}}{(\theta+n)_{m+1}}. \tag{36}$$

Therefore, based on the simplification (33) and (36), the results follows by adding (33) and (36). □

Acknowledgements

The authors are grateful to to Ole Winther for some stimulating discussions. This work is partially supported by MIUR, Grant 2008MK3AFZ and Regione Piemonte.

References

Barger, K. and Bunge, J. (2010). Objective Bayesian estimation of the number of species. *Bayesian Analysis* **5**, 619–639.

Chao, A. (1981). On estimating the probability of discovering a new species. *Annals of Statistics* **9**, 1339–1342.

Charalambides, C.A. (2005). *Combinatorial methods in discrete distributions*, Hoboken, NJ: Wiley.

Christen, J.A. and Nakamura, M. (2003). Sequential stopping rules for species accumulation. *Journal of Agricultural, Biological and Environmental Statistics* **8**, 184–195.

Clayton, M.K. and Frees, E.W. (1987). Nonparametric estimation of the probability of discovering a new species. *Journal of the American Statistical Association* **82**, 305–311.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87-112.

Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. (2009). Bayesian non–parametric inference for species variety with a two–parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society Series B* **71**, 993–1008.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences (New York)* **140**, 376390.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264.

Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.

Guindani, M. and Müller, P. (2010). A Bayesian Semiparametric model for the analysis of SAGE Data. *Technical Report*.

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Jara, A., Lesaffre, E., De Iorio, M. and Quintana F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Annals of Applied Statistics* **4**, 2126–2149.

Jara, A., Hanson, T.E., Quintana, F.A., Müller, P., Rosner, G.L. (2011). DPpackage: Bayesian semi and nonparametric modeling in R. *Journal of Statistical Software* **40**, 1–30.

Kolossiatis, M., Griffin, J.E. and Steel, M.F.J. (2011). Modeling overdispersion with the normalized tempered stable distribution. *Computational Statistics and Data Analysis* **55**, 2288–2301.

Laird, N.M. and Lange, C. (2010). *The Fundamentals of Modern Statistical Genetics*, Springer.

Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association* **100**, 1278–1291.

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

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

Lijoi, A., Mena, R.H. and Prünster, I. (2007c). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society Series B* **69**, 715-740.

Lijoi A., Mena, R.H. and Prünster, I. (2008a). A Bayesian Nonparametric approach for comparing clustering structures in EST libraries. *Journal of Computational Biology* **15**, 1315–1327.

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

Magurran, A. E. (2003). *Measuring biological diversity*, Wiley–Blackwell, Oxford.

Mao, C.X. and Lindsay, B.G. (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669–681.

Mao, C.X. (2004). Prediction of the conditional probability of discovering a new class. *Journal of the American Statistical Association* **99**, 1108–1118.

Navarrete, C., Quintana, F. and Müller, P. (2008). Some issues on nonparametric Bayesian modeling using species sampling models. *Statistical Modelling* **41**, 3–21.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*. **102**, 145–158.

Sepúlveda, N., Sousa, V., Guindani, M., Müller, P., Paulino, C.D. and Carneiro, J. (2010). Biodiversity estimation: Unraveling the T-cell receptor repertoire in the body's cellular ecosystem. *Technical Report*.

Susko, E. and Roger, A.J. (2004). Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys. *Bioinformatics*, **20**, 2279–2287.

Starr, N. (1979). Linear estimation of the probability of discovering a new species. *Annals of Statistics* **7**, 644–652.

Teh, Y.W. (2006). A Hierarchical Bayesian Language Model based on Pitman-Yor Processes". In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 985-992.

Teh, Y.W., Jordan, M.I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C. Müller, P., Walker, S.G. Eds.), pp. 80–136. Cambridge University Press, Cambridge.

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