Quaderni di Dipartimento

Asymptotics for a Bayesian nonparametric estimator of species richness

Stefano Favaro
(Università degli Studi di Torino e Collegio Carlo Alberto)

Antonio Lijoi
(Università di Pavia e Collegio Carlo Alberto)

Igor Prünster
(Università degli Studi di Torino, Collegio Carlo Alberto e ICER)

144 (05-11)

Dipartimento di economia politica
e metodi quantitativi
Università degli studi di Pavia
Via San Felice, 5
I-27100 Pavia

Maggio 2011

Asymptotics for a Bayesian nonparametric estimator of species richness

S. Favaro[1], A. Lijoi[2] and I. Prünster3

[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, Collegio Carlo Alberto and ICER.
E-mail: igor@econ.unito.it

February 2011

Abstract

In Bayesian nonparametric inference, random discrete probability measures are commonly used as priors within hierarchical mixture models for density estimation and for inference on the clustering of the data. Recently it has been shown that they can also be exploited in species sampling problems: indeed they are natural tools for modeling the random proportions of species within a population thus allowing for inference on various quantities of statistical interest. For applications that involve large samples, the exact evaluation of the corresponding estimators becomes impracticable and, therefore, asymptotic approximations are sought. In the present paper we study the limiting behaviour of the number of new species to be observed from further sampling, conditional on observed data, assuming the observations are exchangeable and directed by a normalized generalized gamma process prior. Such an asymptotic study highlights a connection between the normalized generalized gamma process and the two–parameter Poisson–Dirichlet process that was previously known only in the unconditional case.

Key words and phrases: Bayesian Nonparametrics; Species sampling models; Asymptotics; σ–diversity; Polynomially and exponentially tilted random variables; Completely random measures; Normalized generalized gamma process; Two parameter Poisson–Dirichlet process.

1 Introduction

In species sampling problems one is interested in the species composition of a certain population (of plants, animals, genes etc.) containing an unknown number of species and only a sample drawn from it is available. The relevance of such problems in ecology, biology and, more recently, in genomics and bioinformatics is not surprising. From an inferential perspective, one is willing

to use available data in order to evaluate some quantities of practical interest. The available data specifically consist of a so–called *basic sample* of size n, $(X_1, \ldots, X_n)$, which exhibits $K_n \in \{1, \ldots, n\}$ distinct species, $(X_1^*, \ldots, X_{K_n}^*)$, with respective frequencies $(N_1, \ldots, N_{K_n})$, where clearly $\sum_{i=1}^{K_n} N_i = n$. Given a basic sample, interest mainly lies in estimating the number of new species, $K_m^{(n)} := K_{m+n} - K_n$, to be observed in an additional sample $(X_{n+1} \ldots, X_{n+m})$ of size m and not included among the X_j^*'s, $j = 1, \ldots, K_n$.

Most of the contributions in the literature that address this issue rely on a frequentist approach (see [4, 6] for reviews) and only recently an alternative Bayesian nonparametric approach has been set forth (see, e.g., [10, 24, 26, 12]). The latter resorts to a general class of discrete random probability measures, termed *species sampling models* and introduced by J. Pitman in [32]. Given a non–atomic probability measure P_0 on some complete and separable metric space $\mathbb{X}$, endowed with the Borel σ–field $\mathscr{X}$, a (proper) species sampling model on $(\mathbb{X}, \mathscr{X})$ is a random probability measure

$$\tilde{p} = \sum_{i \geq 1} \tilde{p}_i \, \delta_{Y_i},$$

where $(Y_i)_{i \geq 1}$ is a sequence of independent and identically distributed (i.i.d.) random elements taking values in $\mathbb{X}$ and with probability distribution P_0, the non–negative random weights $(\tilde{p}_i)_{i \geq 1}$ are independent from $(Y_i)_{i \geq 1}$ and are such that $\sum_{i \geq 1} \tilde{p}_i = 1$, almost surely. In the species sampling context, the Y_i's act as species tags and $\tilde{p}_i$ is the random proportion with which the i–th species is present in the population. If $(X_n)_{n \geq 1}$ is an exchangeable sequence directed by a species sampling model $\tilde{p}$, i.e. for every $n \geq 1$ and $A_1, \ldots, A_n$ in $\mathscr{X}$ one has

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

almost surely, then $(X_n)_{n \geq 1}$ is termed *species sampling sequence*. Besides being an effective tool for statistical inference, species sampling models have an appealing structural property established in [32]. Indeed, if $(X_n)_{n \geq 1}$ is a species sampling sequence, then there exists a collection of non–negative weights $\{p_{j,n}(n_1, \ldots, n_k) : 1 \leq j \leq k+1, \, 1 \leq k \leq n, \, n \geq 1\}$ such that $\sum_{j=1}^{k+1} p_{j,n}(n_1, \ldots, n_k) = 1$, for any vector of positive integers $(n_1, \ldots, n_k)$ with $\sum_{j=1}^{k} n_j = n$, and

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n] = p_{K_n+1,n}(n_1, \ldots, n_{K_n}) P_0(\,\cdot\,) + \sum_{j=1}^{K_n} p_{j,n}(n_1, \ldots, n_{K_n}) \delta_{X_j^*}(\,\cdot\,)$$

where $X_1, \ldots, X_n$ is a sample with K_n distinct values $X_1^*, \ldots, X_{K_n}^*$. Statistical applications involving species sampling models for different purposes than those of the present paper are provided, e.g., in [29, 30, 23].

The Bayesian nonparametric approach we undertake postulates that the data are exchangeable and generated by a species sampling model. Then, conditionally on the basic sample of size n, inference is to be made on the number $K_m^{(n)}$ of new distinct species that will be observed in the additional sample of size m. Interest lies in providing both a point estimate and a measure of uncertainty, in the form of a credible interval, for $K_m^{(n)}$ given $(X_1, \ldots, X_n)$. Since the conditional distribution of $K_m^{(n)}$ becomes intractable for large sizes m of the additional sample, one is led to studying its limiting behaviour as m increases. Such asymptotic results, in addition to providing useful approximations to the required estimators, are also of independent theoretical interest since they provide useful insight on the behaviour of the models we focus on. The only discrete random probability measure for which a conditional asymptotic result, similar to the one investigated in this paper, is known, is the two–parameter Poisson–Dirichlet process, shortly denoted as PD(σ, θ). According to [32], a PD (σ, θ) process is a species sampling model characterized by

$$p_{K_n+1,i}(n_1, \ldots, n_{K_n}) = \frac{\theta + K_n \sigma}{\theta + n}\,, \qquad p_{j,n}(n_1, \ldots, n_{K_n}) = \frac{n_j - \sigma}{\theta + n} \tag{2}$$

with $j = 1, \ldots, K_n$, $\sigma \in (0, 1)$ and $\theta > -\sigma$. In this case, [10] provide a result describing the conditional limiting behaviour of $K_m^{(n)}$. In the present paper we focus on an alternative species sampling model, termed *normalized generalized gamma process* in [23]. As we shall see in the next section, it depends on two parameters $\sigma \in (0, 1)$ and $\beta > 0$ and, for the sake of brevity, is denoted by NGG(σ, β). Moreover, it is characterized by

$$p_{K_n+1,n}(n_1, \ldots, n_{K_n}) = \frac{\sigma}{n} \frac{\sum_{l=0}^{n} \binom{n}{l} (-1)^l \beta^{l/\sigma} \Gamma\left(K_n + 1 - \frac{l}{\sigma};\, \beta\right)}{\sum_{l=0}^{n-1} \binom{n-1}{l} (-1)^l \beta^{l/\sigma} \Gamma\left(K_n - \frac{l}{\sigma};\, \beta\right)} \tag{3}$$

$$p_{j,n}(n_1, \ldots, n_{K_n}) = (n_j - \sigma) \frac{\sum_{l=0}^{n} \binom{n}{l} (-1)^l \beta^{l/\sigma} \Gamma\left(K_n - \frac{l}{\sigma};\, \beta\right)}{n \sum_{l=0}^{n-1} \binom{n-1}{l} (-1)^l \beta^{l/\sigma} \Gamma\left(K_n - \frac{l}{\sigma};\, \beta\right)} \tag{4}$$

for any $j \in \{1, \ldots, K_n\}$, where $\Gamma(a; x)$ is the incomplete gamma function.The NGG(σ, β) process prior has gained some attention in the Bayesian literature and it has proved to be useful for various applications such as those considered, e.g., in [23, 1, 2, 14, 15, 16]. It is to be noted that the NGG(σ, θ) does not feature a posterior structure that is as tractable as the one associated to the PD(σ, θ) process (see, e.g., [32, 5, 26, 20]). Nonetheless, in terms of practical implementation, it is possible to devise efficient simulation algorithms that allow for a full Bayesian analysis within models based on a NGG(σ, β) prior. See [25] for a review of such algorithms.

In the present manuscript we will specify the asymptotic behaviour of $K_m^{(n)}$, given the basic sample, as m diverges and highlight the interplay between the conditional distributions of the PD(σ, θ) and the NGG(σ, β) processes. Since the posterior characterization of a NGG(σ, β) process is far more involved than the one associated to the PD(σ, θ) process, the derivation of the

conditional asymptotic results considered in this paper is technically more challenging. This is quite interesting since it suggests that it is possibile to study the limiting conditional behaviour of $K_m^{(n)}$ even beyond species sampling models sharing some sort of conjugacy property. For example, one might conjecture that the same asymptotic regime, up to certain transformations of the limiting random variable, should hold also for the wide class of Gibbs–type priors, to be recalled in Section 2. An up to date account of Bayesian Nonparametrics can be found in the monograph [18] and, in particular for asymptotic studies, [11] provides a review of asymptotics of nonparametric models in terms of "frequentist consistency". Yet another type of asymptotic results are obtained in [9, 31].

The outline of the paper is as follows. In Section 2 one can find a basic introduction to species sampling models and a recollection of some results in the literature concerning the asymptotic behaviour of the number K_n of distinct species in the basic sample, as n increases. Section 3 displays the main results, whereas the last section contains some concluding remarks.

2 Species sampling models and Gibbs–type priors

Let us start by providing a succinct description of completely random measures (CRM) before defining the specific models we will consider and which can be derived as suitable transformations of CRMs. See [25] for an overview of discrete nonparametric models defined in terms of CRMs.

Suppose $\tilde{\mu}$ is a random element defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values on the space $\mathcal{M}_{\mathbb{X}}$ of boundedly finite measures on $(\mathbb{X}, \mathscr{X})$ such that for any $A_1, \ldots, A_n$ in $\mathscr{X}$, with $A_i \cap A_j = \varnothing$ for $i \neq j$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent. Then $\tilde{\mu}$ is termed *completely random measure* (CRM). It is well–known that the Laplace functional transform of $\tilde{\mu}$ has a simple representation of the type

$$\mathbb{E}\left[\mathrm{e}^{-\int f \,\mathrm{d}\tilde{\mu}}\right] = \mathrm{e}^{-\psi(f)}$$

where $\psi(f) = \int_{\mathbb{R}^+ \times \mathbb{X}} [1 - \mathrm{e}^{-sf(y)}] \, \nu(\mathrm{d}s, \mathrm{d}y)$ for any measureable function $f : \mathbb{X} \to \mathbb{R}$ such that $\int |f| \,\mathrm{d}\tilde{\mu} < \infty$ almost surely and the measure ν on $\mathbb{R}^+ \times \mathbb{X}$ is known as the Lévy intensity of $\tilde{\mu}$. See, e.g., [21]. Since a CRM is almost surely discrete, any CRM can be represented as $\tilde{\mu} = \sum_{i \geq 1} J_i \, \delta_{Y_i}$ with independent random jump locations $(Y_i)_{i \geq 1}$ and heights $(J_i)_{i \geq 1}$. For our purposes it is enough to focus on the special case of ν factorizing as $\nu(\mathrm{d}s, \mathrm{d}x) = \rho(s) \,\mathrm{d}s \, \alpha(\mathrm{d}x)$, which implies independence of the locations Y_i's and jumps J_i's in the above series representation. Furthermore, α can be taken to be non–atomic and finite, the latter ensuring almost sure finiteness of the corresponding CRM. Now, if $\mathrm{card}(\{J_i : \, i \geq 1\} \cap (0, \epsilon)) = \int_0^\epsilon \rho(s) \,\mathrm{d}s = \infty$ for

any $\epsilon > 0$, one can define a random probability measure on $\mathbb{X}$ as

$$\tilde{p} = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}. \tag{5}$$

This family of random probability measures is known from [19] as homogeneous normalized random measure with independent increments, a subclass of the general class of normalized processes introduced in [35]. Note that an $\mathbb{X}$–valued exchangeable sequence $(X_n)_{n\geq 1}$ generated by $\tilde{p}$ as in (5) is a species sampling sequence.

Here we focus on a specific example where the CRM defining $\tilde{p}$ in (5) is the so–called *generalized gamma process* [3] that is characterized by

$$\rho(s) = \frac{\sigma}{\Gamma(1-\sigma)}\, s^{-1-\sigma}\, \mathrm{e}^{-\tau s}$$

with $\sigma \in (0,1)$ and $\tau > 0$. In this case

$$\psi(f) = \int_{\mathbb{X}} [(f(x)+\tau)^{\sigma} - \tau^{\sigma}]\, \alpha(\mathrm{d}x) \tag{6}$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\int |f|^{\sigma}\, \mathrm{d}\alpha < \infty$. In the sequel the model will be reparameterized, without loss of generality (see, e.g., [33, 23]), by setting $\beta := \tau^{\sigma}$ and α as a probability measure. The corresponding CRM will be denoted by $\tilde{\mu}_{\sigma,\beta}$. Henceforth, the random probability measure $\tilde{p}$ obtained by normalizing $\tilde{\mu}_{\sigma,\beta}$ as in (5) coincides, in distribution, with the NGG(σ,β) process prior. An important special case arises when $\beta = 0$, since $\tilde{\mu}_{\sigma,0}$ reduces to the σ–stable process, which plays a key role within the paper. For example, it is worth noting that $\tilde{\mu}_{\sigma,\beta}$ can also be defined as an exponential tilting of $\tilde{\mu}_{\sigma,0}$, for any $\beta > 0$. Specifically, if $\mathbb{P}_{\sigma,0}$ is the probability distribution of $\tilde{\mu}_{\sigma,0}$ on $\mathcal{M}_{\mathbb{X}}$ and $\mathbb{P}_{\sigma,\beta}$ is a probability measure on $\mathcal{M}_{\mathbb{X}}$ that is absolutely continuous with respect to $\mathbb{P}_{\sigma,0}$ and such that

$$\frac{\mathrm{d}\mathbb{P}_{\sigma,\beta}}{\mathrm{d}\mathbb{P}_{\sigma,0}}(\mu) = \exp\left\{\beta - \beta^{1/\sigma}\, \mu(\mathbb{X})\right\} \tag{7}$$

then $\mathbb{P}_{\sigma,\beta}$ coincides with the probability distribution of $\tilde{\mu}_{\sigma,\beta}$. In a similar fashion one can also define the PD(σ,θ) process as a polynomial tilting of $\tilde{\mu}_{\sigma,0}$, for any $\theta > -\sigma$. Indeed, one introduces another probability measure $\mathbb{Q}_{\sigma,\theta}$ that is still absolutely continuous with respect to $\mathbb{P}_{\sigma,0}$ and whose Radon–Nykodim derivative is

$$\frac{\mathrm{d}\mathbb{Q}_{\sigma,\theta}}{\mathrm{d}\mathbb{P}_{\sigma,0}}(\mu) = \frac{\Gamma(\theta+1)}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\, [\mu(\mathbb{X})]^{-\theta} \tag{8}$$

for any $\sigma \in (0,1)$ and $\theta > -\sigma$. If $\mu^*_{\sigma,\theta}$ is the random measure with probability distribution $\mathbb{Q}_{\sigma,\theta}$ above, then $p^* = \mu^*_{\sigma,\theta}/\mu^*_{\sigma,\theta}(\mathbb{X})$ coincides, in distribution, with a PD(σ,θ) process. See [34]. The different tilting structure featured by the normalized generalized gamma process and the two

parameter Poisson–Dirichlet process will be reflected by the limiting results to be illustrated in the paper.

It is also worth to recall that both the NGG(σ,β) and the PD(σ,θ) processes can be seen as elements of the general class of *Gibbs–type* nonparametric priors introduced in [13]. Gibbs–type priors represent the most tractable subclass of species sampling models. They are characterized by a parameter $\sigma < 1$ and a collection of non–negative quantities $\{V_{n,k} : \ n \geq 1, \ 1 \leq k \leq n\}$ that satisfy the forward recursive relations

$$V_{n,k} = V_{n+1,k+1} + (n - k\sigma)V_{n+1,k}.$$

These $V_{n,k}$'s define the predictive weights that characterize a species sampling sequence governed by a Gibbs–type prior. Indeed, one has

$$p_{K_n+1,n}(n_1,\ldots,n_{K_n}) = \frac{V_{n+1,K_n+1}}{V_{n,K_n}}, \quad p_{j,n}(n_1,\ldots,n_{K_n}) = \frac{V_{n+1,K_n}}{V_{n,K_n}}(n_j - \sigma) \tag{9}$$

for each $j \in \{1,\ldots,K_n\}$. The fundamental simplification involved in (9) is that the probability of observing a "new" or an "old" species depend on the sample size and on the number of already observed distinct species but not on their frequencies: this crucially simplifies explicit calculations. Turning to the two specific processes introduced before, in accordance with (2), the PD(σ,θ) process identifies a Gibbs–type prior with

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}},$$

whereas, in accordance with (3) and (4), a NGG(σ,β) prior is also of Gibbs type with $\sigma \in (0,1)$ and

$$V_{n,k} = \frac{\mathrm{e}^{\beta}\,\sigma^{k-1}}{\Gamma(n)} \sum_{l=0}^{n-1} \binom{n-1}{l} (-1)^l \,\beta^{l/\sigma}\,\Gamma\left(k - \frac{l}{\sigma};\beta\right).$$

As shown in [27], a normalized CRM is a Gibbs–type prior (with $\sigma \in (0,1)$) if and only if it is a NGG(σ,β) process. This result also motivates the focus of the paper on the NGG(σ,β) process, which clearly has a prominent role.

The result on the limiting behaviour of $K_m^{(n)}$ to be determined in the next section parallels known results for the unconditional case where one aims at determining the asymptotics of K_n as the sample size n increases and connects to the conditional asymptotics displayed in [10] for the PD(σ,θ) process. In order to describe the result for the unconditional case, let $T_{\sigma,0} := \tilde{\mu}_{\sigma,0}(\mathbb{X})$ be the random total mass of a σ–stable CRM and denote by f_σ its density function which satisfies $\int_0^\infty \mathrm{e}^{-\lambda s}\, f_\sigma(s)\,\mathrm{d}s = \mathrm{e}^{-\lambda^\sigma}$ for any $\lambda > 0$. Moreover, let $T_{\sigma,\beta} := \tilde{\mu}_{\sigma,\beta}(\mathbb{X})$ be the random total mass of NGG(σ,β) process and recall that its law can be obtained by exponentially tilting the probability distribution of $T_{\sigma,0}$ as in (7). In particular, if

$$S_{\sigma,\beta} \stackrel{\mathrm{d}}{=} T_{\sigma,\beta}^{-\sigma}, \tag{10}$$

then its density function, with respect to the Lebesgue measure on $\mathbb{R}$, coincides with

$$g_{\sigma,\beta}(s) = \frac{\mathrm{e}^{\beta}}{\sigma}\, \mathrm{e}^{-\left(\frac{\beta}{s}\right)^{\frac{1}{\sigma}}}\, s^{-1-\frac{1}{\sigma}}\, f_{\sigma}(s^{-1/\sigma})\, \mathbb{1}_{(0,\infty)}(s).$$

and one has that

$$\frac{K_n}{n^{\sigma}} \xrightarrow{\text{a.s.}} S_{\sigma,\beta}. \tag{11}$$

According to the terminology introduced by [33], the random variable $S_{\sigma,\beta}$ is the so–called σ–*diversity* of the exchangeable random partition induced by a NGG(σ,β) process prior. See also Definition 3.10 in Pitman [34]. Note that a similar result holds true for the PD(σ,θ) process. Indeed, if $T'_{\sigma,\theta} \overset{\mathrm{d}}{=} \mu^*_{\sigma,\theta}(\mathbb{X})$ so that its probability distribution is obtained by polynomially tilting the probability distribution of $T_{\sigma,0}$ as in (8) and

$$S'_{\sigma,\theta} \overset{\mathrm{d}}{=} (T'_{\sigma,\theta})^{-\sigma} \tag{12}$$

admits density function

$$h_{\sigma,\theta}(s) = \frac{\Gamma(\theta+1)}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\, \frac{s^{\frac{\theta}{\sigma}-\frac{1}{\sigma}-1}}{\sigma}\, f_{\sigma}(s^{-1/\sigma})\, \mathbb{1}_{(0,\infty)}(s).$$

Then one has

$$\frac{K_n}{n^{\sigma}} \xrightarrow{\text{a.s.}} S'_{\sigma,\theta}. \tag{13}$$

See [34, Theorem 3.8]. These results are somehow in line with the fact that the NGG(σ,β) and the PD(σ,θ) processes are distributionally equivalent to normalized random measures that are obtained by an exponential and a polynomial tilting, respectively, of a σ–stable CRM as highlighted in (7) and in (8). Finally note that a combination of [13, Theorem 12] and [33, Proposition 13] shows that the unconditional asymptotic results in (11) and (13) can be extended to the whole class of Gibbs–type priors. See also [17] for another contribution at the interface between Bayesian Nonparametrics and Gibbs–type random partitions.

3 Asymptotics of $K_m^{(n)}$ with a NGG(σ,β) process

As mentioned before, inference on $K_m^{(n)}$ is of great importance since it provides a measure of species richness of a community of plants/animals or of a cDNA library for gene discovery. The key quantity for obtaining posterior inferences is given by the probability distribution $\mathbb{P}[K_m^{(n)} = k \,|\, X_1,\ldots,X_n]$ for $k = 0,\ldots,m$. By virtue of predictive sufficiency of the number K_n of distinct species observed among the first n data $X_1,\ldots,X_n$, in [24] it has been shown that in the NGG(σ,β) this distribution coincides with

$$P_m^{(n,j)}(k) := \mathbb{P}\left[K_m^{(n)} = k \,|\, K_n = j\right] \tag{14}$$

$$= \frac{\mathscr{G}(m,k;\sigma,-n+j\sigma)}{(n)_m} \frac{\sum_{l=0}^{n+m-1}\binom{n+m-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(j+k-\frac{l}{\sigma};\beta\right)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(j-\frac{l}{\sigma};\beta\right)}$$

for $k = 0, \ldots, m$, with $\mathscr{G}(n,k;s,r)$ denoting the non–central generalized factorial coefficient. See [7] for a comprehensive account on generalized factorial coefficients. Expression (14) can be interpreted as the "posterior" probability distribution of the number of distinct new species to be observed in a further sample of size m. Now, based on (14), one obtains the expected number of new species as

$$\hat{E}_m^{(n,j)} := \mathbb{E}[K_m^{(n)}|K_n = j] = \sum_{k=0}^{m} k P_m^{(n,j)}(k), \tag{15}$$

which corresponds to the Bayes estimator of $K_m^{(n)}$ under quadratic loss. Moreover, a measure of uncertainty of the point estimate $\hat{E}_m^{(n,j)}$ can be obtained in terms of α–credible intervals i.e. by determining an interval (z_1, z_2) with $z_1 < z_2$ such that $\mathbb{P}[z_1 \leq K_m^{(n)} \leq z_2 \,|\, K_n = j] \geq \alpha$. The interval (z_1, z_2) of shortest length is then typically referred to as highest posterior density interval.

The main advantage of the distribution (14) is that it is explicit. However, since the sum of incomplete gamma functions cannot be further simplified, its computation can become overwhelming even for moderately large sizes of n and m. This fact represents a major problem in the frequent practical situations in which the size of the additional sample of interest is large. For instance, in genomic applications one has to deal with relevant portions of cDNA libraries which typically consist of millions of genes. Hence, it is natural to study the asymptotics for $K_m^{(n)}$, given K_n, as $m \to +\infty$, in order to obtain approximations of (14) and, consequently, also of (15) and of the corresponding highest posterior density intervals. Indeed, if one is able to show that a suitable rescaling of $K_m^{(n)}$, given K_n, converges in law to some random variable, one can use the probability distribution of this limiting random quantity in order to derive the desired approximations.

3.1 Asymptotic distribution.

The statement of the main result in the paper involves a positive random variable Y_q whose density function is, for any $q > 0$,

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

and we $B_{a,b}$ to denote a beta random variable with parameters (a,b). Moreover, set $S_{n,j} \stackrel{\mathrm{d}}{=} B_{j,n/\sigma-j} Y_{n/\sigma}$, with $B_{j,n/\sigma-j}$ and $Y_{n/\sigma}$ independent, and denote by $g_{S_{n,j}}$ the density function of $S_{n,j}$.

Theorem 3.1 *If $(X_n)_{n\geq 1}$ is a species sampling sequence directed by a NGG(σ,β) process prior, conditional on $K_n = j$ one has*

$$\frac{K_m^{(n)}}{m^\sigma} \to Z_{n,j} \qquad a.s. \tag{16}$$

as $m \to +\infty$, where $Z_{n,j}$ is a positive random variable obtained by exponentially tilting the density function of $S_{n,j}$, namely

$$f_{Z_{n,j}}(z) = \frac{\Gamma(j)e^{-(\beta/z)^{1/\sigma}} g_{S_{n,j}}(z)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(j-\frac{l}{\sigma};\beta\right)}.$$

Proof. The first part of the proof exploits a martingale convergence theorem along the same lines of [34, Theorem 3.8]. In particular, let us start by computing the likelihood ratio

$$M_{\sigma,\beta,m}^{(n)} = \left.\frac{\mathrm{d}\mathbb{P}_{\sigma,\beta}^{(n)}}{\mathrm{d}\mathbb{P}_{\sigma,0}^{(n)}}\right|_{\mathscr{F}_m^{(n)}} = \frac{q_{\sigma,\beta}^{(n)}(K_m^{(n)})}{q_{\sigma,0}^{(n)}(K_m^{(n)})}$$

where $\mathscr{F}_m^{(n)} = \sigma(X_{n+1},\ldots,X_{n+m})$, $\mathbb{P}_{\sigma,\beta}^{(n)}$ is the conditional probability distribution of a normalized generalized gamma process with parameter (σ,β) given K_n and, by virtue of [26, Proposition 1],

$$q_{\sigma,\beta}^{(n)}(K_m^{(n)}) = \frac{\sigma^{K_m^{(n)}}}{(n)_m}\,\frac{\sum_{l=0}^{n+m-1}\binom{n+m-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n+K_m^{(n)}-\frac{l}{\sigma};\beta\right)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n-\frac{l}{\sigma};\beta\right)}$$

for any integer $K_n \geq 1$ and

$$q_{\sigma,0}^{(n)}(K_m^{(n)}) = \frac{\sigma^{K_m^{(n)}}\,(K_n)_{K_m^{(n)}}}{(n)_m}.$$

Hence $(M_{\sigma,\beta,m}^{(n)}, \mathscr{F}_m^{(n)})_{m\geq 1}$ is a $\mathbb{P}_{\sigma,0}^{(n)}$–martingale and by a martingale convergence theorem, $M_{\sigma,\beta,m}^{(n)}$ has a $\mathbb{P}_{\sigma,0}^{(n)}$ almost sure limit, say $M_{\sigma,\beta}^{(n)}$, as $m \to +\infty$. Clearly we have that $\mathbb{E}_{\sigma,0}^{(n)}[M_{\sigma,\beta}^{(n)}] = 1$, where $\mathbb{E}_{\sigma,0}^{(n)}$ denotes the expected value with respect to $\mathbb{P}_{\sigma,0}^{(n)}$. Let now $(E_n)_{n\geq 1}$ be a sequence of i.i.d. random variables having a negative exponential distribution with parameter 1. Moreover, suppose the E_n's are independent of $(K_n, K_m^{(n)})$. Set $\mathscr{E}_m^{(n)} := \sum_{i=1}^{K_n+K_m^{(n)}} E_i$ and note that, conditionally on $(K_n, K_m^{(n)})$, $\mathscr{E}_m^{(n)}$ has gamma distribution with expected value $K_n + K_m^{(n)}$. We can then rewrite $M_{\sigma,\beta,m}^{(n)}$ as follows

$$\begin{aligned} M_{\sigma,\beta,m}^{(n)} &= \frac{\Gamma(K_n)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n-\frac{l}{\sigma};\beta\right)}\,\frac{1}{\Gamma(K_n+K_m^{(n)})} \\ &\qquad \times \sum_{l=0}^{n+m-1}\binom{n+m-1}{l}(-1)^l\beta^{l/\sigma}\int_\beta^{+\infty} y^{K_n+K_m^{(n)}-l/\sigma-1}\mathrm{e}^{-y}\mathrm{d}y \\ &= \frac{\Gamma(K_n)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n-\frac{l}{\sigma};\beta\right)}\,\frac{1}{\Gamma(K_n+K_m^{(n)})} \end{aligned}$$

$$\times \int_{\beta}^{+\infty} y^{K_n+K_m^{(n)}-1}\mathrm{e}^{-y}\left(1-\frac{\beta^{1/\sigma}}{y^{1/\sigma}}\right)^{n+m-1} dy$$

$$= \frac{\Gamma(K_n)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n-\frac{l}{\sigma};\beta\right)}$$

$$\times\ \mathbb{E}\left[\mathbb{1}_{(\beta,+\infty)}(\mathscr{E}_m^{(n)})\left(1-\frac{\beta^{1/\sigma}}{(\mathscr{E}_m^{(n)})^{1/\sigma}}\right)^{n+m+1}\middle|\mathscr{F}_m^{(n)}\right].$$

From the strong law of large numbers, $\mathscr{E}_m^{(n)}/(K_n+K_m^{(n)})\to 1$ as $m\to+\infty$ and conditionally on $(K_n,K_m^{(n)})$. Using the dominated convergence theorem, we have

$$\begin{aligned}
M_{\sigma,\beta,m}^{(n)} &\approx \frac{\Gamma(K_n)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n-\frac{l}{\sigma};\beta\right)} \\
&\qquad\qquad \times\left(1-\frac{\beta^{1/\sigma}}{((K_n+K_m^n)(\mathscr{E}_m^n/(K_n+K_m^n)))^{1/\sigma}}\right)^{n+m-1} \\
&\approx \frac{\Gamma(K_n)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n-\frac{l}{\sigma};\beta\right)}\left(1-\frac{\beta^{1/\sigma}}{(K_n+K_m^{(n)})^{1/\sigma}}\right)^{n+m-1} \\
&\approx \frac{\Gamma(K_n)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n-\frac{l}{\sigma};\beta\right)}\ \exp\left\{-m\frac{\beta^{1/\sigma}}{(K_m^{(n)})^{1/\sigma}}\right\}
\end{aligned}$$

as $m\to+\infty$. Since $M_{\sigma,\beta,m}^{(n)}\to M_{\sigma,\beta}^{(n)}$ almost surely (with respect to $\mathbb{P}_{\sigma,0}^{(n)}$), then there exists some positive random variable, say $L_{\sigma,n}$ such that $m/(K_m^{(n)})^{1/\sigma}\to L_{\sigma,n}$ almost surely (with respect to $\mathbb{P}_{\sigma,0}^{(n)}$). In order to identify the probability distribution of $L_{\sigma,n}$ note that it must be such that

$$\mathbb{E}[\mathrm{e}^{-\beta^{1/\sigma}L_{\sigma,n}}]=\frac{1}{\Gamma(K_n)}\int_{\beta}^{+\infty}y^{K_n-1}\left(1-\frac{\beta^{1/\sigma}}{y^{1/\sigma}}\right)^{n-1}\mathrm{e}^{-y}\,\mathrm{d}y. \tag{17}$$

Since $S_{n,K_n}\overset{\mathrm{d}}{=}B_{K_n,n/\sigma-K_n}Y_{n/\sigma}$, we have to prove that $L_{\sigma,n}\overset{\mathrm{d}}{=}S_{n,K_n}^{-1/\sigma}$, i.e. that the density function of $L_{\sigma,n}$ coincides with

$$\begin{aligned}
f_{L_{\sigma,n}}(z) &= \frac{\sigma\Gamma(n)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma}-K_n\right)}\,z^{-\sigma-1} \\
&\qquad\times\int_{z^{-\sigma}}^{+\infty}\frac{1}{v}v^{\frac{n}{\sigma}-1-\frac{1}{\sigma}}f_\sigma\left(v^{-\frac{1}{\sigma}}\right)\left(\frac{z^{-\sigma}}{v}\right)^{K_n-1}\left(1-\frac{z^{-\sigma}}{v}\right)^{\frac{n}{\sigma}-K_n-1}\mathrm{d}v.
\end{aligned} \tag{18}$$

So we simply have to show that the Laplace transform of the density function in (18) is given by (17). By a simple change of variable, $x=v^{-1/\sigma}$, the previous density reduces to

$$f_{L_{\sigma,n}}(z)=\frac{\sigma^2\Gamma(n)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma}-K_n\right)}\int_0^z x^{-n+\sigma}f_\sigma(x)\left(\frac{z^{-\sigma}}{x^{-\sigma}}\right)^{K_n-1}\left(1-\frac{z^{-\sigma}}{x^{-\sigma}}\right)^{\frac{n}{\sigma}-K_n-1}z^{-\sigma-1}\mathrm{d}x$$

and by the change of variable $y = z^{-\sigma}/x^{-\sigma}$

$$f_{L_{\sigma,n}}(z) = \frac{\sigma\Gamma(n)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma} - K_n\right)} z^{-n} \int_0^1 y^{-n/\sigma+1/\sigma+K_n-1}(1-y)^{n/\sigma-K_n-1} f_\sigma\left(zy^{1/\sigma}\right) \mathrm{d}y.$$

The Laplace transform of $f_{L_{\sigma,n}}$ is then given by

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\beta^{1/\sigma}L_{\sigma,n}}] &= \frac{\sigma\Gamma(n)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma} - K_n\right)} \int_0^\infty \mathrm{e}^{\beta^{1/\sigma}z} z^{-n} \\
&\qquad \times \int_0^1 y^{-n/\sigma+1/\sigma+K_n-1}(1-y)^{n/\sigma-K_n-1} f_\sigma\left(zy^{1/\sigma}\right) \mathrm{d}y\, \mathrm{d}z \\
&= \frac{\sigma\Gamma(n)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma} - K_n\right)} \int_0^1 y^{-n/\sigma+1/\sigma+K_n-1}(1-y)^{n/\sigma-K_n-1} \\
&\qquad \times \int_0^\infty \mathrm{e}^{\beta^{1/\sigma}z} z^{-n} f_\sigma\left(zy^{1/\sigma}\right) \mathrm{d}z\mathrm{d}y \\
&= \frac{\sigma\Gamma(n)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma} - K_n\right)} \int_0^1 y^{K_n-1}(1-y)^{n/\sigma-K_n-1} \\
&\qquad \times \int_0^\infty \mathrm{e}^{(\beta/y)^{1/\sigma}h} h^{-n} f_\sigma(h)\, \mathrm{d}h\, \mathrm{d}y \\
&= \frac{\Gamma\left(\frac{n}{\sigma}\right)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma} - K_n\right)} \int_0^1 y^{K_n-1}(1-y)^{n/\sigma-K_n-1} \\
&\qquad \times \frac{\sigma\Gamma(n)}{\Gamma\left(\frac{n}{\sigma}\right)} \int_0^\infty \mathrm{e}^{(\beta/y)^{1/\sigma}h} h^{-n} f_\sigma(h)\, \mathrm{d}h\, \mathrm{d}y.
\end{aligned}$$

According to the well–known gamma identity, we can write

$$\frac{\sigma\Gamma(n)}{\Gamma\left(\frac{n}{\sigma}\right)} \int_0^\infty \frac{\mathrm{e}^{(\beta/y)^{1/\sigma}h}}{h^n} f_\sigma(h)\, \mathrm{d}h = \frac{\sigma}{\Gamma\left(\frac{n}{\sigma}\right)} \int_0^\infty u^{n-1} \int_0^\infty \mathrm{e}^{-h\left(\frac{\beta^{1/\sigma}}{y^{1/\sigma}}+u\right)} f_\sigma(h)\, \mathrm{d}h\, \mathrm{d}u$$

obtaining

$$\begin{aligned}
&\frac{\Gamma\left(\frac{n}{\sigma}\right)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma} - K_n\right)} \int_0^1 y^{K_n-1}(1-y)^{n/\sigma-K_n-1} \\
&\qquad \times \frac{\sigma}{\Gamma\left(\frac{n}{\sigma}\right)} \int_0^{+\infty} u^{n-1} \int_0^{+\infty} \mathrm{e}^{-h\left(\frac{\beta^{1/\sigma}}{y^{1/\sigma}}+u\right)} f_\sigma(h)\mathrm{d}h\, \mathrm{d}u \\
&= \frac{\Gamma\left(\frac{n}{\sigma}\right)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma} - K_n\right)} \int_0^1 y^{K_n-1}(1-y)^{n/\sigma-K_n-1} \\
&\qquad \times \frac{\sigma}{\Gamma\left(\frac{n}{\sigma}\right)} \int_0^{+\infty} u^{n-1} \mathrm{e}^{-\left(\frac{\beta^{1/\sigma}}{y^{1/\sigma}}+u\right)^\sigma} \mathrm{d}u
\end{aligned}$$

$$= \frac{\Gamma\left(\frac{n}{\sigma}\right)}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma}-K_n\right)} \int_0^1 y^{K_n-1}(1-y)^{n/\sigma-K_n-1} \times \frac{1}{\Gamma\left(\frac{n}{\sigma}\right)} \int_\beta^{+\infty} z^{n/\sigma-1}\left(1-\left(\frac{\beta}{zy}\right)^{1/\sigma}\right)^{n-1} \mathrm{e}^{-z}\mathrm{d}z\,\mathrm{d}y$$

$$= \frac{1}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma}-K_n\right)} \int_\beta^{+\infty} \mathrm{e}^{-z} \times \int_0^z w^{K_n-1}(z-w)^{n/\sigma-K_n-1}\left(1-\left(\frac{\beta}{w}\right)^{1/\sigma}\right)^{n-1} \mathrm{d}w\,\mathrm{d}z$$

$$= \frac{1}{\Gamma(K_n)\Gamma\left(\frac{n}{\sigma}-K_n\right)} \int_\beta^{\infty} w^{K_n-1}\left(1-\left(\frac{\beta}{w}\right)^{1/\sigma}\right)^{n-1} \times \int_w^{+\infty} \mathrm{e}^{-z}(z-w)^{n/\sigma-K_n-1}\,\mathrm{d}z\,\mathrm{d}w$$

which corresponds to (17). Finally, since the probability measures $\mathbb{P}^{(n)}_{\beta,\sigma}$ and $\mathbb{P}^{(n)}_{0,\sigma}$ are mutually absolutely continuous, almost sure convergence holds true with respect to $\mathbb{P}^{(n)}_{\beta,\sigma}$, as well. In order to deduce the $\mathbb{P}^{(n)}_{\beta,\sigma}$-law of Z_{n,K_n}, it is sufficient to exploit a change of measure suggested by

$$\mathbb{P}^{(n)}_{\sigma,\beta}(A) = \int_A \frac{\mathrm{d}\mathbb{P}^{(n)}_{\sigma,\beta}}{\mathrm{d}\mathbb{P}^{(n)}_{\sigma,0}} \mathrm{d}\mathbb{P}^{(n)}_{\sigma,0}$$

and by the fact that

$$\frac{\mathrm{d}\mathbb{P}^{(n)}_{\sigma,\beta}}{\mathrm{d}\mathbb{P}^{(n)}_{\sigma,0}} = M^{(n)}_{\sigma,\beta} = \frac{\Gamma(K_n)}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(K_n-\frac{l}{\sigma};\beta\right)} \mathrm{e}^{-\beta^{1/\sigma}L_{\sigma,n}}.$$

This completes the proof. □

It is worth stressing that the limit random variable in the conditional case is the same as in the unconditional case but with updated parameters and a rescaling induced by a beta-distributed random variable. The density of $Z_{n,j}$ in (16) can be formally represented as

$$f_{Z_{n,j}}(z) = \frac{\Gamma(j)\mathrm{e}^{-(\beta/z)^{1/\sigma}}}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l\beta^{l/\sigma}\Gamma\left(j-\frac{l}{\sigma};\beta\right)} \times \frac{\Gamma(n)}{\Gamma(j)\Gamma\left(\frac{n}{\sigma}-j\right)} z^{j-1}\int_z^{+\infty} v^{-1/\sigma}(v-z)^{n/\sigma-j-1} f_\sigma(v^{-1/\sigma})\mathrm{d}v, \tag{19}$$

where we recall that f_σ is the density function of a positive stable random variable and, then, coincides with a density function of the random total mass of a σ–stable CRM $T_{\sigma,0} := \tilde{\mu}_{\sigma,0}(\mathbb{X})$. Theorem 3.1 can be compared with an analogous result recently obtained in [10, Proposition

2] for the PD (σ,θ) process, where it is shown that the number of new distinct species $K_m^{(n)}$ induced by the PD(σ,θ) process is such that

$$\frac{K_m^{(n)}}{m^\sigma} \xrightarrow{\text{a.s.}} Z'_{n,j} \tag{20}$$

as $m \to +\infty$, where $Z'_{n,j} \stackrel{\text{d}}{=} B_{j+\theta/\sigma,\, n/\sigma-j}\, Y_{(\theta+n)/\sigma}$ and the random variables $B_{j+\theta/\sigma,\, n/\sigma-j}$ and $Y_{(\theta+n)/\sigma}$ are independent. This can be paralleled with the unconditional limit since it is known that $K_n/n^\sigma \to Y_{\theta/\sigma}$, almost surely, as $n \to \infty$. See, e.g., [34, Theorem 3.8].

Remark. Note that a normalized σ–stable process coincides, in distribution, with both a NGG$(\sigma,0)$ and a PD$(\sigma,0)$ process. Hence, it is no surprise that the two limits (16) and (20) are the same, in distribution, when $\beta = \theta = 0$. Another interesting case is represented by the normalized generalized gamma process with parameter $(1/2,\beta)$ which yield to the so–called normalized inverse–gaussian processes ([22]). In particular, for the NGG$(1/2,\beta)$ process the density $f_{1/2}$ in (19) is known explicitly and the previous expression can be simplified to

$$f_{Z_{n,j}}(z) = \frac{\mathrm{e}^{-(\beta/z)^{1/\sigma}}}{\sum_{l=0}^{n-1}\binom{n-1}{l}(-1)^l \beta^{l/\sigma}\Gamma\left(j-\frac{l}{\sigma};\beta\right)} \frac{\Gamma(n) 4^{n-1} z^{j/2-1}}{\pi^{1/2}\,\Gamma(2n-j)} \times \sum_{l=0}^{2n-j-1}\binom{2n-j-1}{l}(-z)^{l/2}\Gamma\left(n-\frac{j-1+l}{2};z\right).$$

3.2 Sampling from the limiting random variable.

Since the above described limiting distributions cannot be easily handled for practical purposes, it is useful to devise a simulation algorithm. In this respect, one can adapt, similarly to [28], an exact sampling algorithm recently devised by [8] for random variate generation from polynomially and exponentially tilted σ–stable distributions. This will allow to sample the limiting random variables $Z'_{n,j}$ and $Z_{n,j}$ corresponding to the PD(σ,θ) and to the NGG(σ,β) case, respectively. Indeed, note that $Z'_{n,j}$ is a scale mixture involving a beta random variable $B_{j+\theta/\sigma,\, n/\sigma-j}$ and a positive random variable $Y_{(\theta+n)/\sigma}$. The latter is such that its transformation $Y_{(\theta+n)/\sigma}^{-1/\sigma}$ admits density function of the form

$$f_{Y_{(\theta+n)/\sigma}^{-1/\sigma}}(y) = \frac{\Gamma(\theta+n+1)}{\Gamma(\frac{\theta+n}{\sigma}+1)} y^{-\theta-n} f_\sigma(y)\, \mathbb{1}_{(0,\infty)}(y), \tag{21}$$

which is precisely the density function of a polynomially tilted σ–stable distribution. Therefore, random variate generation from $Z'_{n,j}$ can be easily done by independently sampling from a beta random variable with parameter $(j+\theta/\sigma, n/\sigma - j)$ and from a random variable with density function (21) by means of the algorithm devised in [8]. We refer to [10] for an alternative

sampling algorithm for $Z'_{n,j}$ via augmentation. Similar arguments can be applied in order to sample from the limit random variable $Z_{n,j}$. Indeed, observe that $Z_{n,j}$ is characterized by a density function proportional to

$$\mathrm{e}^{-(\beta/z)^{1/\sigma}} g_{S_{n,j}}(z)$$

with $g_{S_{n,j}}$ being the density function of the random variable $S_{n,j} \stackrel{\mathrm{d}}{=} B_{j,n/\sigma-j} Y_{n/\sigma}$. Therefore, in order to sample from the distribution of $Z_{n,j}$ one can apply a simple rejection sampling. In particular, the sampling scheme would work as follows

(1) Generate $B \sim B_{j,n/\sigma-j}$.

(2) Sample $Y \sim Y_{n/\sigma}^{-1/\sigma}$ according to Devroye's algorithm.

(3) Set $S = B\, Y^{-\sigma}$

(4) Sample U from a uniform on the interval $(0,1)$

(4.a) If $U \leq \exp\{-(\beta/S)^{1/\sigma}\}$ set $Z = S$

(4.b) If $U > \exp\{-(\beta/S)^{1/\sigma}\}$ restart from (1).

3.3 Interpretation of asymptotic quantities.

In this final section we provide a result that gives an interesting representation of the key random variable $L_{\sigma,n} \stackrel{\mathrm{d}}{=} S_{n,j}^{-1/\sigma}$. To this end, we need to provide a representation for the posterior Laplace transform of the total mass of the σ–stable CRM $\tilde{\mu}_{\sigma,0}$ or, equivalently, of the un–normalized NGG$(\sigma,0)$ or PD$(\sigma,0)$ processes. Indeed one has

Proposition 3.1 *Let $(X_i)_{i\geq 1}$ be a species sampling sequence directed by a normalized σ–stable process prior and suppose that the sample $X_1,\ldots,X_n$ is such that $K_n = j$. Then*

$$\mathbb{E}\left[\mathrm{e}^{-\lambda\,\tilde{\mu}_{\sigma,0}(\mathbb{X})} \,|\, X_1,\ldots,X_n\right] = \frac{1}{\Gamma(j)} \int_{\lambda^\sigma}^{\infty} y^{j-1} \left(1 - \frac{\lambda^{1/\sigma}}{y^{1/\sigma}}\right)^{n-1} \mathrm{e}^{-y}\, \mathrm{d}y \tag{22}$$

for any $\lambda > 0$.

Proof. Set $T_{\sigma,0} \stackrel{\mathrm{d}}{=} \tilde{\mu}_{\sigma,0}(\mathbb{X})$. Since the joint distribution of $(K_n, N_1,\ldots,N_{K_n})$, also known as exchangeable partition probability function (see [34]), of a normalized σ–stable process coincides with $\mathbb{P}[(K_n,N_1,\ldots,N_{K_n}) = (k,n_1,\ldots,n_k)] = \sigma^{j-1}\Gamma(j)\,\prod_{i=1}^k (1-\sigma)_{n_i-1}/\Gamma(n)$, one has

$$\mathbb{E}\left[\mathrm{e}^{-\lambda\,T_{\sigma,0}} \,|\, X_1,\ldots,X_n\right] = \frac{\Gamma(n)}{\sigma^{j-1}\,\Gamma(j)\,\prod_{i=1}^k (1-\sigma)_{n_i-1}}\,\frac{1}{\Gamma(n)}$$

$$\times \int_0^\infty u^{n-1} \mathrm{e}^{-(\lambda+u)^\sigma} \sigma^j \prod_{i=1}^k \frac{\Gamma(n_i-\sigma)}{\Gamma(1-\sigma)} (u+\lambda)^{-n_i+\sigma} \, \mathrm{d}u$$

and a simple change of variable $(u+\lambda)^\sigma = y$ yields the representation in (22). □

Proposition 3.1 allows one to draw an interesting comparison between unconditional and conditional limits of the number of distinct species. As we have already highlighted in Section 2, the probability distribution of the σ–diversities for the NGG(σ, β) process and the PD(σ, θ) process arise as a power transformation (involving the parameter σ) of a suitable tilting of the probability distribution of $T_{\sigma,0} := \tilde{\mu}_{\sigma,0}(\mathbb{X})$. We are now in the position to show that a similar structure carries over when one deals with the conditional case. Resorting to the notation set forth in Theorem 3.1, let $T_{\sigma,0,K_n}$ to be a random variable whose law coincides with the probability distribution of the conditional total mass $T_{\sigma,0}$ of a σ–stable process given a sample of size n containing K_n distinct species. Hence, from the Laplace transform (17) in the proof of Theorem 3.1 one can easily spot the following identity

$$L_{\sigma,n} \stackrel{\mathrm{d}}{=} T_{\sigma,0,K_n}. \tag{23}$$

Let now $\mathbb{P}^{(n)}_{\sigma,0}$ and $\mathbb{P}^{(n)}_{\sigma,\beta}$ be the conditional probability distributions of, respectively, the σ–stable $\tilde{\mu}_{\sigma,0}$ and the generalized gamma $\tilde{\mu}_{\sigma,\beta}$ processes. According to Theorem 3.1 the probability distributions $\mathbb{P}^{(n)}_{\sigma,0}$ and $\mathbb{P}^{(n)}_{\sigma,\beta}$ are mutually absolutely continuous giving rise to the conditional counterpart of the identity (7), i.e.

$$\frac{\mathrm{d}\mathbb{P}^{(n)}_{\sigma,\beta}}{\mathrm{d}\mathbb{P}^{(n)}_{\sigma,0}}(\mu) = \frac{\Gamma(j)}{\sum_{l=0}^{n-1} \binom{n-1}{l} (-1)^l \beta^{l/\sigma} \Gamma\left(j - \frac{l}{\sigma}; \beta\right)} \exp\{-\beta^{1/\sigma} \mu(\mathbb{X})\}, \tag{24}$$

for any $\sigma \in (0,1)$ and $\beta > 0$. In particular, if we denote by T_{σ,β,K_n} the random variable whose probability distribution is obtained by exponentially tilting the probability distribution of $T_{\sigma,0,K_n}$ as in (24), then one can establish that

$$Z_{n,j} \stackrel{\mathrm{d}}{=} (T_{\sigma,\beta,K_n})^{-\sigma}. \tag{25}$$

In other terms, one can easily verify that the probability distribution of the limit random variable $Z_{n,j}$ in (16) can be also derived by applying to the probability distribution of T_{σ,β,K_n} the same transformation characterizing the corresponding unconditional case. In a similar fashion, one can also derive the conditional counterpart of the identity (8) for the two parameter Poisson–Dirichlet process. Indeed, according to [10, Proposition 2], one can introduce a probability measure $\mathbb{Q}^{(n)}_{\sigma,\theta}$ on $\mathcal{M}_{\mathbb{X}}$ whose Radon–Nikodým derivative with respect to the dominating measure $\mathbb{P}^{(n)}_{\sigma,0}$ is given by

$$\frac{\mathrm{d}\mathbb{Q}^{(n)}_{\sigma,\theta}}{\mathrm{d}\mathbb{P}^{(n)}_{\sigma,0}}(\mu) = \frac{\Gamma(\theta+n)\Gamma(j)}{\Gamma(n)\Gamma\left(\frac{\theta}{\sigma}+j\right)} \left[\mu(\mathbb{X})\right]^{-\theta}, \tag{26}$$

for any $\sigma \in (0,1)$ and $\theta > -\sigma$ with $\mathbb{Q}^{(n)}_{\sigma,\theta}$ being the probability measure of the random measure $\mu^*_{\sigma,\theta}$ conditional on the sample. In particular, if we denote by $T^{'}_{\sigma,\theta,K_n}$ the random variable whose probability distribution is obtained by polynomially tilting the probability distribution of $T_{\sigma,0,K_n}$ as in (26), then one can easily verify that

$$Z^{'}_{n,j} \stackrel{\mathrm{d}}{=} (T^{'}_{\sigma,\theta,K_n})^{-\sigma}. \tag{27}$$

This suggests that the probability distribution of the limiting random variable $Z^{'}_{n,j}$ in (20) can also be derived by applying to the probability distribution of $T^{'}_{\sigma,\theta,K_n}$ the same transformation characterizing the corresponding unconditional case.

4 Concluding remarks

The identities (25) and (27) represent the conditional counterparts of the identities (10) and (12), respectively, given a sample containing K_n distinct species. Hence, in the same spirit of [33, Proposition 13], we have provided a characterization of the distribution of the limiting random variables $Z_{n,j}$ and $Z^{'}_{n,j}$ in terms of a power transformation (involving the parameter σ) applied to a suitable tilting for the conditional distribution of the total mass of the σ–stable process. In particular, the identities (25) and (27) characterize the distribution of the limit random variables $Z_{n,j}$ and $Z^{'}_{n,j}$ via the same transformation characterizing the unconditional case and applied to an exponential tilting and polynomial tilting, respectively, for a scale–mixture distribution involving the beta distribution and the σ–stable distribution. To conclude, there is a connection between the prior, and posterior, total mass of a σ–stable CRM that we conjecture can be extended to any Gibbs–type random probability measure and will be object of future research.

References

[1] Argiento, R., Guglielmi, A. and Pievatolo, A. (2009). A comparison of nonparametric priors in hierarchical mixture modelling for AFT regression. *J. Statist. Plann. Inference* **139**, 3989–4005.

[2] Argiento, R., Guglielmi, A. and Pievatolo, A. (2010). Bayesian density estimation and model selection using nonparametric hierarchical mixtures. *Comput. Stat. Dat. An.* **54**, 816–832.

[3] Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.* **31**, 929–953.

[4] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Am. Statist. Assoc.* **88**, 364–73.

[5] Carlton, M.A. (2002). A family of densities derived from the three-parameter Dirichlet process. *J. Appl. Probab.* **39**, 764-774.

[6] Chao, A. (2005). Species estimation and applications. In *Encyclopedia of Statistical Sciences* (N. Balakrishnan, C.B. Read and B. Vidakovic, eds), Vol. 12, 7907-7916. Wiley, New York.

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

[8] Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Trans. Model. Comp. Simul.* **19**, 4.

[9] De Blasi, P., Peccati, G. and Prünster, I. (2009). Asymptotics for posterior hazards. *Ann. Statist.* **37**, 1906–1945.

[10] Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B* **71**, 993–1008.

[11] Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C. Müller, P., Walker, S.G. Eds.), pp. 35–79. Cambridge University Press, Cambridge.

[12] Gnedin, A. (2010). A Species Sampling Model with Finitely Many Types. *Elect. Comm. Probab.* **15**, 79–88.

[13] 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.

[14] Griffin, J., Kolossiatis and M. F. J. Steel (2010a). Modelling overdispersion with the Normalized Tempered Stable distribution. CRiSM Working Paper N. 10-01.

[15] Griffin, J., Kolossiatis and M. F. J. Steel (2010b). Comparing Distributions Using Dependent Normalized Random Measure Mixtures. *CRiSM Working Paper* N. 10-24.

[16] Griffin, J. and Walker, S.G. (2011). Posterior Simulation of Normalized Random Measure Mixtures. *J. Comput. Graph. Stat.*, doi:10.1198/jcgs.2010.08176.

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

[18] Hjort, N.L., Holmes, C.C. Müller, P., Walker, S.G. (Eds.) (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge.

[19] James, L.F., Lijoi, A. and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.* **33**, 105–120.

[20] James, L.F., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.* **36**, 76-97.

[21] Kingman J.F.C. (1993). *Poisson processes*. Oxford University Press, Oxford.

[22] Lijoi, A., Mena, R.H. and Prünster, I. (2005a). Hierarchical mixture modelling with normalized inverse-Gaussian priors. *J. Amer. Stat. Assoc.* **100**, 1278–1291.

[23] Lijoi, A., Mena, R.H. and Prünster, I. (2007a). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 715-740.

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

[25] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C. Müller, P., Walker, S.G. Eds.), pp. 80–136. Cambridge University Press, Cambridge.

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

[27] Lijoi, A., Prünster, I. and Walker, S.G. (2008b). Investigating nonparametric priors with Gibbs structure. *Statist. Sinica* **18**, 1653–1668.

[28] Montagna, S. (2009). Random probability measures and their applications to Bayesian Nonparametrics. *Master thesis*, Collegio Carlo Alberto, Moncalieri, Italy.

[29] Navarrete, C. and Quintana, F. A. (2010). Similarity analysis in Bayesian random partition models. *Comput. Statist. Data Anal*, to appear.

[30] Navarrete, C., Quintana, F.A., and Müller, P. (2008). Some issues on nonparametric Bayesian modeling using species sampling models. *Statist. Modelling* **8**, 3–21.

[31] Peccati, G. and Prünster, I. (2008). Linear and quadratic functionals of random hazard rates: an asymptotic analysis. *Ann. Appl. Probab.* **18**, 1910-1943.

[32] Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (T.S. Ferguson, L.S. Shapley and J.B. MacQueen, Eds.). IMS Lecture Notes Monogr. Ser., Vol. **30**. Inst. Math. Statist., Hayward, 245–267.

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

[34] Pitman, J. (2006). *Combinatorial stochastic processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875, Springer, New York.

[35] Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560–585.