Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 5, *pp.* 993–1008

Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior

Stefano Favaro,

Università degli Studi di Torino and Collegio Carlo Alberto, Turin, Italy

Antonio Lijoi,

Università degli Studi di Pavia, and Istituto di Matematica Applicata e Tecnologie Informatiche–Consiglio Nazionale delle Ricerche, Milan, Italy

Ramsés H. Mena

Universidad Nacional Autónoma de México, Mexico City, Mexico

and Igor Prünster

Università degli Studi di Torino, Collegio Carlo Alberto and International Center for Economic Research, Turin, Italy

[Received July 2008. Revised March 2009]

Summary. A Bayesian non-parametric methodology has been recently proposed to deal with the issue of prediction within species sampling problems. Such problems concern the evaluation, conditional on a sample of size n, of the species variety featured by an additional sample of size m. Genomic applications pose the additional challenge of having to deal with large values of both n and m. In such a case the computation of the Bayesian non-parametric estimators is cumbersome and prevents their implementation. We focus on the two-parameter Poisson–Dirichlet model and provide completely explicit expressions for the corresponding estimators, which can be easily evaluated for any sizes of n and m. We also study the asymptotic behaviour of the number of new species conditionally on the observed sample: such an asymptotic result, combined with a suitable simulation scheme, allows us to derive asymptotic highest posterior density intervals for the estimates of interest. Finally, we illustrate the implementation of the proposed methodology by the analysis of five expressed sequence tags data sets.

Keywords: Asymptotics; Bayesian non-parametrics; Expressed sequence tags analysis; Posterior probability of discovering a new species; Sample coverage; Species sampling; Two-parameter Poisson–Dirichlet process

1. Introduction

Species sampling problems have a long history in ecological and biological studies. Given the information that is yielded by an initial sample of size n, most of the statistical issues to be faced are related to the concept of species richness, which can be quantified in different ways. For example, given an initial sample of size n, species richness might be measured by the estimated

Address for correspondence: Ramsés H. Mena, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, AP 20-726, Admón. 20, 01000 México.
E-mail: ramses@sigma.iimas.unam.mx

 1369–7412/09/71993

number of new species to be observed in an additional sample of size m. It can be alternatively evaluated in terms of the probability of discovering a new species at the $(n+m+1)$th draw, which yields the discovery rate as a function of the size of the additional sample m. Or it can be seen as the sample coverage that is achievable by means of a sample of size $n+m$ which, in other words, is the proportion of distinct species that are detectable in a sample of size $n+m$. Recently there has been renewed interest in the area due to its importance in genomics as witnessed by the recent contributions of, for example, Mao and Lindsay (2002), Mao (2004), Susko and Roger (2004) and Wang and Lindsay (2005). In such inferential problems we are interested in the species composition of a certain population containing an unknown number of species and only a sample drawn from it is available. Specifically, a sample of size n, $X_1, \ldots, X_n$, will exhibit $K_n \in \{1, \ldots, n\}$ distinct species with frequencies $(N_1, \ldots, N_{K_n})$, where clearly $\Sigma_{i=1}^{K_n} N_i = n$. Given such a *basic sample*, interest lies in estimating the number of new species, $K_m^{(n)} := K_m - K_n$, to be observed in an additional sample of size m and in determining the decay of the discovery probability as a function of the sample size m. Genomic applications, such as the analysis of expressed sequence tags (ESTs) that are generated by sequencing complementary deoxyribonucleic acid ('cDNA') libraries consisting of millions of genes, have the distinctive feature of requiring estimation of $K_m^{(n)}$ for very large additional samples.

In recent years there has been an enormous growth in the proposal of Bayesian non-parametric methods for several applied problems. See Müller and Quintana (2004) and Dunson (2008) for interesting reviews, the latter with emphasis on biostatistics applications. As far as species sampling problems are concerned, a Bayesian non-parametric approach has been laid out in Lijoi *et al.* (2007a). Assuming that the data form an exchangeable sequence $(X_n)_{n \geqslant 1}$, by de Finetti's representation theorem $(X_n)_{n \geqslant 1}$ can be characterized by a hierarchical model, with the X_ns as a random sample from some distribution $\tilde{P}$ and a prior Π on $\tilde{P}$, i.e.

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

Their idea then consists in deriving estimators for quantities that are related to the additional sample $X_{n+1}, \ldots, X_{n+m}$ conditionally on the observed basic sample $X_1, \ldots, X_n$. See also Lijoi *et al.* (2008b) for a theoretical study and Lijoi *et al.* (2007b) for a practitioner-oriented illustration. Although the Bayesian non-parametric estimators can be exactly evaluated, there are situations of practical interest, such as the analysis of EST data, where the size of the additional sample of interest is very large and the computational burden makes the evaluation of these estimators almost impossible.

In this paper we focus attention on the two-parameter Poisson–Dirichlet model (Pitman, 1995) which stands out for its tractability and, hence, represents the natural candidate for applications within the large class of priors that was considered in Lijoi *et al.* (2007a). Our primary aim is the achievement of a considerable simplification of the estimators that were proposed in Lijoi *et al.* (2007a), which makes them of straightforward use for any size, no matter how large, of the additional sample. In particular, we obtain an explicit and simple expression for both the expected number of new species and the discovery probability. Moreover, to be able to combine the estimators with measures of uncertainty, we study the asymptotic behaviour of $K_m^{(n)}$: this allows us to deduce asymptotic highest posterior density (HPD) intervals to be associated with the point estimates. The results that we obtain are also of interest beyond the species sampling framework since they shed some light on conditional properties of the two-parameter Poisson–Dirichlet process, which appears in many contexts that are not related to Bayesian non-parametrics such as combinatorics, excursion theory and population genetics. See Pitman (2006) and references therein.

In Section 2 we recall the definition of the two-parameter Poisson–Dirichlet process, derive the moments of any order of $K_m^{(n)}$ conditionally on a basic sample and study its asymptotic behaviour: it will be shown that, given K_n, $K_m^{(n)}/m^\sigma$ converges, as $m \to \infty$, to a random variable. Moreover, we devise a simulation algorithm for drawing samples from this limiting random variable. In Section 3 we show how to implement the results by analysing five real EST data sets. Proofs are postponed to Appendix A.

2. Conditional formulae for species sampling problems

We start this section by introducing the two-parameter Poisson–Dirichlet process (Pitman, 1995). Among the various possible definitions, a simple and intuitive one follows from the so-called stick breaking construction. For a pair of parameters (σ, θ) such that $\sigma \in (0,1)$ and $\theta > -\sigma$, let $(V_k)_{k \geqslant 1}$ denote a sequence of independent random variables, with $V_k \sim \text{beta}(\theta + k\sigma, 1-\sigma)$. Define the stick breaking weights as $\tilde{p}_1 = V_1$ and

$$\tilde{p}_j = V_j \prod_{i=1}^{j-1} (1 - V_i) \qquad j \geqslant 2$$

and suppose that $(Y_n)_{n \geqslant 1}$ is a sequence of independent and identically distributed (IID) random variables, which are independent of the $\tilde{p}_i$s and whose common probability distribution P_0 is non-atomic. If δ_a is the point mass at a, the discrete random-probability measure $\tilde{P}_{\sigma,\theta} = \Sigma_{j \geqslant 1} \tilde{p}_j \delta_{Y_j}$ is a Poisson–Dirichlet process with parameters (σ, θ). For brevity we write $\text{PD}(\sigma, \theta)$. See Pitman (2006) for a detailed account on general theoretical aspects and, for example, Ishwaran and James (2001), Navarrete *et al.* (2008) and Jara *et al.* (2008) for applications in Bayesian non-parametrics.

Under model (1) with $\tilde{P}$ being a $\text{PD}(\sigma, \theta)$ process, the sample coverage, which is defined as the proportion of species represented in a basic sample of size n featuring j distinct species, is given by

$$\hat{C}_1^{(n,j)} = 1 - \frac{\theta + j\sigma}{\theta + n}.$$

Moreover, the distribution of the number of new distinct species $K_m^{(n)}$ that will be observed in an additional sample of size m, conditionally on a basic sample of size n featuring K_n distinct species, is given by

$$P_m^{(n,j)}(k) := \text{pr}(K_m^{(n)} = k \mid K_n = j) = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{\prod_{i=j}^{j+k-1} (\theta + i\sigma)}{\sigma^k} \mathscr{C}(m, k; \sigma, -n + j\sigma) \qquad (2)$$

for $k = 0, \ldots, m$, where $\mathscr{C}(m, k; \sigma, -n + j\sigma)$ is the non-central generalized factorial coefficient whose definition is recalled in equation (16) in Appendix A. Such an expression is the key for evaluating Bayesian estimators that are useful for inference with species sampling problems. In Lijoi *et al.* (2007a) it was deduced, resorting to combinatorial arguments, as a particular case of a general class of priors. In Appendix A we provide an alternative proof of result (2) since it introduces the way of reasoning that we shall resort to for proving proposition 1.

On the basis of result (2), the estimators of interest can be derived: the expected number of new species is

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

whereas the discovery probability, which is interpreted as the probability that the $(n+m+1)$th observation will yield a new species, without observing the m intermediate records, is given by

$$\hat{D}_m^{(n,j)} = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \sum_{k=0}^{m} \frac{\prod_{i=j}^{j+k}(\theta+i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma). \tag{3}$$

Hence, the estimated sample coverage after $n+m$ draws is given by $\hat{C}_m^{(n,j)} = 1 - \hat{D}_m^{(n,j)}$. The advantage of the formulae yielding $\hat{E}_m^{(n,j)}$ and $\hat{D}_m^{(n,j)}$ is that they are explicit and can be exactly evaluated. There are, however, situations of practical interest where the size of the additional sample of interest is very large and the computational burden for evaluating equations (2) and (3) becomes overwhelming. This happens, for instance, in genomic applications where one must deal with relevant portions of cDNA libraries which typically consist of millions of genes. Our first aim is the achievement of a considerable simplification of the two above-mentioned estimators. Moreover, since equation (2) is still required for determining the corresponding HPD intervals, we shall study the asymptotics of $K_m^{(n)}$, given K_n, as $m \to \infty$: this allows us to use the distribution of the limiting random quantity to approximate the HPD intervals.

The first important result concerns the moments of $K_m^{(n)}$, given K_n, which will be expressed in terms of non-central Stirling numbers of the second kind,

$$S(r,i;\gamma) = \frac{1}{i!} \sum_{l=0}^{i} (-1)^{i-l} \binom{i}{l} (l+\gamma)^r \tag{4}$$

for $r = 0, 1, \ldots$ and $i = 0, \ldots, r$, and $S(r,i;\gamma) = 0$ for $i = r+1, r+2, \ldots$. See Charalambides (2005) for an account on non-central Stirling numbers. Such moments allow us to derive completely explicit expressions for the estimators of interest, which can be easily evaluated for any choice of n and m.

Proposition 1. Under the two-parameter Poisson–Dirichlet model, we have

$$E[(K_m^{(n)})^r | K_n = j] = \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(j + \frac{\theta}{\sigma}\right)_{\nu} S\left(r, \nu; \frac{\theta}{\sigma} + j\right) \frac{(\theta+n+\nu\sigma)_m}{(\theta+n)_m} \tag{5}$$

where, for any non-negative integer N, $(a)_N = \Gamma(a+N)/\Gamma(a)$ is the Nth ascending factorial of a. In particular, a Bayesian non-parametric estimator of $K_m^{(n)}$ coincides with

$$E[K_m^{(n)} | K_n = j] = \left(j + \frac{\theta}{\sigma}\right) \left\{ \frac{(\theta+n+\sigma)_m}{(\theta+n)_m} - 1 \right\}, \tag{6}$$

the discovery probability is equal to

$$\hat{D}_m^{(n:j)} = \frac{\theta + j\sigma}{\theta+n} \frac{(\theta+n+\sigma)_m}{(\theta+n+1)_m} \tag{7}$$

and the sample coverage after $n+m$ draws is given by

$$\hat{C}_m^{(n,j)} = 1 - \frac{\theta + j\sigma}{\theta+n} \frac{(\theta+n+\sigma)_m}{(\theta+n+1)_m}. \tag{8}$$

Note that the estimator in equation (6) admits an interesting probabilistic interpretation. Indeed, we have that

$$E[K_m^{(n)} | K_n = j] = \mathrm{pr}(X_{n+1} = \text{new} | K_n = j)\, E_{\sigma,\theta+n}[K_m]$$

where $E_{\sigma,\theta+n}[K_m]$ stands for the unconditional expected number of distinct species, among m observations, with respect to the probability distribution of a Poisson–Dirichlet process with

parameter $(\sigma, \theta+n)$. Moments of any order of the unconditional distribution, i.e. $E[K_n^r]$, have been determined by Pitman (1996a) and Yamato and Sibuya (2000) and are recovered from equation (5) by setting $n=j=0$.

The formulae that were obtained in proposition 1 provide point estimators for quantities of interest in species sampling problems. Besides them, we would also like to determine HPD intervals since they provide a measure of uncertainty related to the point estimates. However, for large values of m this represents a difficult task. To overcome this drawback, we analyse the asymptotic behaviour of $K_m^{(n)}$, for fixed n and as $m \to \infty$, and use the appropriate quantiles of the limiting random variable to obtain an HPD interval. Results of this type for the unconditional distribution have been determined by Pitman (1996a, 1999). See also Pitman (2006). To recall Pitman's result, let f_σ be the density function of a positive σ-stable random variable and Y_q be, for any $q \geqslant 0$, a positive random variable with density function

$$f_{Y_q}(y) = \frac{\Gamma(q\sigma+1)}{\sigma\,\Gamma(q+1)} y^{q-1-1/\sigma} f_\sigma(y^{-1/\sigma}). \tag{9}$$

We, then, have that $K_n/n^\sigma \to Y_{\theta/\sigma}$ almost surely, as $n \to \infty$. As we shall now see, conditioning on the outcome of a basic sample leads to a different limiting result.

Proposition 2. Under the two-parameter Poisson–Dirichlet model, conditional on $K_n = j$ we have

$$K_m^{(n)}/m^\sigma \to Z_{n,j} \qquad \text{almost surely} \tag{10}$$

and in the pth mean, for any $p>0$, where $Z_{n,j} =^{\mathrm{d}} B_{j+\theta/\sigma, n/\sigma-j} Y_{(\theta+n)/\sigma}$, $B_{a,b}$ is a beta random variable with parameters (a,b) and the random variables $B_{j+\theta/\sigma,n/\sigma-j}$ and $Y_{(\theta+n)/\sigma}$ are independent. Moreover,

$$E[(Z_{n,j})^r] = \left(j+\frac{\theta}{\sigma}\right)_r \frac{\Gamma(\theta+n)}{\Gamma(\theta+n+r\sigma)}. \tag{11}$$

It is worth stressing that the limiting random variable in the conditional case is the same as in the unconditional case but with updated parameters and a rescaling that is induced by a beta random variable. The density of $Z_{n,j}$ in expression (10) can be formally represented as

$$f_{Z_{n,j}}(z) = \frac{\Gamma(\theta+n)}{\Gamma(\theta/\sigma+j)\,\Gamma(n/\sigma-j)} z^{\theta/\sigma+j-1} \int_z^\infty v^{-1/\sigma}(v-z)^{n/\sigma-j-1}\, f_\sigma(v^{-1/\sigma})\,\mathrm{d}v.$$

When $\sigma = \frac{1}{2}$, the density $f_{1/2}$ is known explicitly and the previous expression can be simplified to

$$f_{Z_{n,j}}(z) = \frac{4^{n+\theta-1}\Gamma(\theta+n) z^{\theta+k/2-1}}{\pi^{1/2}\Gamma(k+2\theta)\,\Gamma(2n-k)} \sum_{j=0}^{2n-k-1} \binom{2n-k-1}{j} (-z)^{j/2}\, \Gamma\left(n-\frac{k-1+j}{2}; z\right).$$

Nonetheless, even in the latter case we cannot easily determine the quantiles of $Z_{n,j}$ that we need to use to determine HPD intervals. Hence, we resort to a simulation algorithm for generating values of $Z_{n,j}$ and use the output to evaluate quantiles. The demanding part of this simulation is the generation of samples from the probability distribution of Y_q. Note that the sampling strategy that we shall outline is also useful in the unconditional case, where the same tractability issue in deriving properties of Y_q is to be faced. The basic idea consists in setting $W_q = Y_q^{-1/\sigma}$ so that W_q has density function given by

$$f(w) = \frac{\sigma\,\Gamma(q\sigma)}{\Gamma(q)} w^{-q\sigma} f_\sigma(w) = \frac{\sigma}{\Gamma(q)} f_\sigma(w) \int_0^\infty u^{q\sigma-1} \exp(-uw)\,\mathrm{d}u.$$

Via augmentation, we then have

$$f(u,w) = \frac{\sigma}{\Gamma(q)}\, f_\sigma(w) u^{q\sigma-1}\exp(-uw) = f(u)\, f_\sigma(w|u)$$

where $f(u)$ is the density function of a random variable U_q such that $U_q^\sigma \sim \text{gamma}(q,1)$, and

$$f_\sigma(w|u) = f_\sigma(w)\exp(-uw+u^\sigma).$$

This means that, conditional on U_q, W_q is a positive tempered stable random variable, according to the terminology that was adopted in Rosiński (2007). To draw samples from it, a convenient strategy is to resort to the series representation that was derived in Rosiński (2007), which, in our case, yields

$$W_q|U_q \stackrel{\text{d}}{=} \sum_{i=1}^{\infty} \min[\{a_i\,\Gamma(1-\sigma)\}^{-1/\sigma}, e_i v_i^{1/\sigma}] \tag{12}$$

where $e_i \sim^{\text{IID}} \exp(U_q)$, $v_i \sim^{\text{IID}} U(0,1)$ and $a_1 > a_2 > \ldots$ are the arrival times of a Poisson process with unit intensity. Other possibilities for simulating from a tempered stable random variable are the inverse Lévy measure method as described in Ferguson and Klass (1972) and a compound Poisson approximation scheme that was proposed in Cont and Tankov (2004).

Summarizing the above considerations, an algorithm for simulating from the limiting random variable $Z_{n,j} \stackrel{\text{d}}{=} B_{j+\theta/\sigma,n/\sigma-j} Y_{(\theta+n)/\sigma}$ is as follows.

Step 1: generate $B \sim \text{beta}(j+\theta/\sigma, n/\sigma - j)$.
Step 2: To sample from $Y_{(\theta+n)/\sigma}$:

(a) generate $X \sim \text{Ga}\{(\theta+n)/\sigma, 1\}$ and set $U = X^{1/\sigma}$;
(b) for a given truncation N and U sampled in step 2(a), generate: $\{e_i\} \sim^{\text{IID}} \exp(U)$, $\{v_i\} \sim^{\text{IID}} U(0,1)$ and $\xi_j \sim^{\text{IID}} \exp(1)$ and take $a_i = \Sigma_{j=1}^{i} \xi_j$, for $i = 1, \ldots, N$;
(c) compute W according to expression (12) and set $Y = W^{-\sigma}$.

Step 3: take $Z = BY$.

Note that, to establish whether a chosen truncation threshold N for the series in step 2(b) is sufficiently large, one can compare the sample moments with the simple exact moments of $Z_{n,j}$ given in equation (11).

3. Applications to genomics

We now show how to use the results of the previous section by applying them to five real EST data sets. As briefly mentioned in Section 1, EST data arise by sequencing cDNA libraries consisting of millions of genes and one of the main quantities of interest is the number of distinct genes. Typically, owing to cost constraints, only a small portion of the cDNA has been sequenced and, given this basic sample, estimation of the number of new genes $K_m^{(n)}$ to appear in a hypothetical additional sample is required. On the basis of such estimates, geneticists must decide whether it is worth proceeding with sequencing and, if so, also the size of the additional sample. Here, we consider

(a) a tomato flower cDNA library (Quackenbush *et al.*, 2000), which was previously analysed in Mao and Lindsay (2002), Mao (2004) and Lijoi *et al.* (2007a),
(b) two cDNA libraries of the amitochondriate protist *Mastigamoeba balamuthi* (Susko and Roger, 2004) (the first is *non-normalized*, whereas the second is *normalized*, i.e. it undergoes a normalization protocol which aims at making the frequencies of genes in the library more uniform to increase the discovery rate) and

(c) two *Naegleria gruberi* cDNA libraries prepared from cells grown under different culture conditions, aerobic and anaerobic (Susko and Roger, 2004).

To implement the PD(σ, θ) model, the first issue to face is represented by the specification of its parameters. The first possibility is to adopt an empirical Bayes approach. Since the basic sample consists of n observations featuring K_n distinct species with corresponding frequencies $(N_1, \ldots, N_{K_n})$, the joint distribution of K_n and $(N_1, \ldots, N_{K_n})$ is given by

$$\text{pr}(K_n = k, \mathbf{N} = \mathbf{n}) = \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1}. \tag{13}$$

This distribution is known as Pitman's sampling formula (Pitman, 1995) and represents a generalization of Ewens's sampling formula (Ewens, 1972), which is a cornerstone in population genetics. The empirical Bayes rule then suggests that we fix (σ, θ) to maximize expression (13) corresponding to the observed sample $(k, n_1, \ldots, n_k)$, i.e.

$$(\hat{\sigma}, \hat{\theta}) = \arg\max_{(\sigma, \theta)} \left\{ \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1} \right\}. \tag{14}$$

An alternative way of eliciting (σ, θ) is by placing a prior distribution on it. Such an approach is useful when we are interested in testing the compatibility of clustering structures among different populations (Lijoi *et al.*, 2008). However, in terms of estimates there are typically no relevant differences given that the posterior distribution of (σ, θ) is highly concentrated. Hence, to keep the exposition as simple as possible, in what follows we focus on PD(σ, θ) models with empirical Bayes prior specification. The extension to the case of priors on (σ, θ) is straightforward.

The computation of the estimators for the number of new genes (6), for the discovery probability (7) and for the sample coverage (8) is immediate. For each of the five EST data sets, the corresponding estimates for additional samples of size $m \in \{n, 10n, 100n\}$ are reported in Table 1 together with the corresponding values n and j of the basic sample and the empirical Bayes specifications of (σ, θ).

The use of proposition 2 is slightly more delicate. Here, we show it only for the estimator of the number of new genes; for the estimators of the discovery probability and the coverage one can proceed along the same lines. To combine the point estimate for $K_m^{(n)}$ with an asymptotic 95% HPD interval, we can simulate from the limiting random variable $Z_{n,j}$ and determine the 95% HPD interval, (z_1, z_2), for $Z_{n,j}$. Then, given that the normalizing rate function for $K_m^{(n)}$ in proposition 2 is m^σ, we obtain an asymptotic 95% HPD interval for $K_m^{(n)}$ as $(z_1 m^\sigma, z_2 m^\sigma)$. Table 2 reports both the exact and the simulated mean and variance of the limiting random variable $Z_{n,j}$ that is associated with each of the five EST data sets as well as the simulated 95% and 99% HPD intervals. The sampled values are obtained by generating 2000 random variates according to the algorithm that was devised in Section 2 with truncation of the series in expression (12) given by $N = 3 \times 10^7$. In fact, it is important to obtain accurate samples from $Z_{n,j}$: a small bias could heavily affect the asymptotic HPD intervals for $K_m^{(n)}$, $(z_1 m^\sigma, z_2 m^\sigma)$, since a large m^σ would amplify the bias. It should be emphasized that it is sufficient to run the simulation of $Z_{n,j}$ only once to obtain the HPD intervals for any choice of the additional sample size m. Hence, it seems definitely worth pursuing a high precision, which can be easily verified by comparing exact moments in equation (11) with the sampled moments.

Table 1. Analysis of the five EST data sets†

Library	n	j	$\hat{\sigma}$	$\hat{\theta}$	m	$\hat{E}_m^{(n,j)}$	$\hat{D}_m^{(n,j)}$	$\hat{C}_m^{(n,j)}$
Tomato flower	2586	1825	0.612	741.0	n	1281	0.447	0.553
					$10n$	8432	0.240	0.760
					$100n$	40890	0.103	0.897
Mastigamoeba	715	460	0.770	46.0	n	346	0.452	0.548
					$10n$	2634	0.307	0.693
					$100n$	16799	0.185	0.815
Mastigamoeba—normalized	363	248	0.700	57.0	n	180	0.456	0.544
					$10n$	1280	0.278	0.722
					$100n$	7205	0.144	0.856
Naegleria aerobic	959	473	0.670	46.3	n	307	0.290	0.710
					$10n$	2085	0.166	0.834
					$100n$	11031	0.080	0.920
Naegleria anaerobic	969	631	0.660	155.5	n	440	0.412	0.588
					$10n$	2994	0.236	0.764
					$100n$	15673	0.111	0.889

†Size of the basic sample n, number of distinct genes j in the basic sample and empirical Bayes specifications for (σ, θ). Exact estimators for the number of new genes $\hat{E}_m^{(n,j)}$ are rounded to the nearest integer, for the discovery probability $\hat{D}_m^{(n,j)}$ and the coverage $\hat{C}_m^{(n,j)}$ for sizes of the additional sample $m \in \{n, 2n, 3n\}$.

Table 2. Characteristics of the limiting random variable $Z_{n,j}$ for the five cDNA libraries: exact mean $E[Z_{n,j}]$, exact variance $\mathrm{var}(Z_{n,j})$, sample mean $\bar{Z}_{n,j}$, sample variance S^2 and sample 95% and 99% HPD intervals

Library	$E[Z_{n,j}]$	$var(Z_{n,j})$	$\bar{Z}_{n,j}$	S^2	95% *HPD*	99% *HPD*
Tomato flower	21.222	0.098	21.251	0.096	(20.62,21.83)	(20.46,22.02)
Mastigamoeba	3.142	0.011	3.176	0.012	(2.95,3.37)	(2.89,3.44)
Mastigamoeba—normalized	4.804	0.043	4.823	0.044	(4.43,5.24)	(4.28,5.36)
Naegleria aerobic	5.279	0.039	5.304	0.039	(4.93,5.69)	(4.78,5.82)
Naegleria anaerobic	8.400	0.054	8.419	0.054	(7.97,8.88)	(7.80,8.98)

Having the asymptotic 95% HPD intervals for $Z_{n,j}$ at hand, the candidate approximate 95% HPD intervals for $K_m^{(n)}$ are $(z_1 m^\sigma, z_2 m^\sigma)$. As apparent from Table 3, the HPD interval that is constructed through such a procedure is not centred on and, in most cases, does not even include the estimated number of new genes $E[K_m^{(n)} | K_n = j]$. Indeed, if we look at the exact estimator for $K_m^{(n)}$ that is given in equation (6), it is clearly much smaller than its asymptotic approximation $m^\sigma E[Z_{n,j}]$. This is because, when θ and n are moderately large and not overwhelmingly smaller than m, a finer normalization constant is to be used for approximating $K_m^{(n)}$: by close inspection of the derivation of the moments of the limiting random variable $Z_{n,j}$ in expression (17) in Appendix A, we see that an equivalent, though less rough, normalization rate is given by

$$r_{\sigma,\theta,n}(m) := (\theta + n + m)^\sigma - (\theta + n)^\sigma.$$

Obviously, in terms of asymptotics, $r_{\sigma,\theta,n}(m)/m^\sigma \to 1$ as $m \to \infty$, but, importantly, as far as approximations of $K_m^{(n)}$ for finite m are concerned, it overcomes the above-mentioned problems. In fact, we have that, for any m, $E[K_m^{(n)} | K_n = j] \approx r_{\sigma,\theta,n}(m)\, E[Z_{n,j}]$ and the asymptotic HPD interval $(r_{\sigma,\theta,n}(m) z_1, r_{\sigma,\theta,n}(m) z_2)$ is approximately centred on the estimator $E[K_m^{(n)} | K_n = j]$, as desired. Table 3 displays, for the five data sets, the exact estimator for $K_m^{(n)}$, its asymptotic

Table 3. Exact estimates $\hat{E}_m^{(n,j)}$ of the number of new genes $K_m^{(n)}$ and its asymptotic approximation $f(m)\,E[Z_{n,j}]$, with rate functions $f(m) = m^\sigma$ and $f(m) = r_{\sigma,\theta,n}$†

Library	m	$\hat{E}_m^{(n,j)}$	*Results for rate* m^σ		*Results for rate* $r_{\sigma,\theta,n}(m)$	
			$m^\sigma\,E[Z_{n,j}]$	*Asymptotic 95% HPD*	$r_{\sigma,\theta,n}\,E[Z_{n,j}]$	*Asymptotic 95% HPD*
Tomato flower, $n = 2586$	n	1281	2602	(2528,2677)	1281	(1244,1318)
	$10n$	8432	10649	(10347,10956)	8432	(8192,8675)
	$100n$	40890	43583	(42345,44838)	40890	(39728,42067)
Mastigamoeba, $n = 715$	n	346	495	(465,531)	346	(325,371)
	$10n$	2634	2917	(2739,3129)	2634	(2473,2825)
	$100n$	16799	17179	(16130,18427)	16799	(15774,18020)
Mastigamoeba—normalized, $n = 363$	n	180	298	(274,324)	180	(166,196)
	$10n$	1280	1491	(1375,1625)	1280	(1181,1396)
	$100n$	7205	7474	(6893,8146)	7205	(6644,7852)
Naegleria aerobic, $n = 959$	n	307	525	(491,566)	307	(287,331)
	$10n$	2085	2457	(2295,2648)	2085	(1947,2247)
	$100n$	11031	11492	(10735,12387)	11031	(10304,11889)
Naegleria anaerobic, $n = 969$	n	440	786	(745,831)	440	(417,465)
	$10n$	2994	3591	(3407,3797)	2994	(2841,3166)
	$100n$	15673	16414	(15572,17355)	15672	(14869,16571)

†The size m of the additional sample varies in $\{n, 10n, 100n\}$. The asymptotic 95% HPD intervals are evaluated for both rate functions, m^σ and $r_{\sigma,\theta,n}(m)$. All values are rounded to the nearest integer.

approximation and the 95% asymptotic HPD intervals using both m^σ and $r_{\sigma,\theta,n}(m)$ as rate functions for sizes of the additional sample $m \in \{n, 10n, 100n\}$.

For the tomato flower library we have that, even for $m = 100n = 258\,600$, the asymptotic approximation of the number of new genes with m^σ is about 6.6% larger than the asymptotic approximation with $r_{\sigma,\theta,n}(m)$, which coincides with the exactly estimated number. Hence, for the finite sample size approximation it is definitely necessary to use $r_{\sigma,\theta,n}(m)$ as the rate function.

We now move on to comparing the asymptotic HPD intervals that are obtained with the rate function $r_{\sigma,\theta,n}(m)$ with the exact HPD intervals that are determined by using the probability distribution in expression (15) in Appendix A. Hence, we consider $m \in \{n, 2n, 3n\}$, because otherwise the computational burden that is involved in expression (15) would become too heavy. Table 4 reports, for the five data sets, the exact estimator for $K_m^{(n)}$, the exact 95% HPD and both the 95% and the 99% asymptotic HPD intervals. Table 4 shows that the length of the asymptotic 95% HPD intervals is shorter than the exact interval, although the difference is not big.

Indeed, such a finding is not surprising in the species sampling context. Obviously, the variability of $K_m^{(n)}$ increases as m increases. However, the variability of $K_m^{(n)}/r_{\sigma,\theta,n}(m)$, which can be interpreted as an average variability over the additional sample of size m, is necessarily decreasing as m increases, since the more distinct species are collected the lower the probability of detecting additional new species will become. Hence, if we approximate $K_m^{(n)}/r_{\sigma,\theta,n}(m)$ by its asymptotic random variable $Z_{n,j}$, we shall necessarily underestimate its variability, which is reflected in the length of the HPD intervals. Nonetheless the possibility of resorting to the asymptotic HPD intervals is extremely useful from a practical point of view:

(a) the HPD intervals of $Z_{n,j}$ automatically yield HPD intervals of $K_m^{(n)}$ for any choice of m, whereas the exact HPD intervals must be recomputed for any m of interest and cannot even be calculated for large m;

Table 4. Estimates $\hat{E}_m^{(n,j)}$ of the number of new genes $K_m^{(n)}$ together with the exact 95% HPD intervals and the 95% and 99% asymptotic HPD intervals†

Library	m	$\hat{E}_m^{(n,j)}$	*Exact 95% HPD*	*Asymptotic 95% HPD*	*Asymptotic 99% HPD*
Tomato flower, $n=2586$	n	1281	(1221,1341)	(1244,1318)	(1234,1329)
	$2n$	2354	(2263,2449)	(2287,2422)	(2269,2442)
	$3n$	3305	(3181,3434)	(3211,3400)	(3186,3430)
Mastigamoeba, $n=715$	n	346	(312,382)	(325,371)	(318,379)
	$2n$	654	(599,711)	(614,701)	(601,716)
	$3n$	939	(865,1015)	(881,1007)	(863,1028)
Mastigamoeba—normalized, $n=363$	n	180	(156,206)	(166,196)	(160,201)
	$2n$	336	(299,375)	(310,366)	(299,375)
	$3n$	477	(428,528)	(440,520)	(425,533)
Naegleria aerobic, $n=959$	n	307	(271,345)	(287,331)	(278,338)
	$2n$	566	(510,624)	(529,610)	(513,624)
	$3n$	798	(725,873)	(746,861)	(723,880)
Naegleria anaerobic, $n=969$	n	440	(402,478)	(417,465)	(408,470)
	$2n$	812	(753,873)	(771,859)	(755,869)
	$3n$	1146	(1069,1225)	(1088,1212)	(1065,1226)

†All values are rounded to the nearest integer.

(b) the fact that the length of the asymptotic HPD intervals is always shorter than the exact length (and not oscillating) allows us to interpret it as a 'lower bound' on the length of the exact intervals and, moreover, the underestimation will decrease as m increases.

Given such a lower bound, it would be also of interest to have an 'upper bound' on the length of the exact HPD interval. Indeed, if we consider the asymptotic 99% HPD intervals, by proposition 2, there is an $\bar{m}$ such that for any $m > \bar{m}$ the asymptotic 99% HPD interval for $K_m^{(n)}$ covers the exact 95% HPD interval. Hence, for sufficiently large m, the asymptotic 99% HPD interval acts as an upper bound for the exact interval. Although the determination of such a suitable m, for any choice of parameters and basic samples, is not possible we can proceed empirically. From Table 4, where the 99% asymptotic HPD intervals are reported as well, we see that for the *Mastigamoeba* and *Naegleria* libraries the asymptotic 99% HPD interval covers the exact 95% HPD interval already starting from $m = 3n$. As for the tomato flower library, whose distinctive feature with respect to the other libraries is represented by a larger basic sample, such a covering has not yet been achieved for $m = 3n$ but it is very close to happen. Hence, by the combination of the asymptotic 95% and 99% HPD intervals, we obtain a useful device for assessing uncertainty of species richness estimates. Fig. 1 shows, for the *Naegleria* anaerobic cDNA library, how the 95% and 99% asymptotic HPD intervals provide an envelope around the exact HPD interval from $m \approx 2500$ onwards. Given that the two asymptotic HPD intervals are quite close, we thus achieve a satisfactorily accurate estimate of the uncertainty.

Finally, we perform a cross-validation study in terms of out-of-sample predictive performance and at the same time we compare the behaviour of the Poisson–Dirichlet process estimator with other widely used estimators. Specifically, we consider the tomato flower library, which, among the data sets considered, has the largest observed sample ($n = 2586$), thus allowing an effective cross-validation study. We take subsamples of size $n = 1034$ and make predictions over an additional sample of size $m = 1552$. This amounts to predictions for an additional sample 1.5 times the basic sample, which allows us to compare the results also with estimators which become unstable for larger sizes of the additional sample such as the popular estimator of Efron

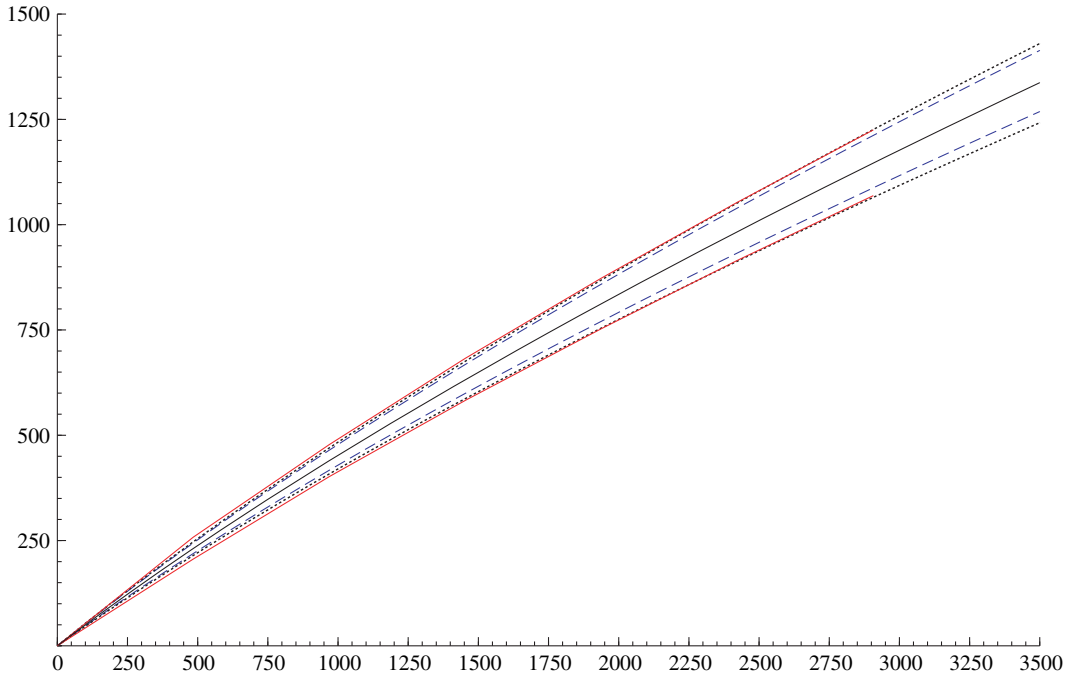

Fig. 1. Exact estimator $\hat{E}_m^{(n,j)}$ (———) and corresponding exact 95% HPD intervals (———), and asymptotic 95% HPD intervals (– – – –) and asymptotic 99% HPD intervals (·····) for the *Naegleria* anaerobic library

and Thisted (1976). The subsamples are obtained by sampling 1034 genes without replacement from the 2586 observed genes and by recording K_n and the frequencies of the observed genes. The true value for the number of new genes in the additional sample, $K_m^{(n)}$, is then equal to $1825 - K_n$, since 1825 are the distinct genes in the observed sample of size 2586. Predictions of $K_m^{(n)}$ are derived using, in addition to the Poisson–Dirichlet process estimator, the following estimators:

(a) the estimator of Efron and Thisted (1976) which is based on a gamma–mixed Poisson model;
(b) the plug-in estimator of Solow and Polasky (1999);
(c) the non-parametric estimator of Chao and Shen (2004);
(d) the penalized non-parametric maximum likelihood estimator of Wang and Lindsay (2005).

Estimators (a)–(c) are computed by using the SPADE software that is available from `http://chao.stat.nthu.edu.tw`, whereas estimator (d) is calculated by using the EST-stat Java program that is available at `http://bioinfo.stats.northwestern.edu/~jzwang`. To make the comparison on representative samples, we generated 10000 subsamples of size 1034 from the whole sample of 2586 units and recorded the frequency distribution of the number of distinct genes K_n within each subsample: the corresponding empirical deciles are 839, 844, 847, 850, 852, 855, 858, 861 and 865. Samples with number of distinct genes belonging to the low and high deciles correspond to situations respectively of underrepresentation and over-representation of distinct genes with respect to the distinct genes in the whole sample. Table 5 displays the results for 10 samples, where each sample corresponds to a different decile. The Chao–Shen estimator, which allows us to tune a cut-off point (see Chao and Shen (2004)), is

Table 5. Cross-validation study with basic sample of size $n = 1034$ and prediction for an additional sample of size $m = 1.5n$ based on the tomato flower library data (2586 genes with 1825 distinct genes)†

Estimator	*Results for the following samples:*				
	1	*2*	*3*	*4*	*5*
K_n	837	842	845	849	851
True $K_m^{(n)}$	*988*	*983*	*980*	*976*	*974*
PD(σ, θ)	952	982	975	972	991
	(904,999)	(934,1031)	(928,1022)	(925,1019)	(944,1039)
Efron–Thisted	670	1000	900	790	800
	(360,980)	(760,1300)	(570,1200)	(510,1100)	(510,1100)
Solow–Polasky	899	926	907	928	940
	(818,980)	(844,1007)	(826,988)	(848,1008)	(860,1021)
Chao–Shen	952	977	968	968	987
	(872,1033)	(897,1056)	(886,1049)	(885,1051)	(905,1069)
Wang–Lindsay	909	927	918	933	948
	(834,955)	(857,983)	(842,966)	(849,983)	(881,1004)
	6	*7*	*8*	*9*	*10*
K_n	853	856	859	862	865
True $K_m^{(n)}$	*972*	*969*	*966*	*963*	*960*
PD(σ, θ)	981	991	990	1003	1013
	(934,1027)	(944,1037)	(944,1036)	(957,1049)	(967,1059)
Efron–Thisted	650	640	1000	880	850
	(280,1000)	(58,1200)	(890,1200)	(690,1100)	(670,1000)
Solow–Polasky	927	941	939	956	950
	(846,1007)	(861,1021)	(859,1018)	(876,1036)	(869,1030)
Chao–Shen	984	989	981	1000	1010
	(900,1068)	(901,1076)	(899,1062)	(916,1082)	(929,1092)
Wang–Lindsay	933	950	939	964	960
	(867,1002)	(886,1013)	(866,1010)	(881,1026)	(893,1028)

†K_n reports the observed distinct genes in the subsamples; the true $K_m^{(n)}$ (values in italics) is then given by $1825 - K_n$. Point and 95% uncertainty estimates are displayed for the Poisson–Dirichlet estimator, the Efron–Thisted estimator, the Solow–Polasky estimator, the Chao–Shen estimator and the Wang–Lindsay estimator. All values are rounded to the nearest integer.

reported with cut-off point equal to the gene(s) with highest frequency. Lower cut-off points worsen the resulting estimates.

Table 5 shows that the Poisson–Dirichlet and Chao–Shen estimators exhibit the overall best performances, whereas the Efron–Thisted and Solow–Polasky estimators are less accurate. The Wang–Lindsay estimator performs very well for samples with large K_n but underestimates $K_m^{(n)}$ significantly in the other cases. Compared with the other estimators the Poisson–Dirichlet estimator exhibits narrower uncertainty estimates: their average length is 92 genes, whereas for the Chao–Shen estimator it is 165. In cases where the point estimate is accurate this represents an advantage but when this is not so it may lead to missing the correct value as happens for sample 10. If one prefers larger HPDs with the Poisson–Dirichlet model, then it is advisable to put priors on (σ, θ). For instance, for sample 10 with independent uniform priors on σ and θ, the estimate for $K_m^{(n)}$ is 1016 with HPD (949,1087): the point estimate is essentially the same but the larger HPD allows us to capture the true value. It also worth noting that the extreme situation with

underrepresentation or overrepresentation of distinct genes in the basic sample seem to be less likely in real EST sequencing than in sampling without replacing, since in EST sequencing there is a constant sequencing error rate which prevents such abrupt changes in the discovery rate. A repeated analysis, which is not reported here, for various samples belonging to the different deciles shows essentially the same behaviour for the various estimators and, hence, confirms the patterns that were nicely highlighted by the grouping according to K_n presented in Table 5.

4. Concluding remarks

In this paper we have derived results which allow the implementation of the two-parameter Poisson–Dirichlet model in species sampling problems for any sizes of the basic and the additional sample. This is of particular importance in genomics problems, where prediction over large unobserved portions of cDNA libraries is required. Specifically, the derived estimators for the number of new genes, the discovery rate and the sample coverage are completely explicit. Moreover, the conditional asymptotic result concerning the number of new species yields also measures of uncertainty of the estimates in the form of asymptotic HPD intervals, which can be readily used as approximate HPD intervals. Given that the 95% asymptotic HPD interval is always included in the 95% exact HPD interval and that, for sufficiently large m, the 99% asymptotic HPD covers the exact 95% HPD interval, the combination of the 95% and 99% asymptotic HPD intervals provides a simple and valuable measure of uncertainty.

Acknowledgements

The authors are grateful to the Joint Editor, an Associate Editor and two referees for their constructive comments and suggestions. Special thanks are due to Ole Winther for some useful discussions. Moreover, the hospitality of the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, where this project started during the Isaac Newton Institute programme 'Bayesian nonparametric regression', is acknowledged. Antonio Lijoi and Igor Prünster are partially supported by the Italian Ministry of University and Research project 'Bayesian methods: theoretical developments and novel applications'. Ramsés H. Mena was partially supported by Consejo Nacional de Ciencia y Tecnología grant J50160-F.

Appendix A

A.1. Alternative derivation of the distribution in expression (2)

An important result that was proved in Pitman (1996b) concerns the representation of the posterior distribution of $\tilde{P}_{\sigma,\theta}$, given a sample $X_1, \ldots, X_n$ of data governed by $\tilde{P}_{\sigma,\theta}$. Indeed, if the observations X_i are, conditional on $\tilde{P}_{\sigma,\theta}$, IID from $\tilde{P}_{\sigma,\theta}$ and the sample $X_1, \ldots, X_n$ contains $j \leqslant n$ distinct values $X_1^*, \ldots, X_j^*$, then

$$\tilde{P}_{\sigma,\theta}|(X_1, \ldots, X_n) \stackrel{\mathrm{d}}{=} \sum_{i=1}^{j} w_i \delta_{X_i^*} + w_{j+1} \tilde{P}_{\sigma,\theta+j\sigma} \tag{15}$$

where $(w_1, \ldots, w_j)$ is distributed according to a j-variate Dirichlet distribution with parameters $(n_1 - \sigma, \ldots, n_j - \sigma, \theta + j\sigma)$, $n_i = \mathrm{card}\{r : X_r = X_i^*\}$ is the frequency of X_i^* in the sample and $w_{j+1} = 1 - \sum_{i=1}^{j} w_i$.

To derive expression (2), we shall make use of the posterior representation that is given in expression (15) and of the distributional properties of K_n. Indeed, from expression (15) we note that, given $w \sim \mathrm{beta}(\theta + j\sigma, n - j\sigma)$, an observation X_{n+i}, with $i = 1, \ldots, m$, does not coincide with any of the $K_n = j$ distinct species that are observed in the basic sample with probability w. Consequently

$$\mathrm{pr}(K_m^{(n)} = k | K_n = j) = \frac{\Gamma(\theta + n)}{\Gamma(\theta + j\sigma)\,\Gamma(n - j\sigma)} \int_0^1 \mathrm{pr}(K_m^{(n)} = k | K_n = j, w) w^{\theta + j\sigma - 1} (1 - w)^{n - j\sigma - 1} \mathrm{d}w.$$

To have $K_m^{(n)} = k$, at least k of the m data $X_{n+1}, \ldots, X_{n+m}$ must be allocated to the k new distinct species that are not observed among the $K_n = j$ species of the basic sample. Hence we have

$$\text{pr}(K_m^{(n)} = k | K_n = j, w) = \sum_{i=k}^{m} \binom{m}{i} w^i (1-w)^{m-i} \, \text{pr}(K_i = k)$$

where it is to be noted that K_i is, now, the number of distinct species among the i observations that are generated by a PD$(\sigma, \theta + j\sigma)$ process. Such a probability distribution was derived in Pitman (1999) (see also Pitman (2006)) and in this case yields

$$\text{pr}(K_i = k) = \frac{\prod_{l=1}^{k-1} (\theta + j\sigma + l\sigma)}{\sigma^k (\theta + j\sigma + 1)_{i-1}} \mathscr{C}(i, k; \sigma) \qquad i = k, \ldots, m$$

with

$$\mathscr{C}(i, k; \sigma) = \frac{1}{k!} \sum_{r=0}^{k} (-1)^r \binom{k}{r} (-r\sigma)_i$$

being the generalized factorial coefficient. Summing up the previous considerations we obtain expression (2) by noting that

$$P_m^{(n,j)}(k) = \frac{(\theta/\sigma + j)_k}{(\theta + n)_m} \sum_{i=k}^{m} \binom{m}{i} \mathscr{C}(i, k; \sigma)(n - j\sigma)_i = \frac{(\theta/\sigma + j)_k}{(\theta + n)_m} \mathscr{C}(m, k; \sigma, -n + j\sigma)$$

where the second equality follows from expression (2.56) in Charalambides (2005) and

$$\mathscr{C}(m, k; \sigma, -n + j\sigma) = \frac{1}{k!} \sum_{r=0}^{k} (-1)^r \binom{k}{r} (n - \sigma(r + j))_m \tag{16}$$

is the non-central generalized factorial coefficient. See Charalambides (2005) for a detailed account on generalized factorial coefficients.

A.2. Proof of proposition 1

Indeed, we have

$$E[(K_m^{(n)})^r | K_n = j, w] = \sum_{i=0}^{m} \binom{m}{i} w^i (1-w)^{m-i} E[K_i^r]$$

where the unconditional moment $E[K_i^r]$ is evaluated with respect to a $\tilde{P}_{\sigma,\theta+j\sigma}$-prior. Such an expression is already available from Pitman (1996a) and Yamato and Sibuya (2000) and it is given by

$$E[K_i^r] = \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(1 + \frac{\theta + j\sigma}{\sigma}\right)_\nu S\left(r, \nu; \frac{\theta + j\sigma}{\sigma}\right) \frac{(\theta + j\sigma + \nu\sigma + 1)_{i-1}}{(\theta + 1)_{i-1}}$$

where S is the non-central Stirling number of the second kind. Hence, we have

$$\begin{aligned}
E[(K_m^{(n)})^r | K_n = j] &= \frac{\Gamma(\theta + n)}{\Gamma(\theta + j\sigma)\,\Gamma(n - j\sigma)} \int_0^1 w^{\theta + j\sigma - 1} (1-w)^{n - j\sigma - 1} E[(K_m^{(n)})^r | K_n = j, w] \, \mathrm{d}w \\
&= \frac{\Gamma(\theta + n)}{\Gamma(\theta + j\sigma)\,\Gamma(n - j\sigma)} \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(1 + \frac{\theta + j\sigma}{\sigma}\right)_\nu S\left(r, \nu; \frac{\theta + j\sigma}{\sigma}\right) \\
&\quad \times \sum_{i=0}^{m} \binom{m}{i} \frac{(\theta + j\sigma + \nu\sigma + 1)_{i-1}}{(\theta + 1)_{i-1}} \int_0^1 w^{\theta + j\sigma + i - 1} (1-w)^{n - j\sigma + m - i - 1} \, \mathrm{d}w \\
&= \frac{1}{(\theta + n)_m} \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(1 + \frac{\theta + j\sigma}{\sigma}\right)_\nu S\left(r, \nu; \frac{\theta + j\sigma}{\sigma}\right) \frac{\theta + j\sigma}{\theta + j\sigma + \nu\sigma} \\
&\quad \times \sum_{i=0}^{m} \binom{m}{i} (\theta + j\sigma + \nu\sigma)_i (n - j\sigma)_{m-i} \\
&= \frac{1}{(\theta + n)_m} \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(\frac{\theta}{\sigma} + j\right)_\nu S\left(r, \nu; \frac{\theta + j\sigma}{\sigma}\right) (\theta + n + \nu\sigma)_m,
\end{aligned}$$

where the last equality follows by an application of the Chu–Vandermonde formula. See, for example, Charalambides (2005).

The expression for the discovery probability in equation (7) is obtained by inserting equation (6) into equation (9) of Lijoi *et al.* (2007b) and some simple algebra.

A.3. Proof of proposition 2

The proof strategy for proposition 2 is as follows: we first adopt a technique that is similar to that suggested in Pitman (2006), theorem 3.8, for the unconditional case to establish that $K_m^{(n)}/m^\sigma$ converges almost surely and in the pth mean for any $p>0$. Then, we determine the moments of the limiting random variable and show that the limiting random variable is characterized by its moments.

Let us start by computing the likelihood ratio

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

where $\mathscr{F}_m^{(n)} = \sigma(X_{n+1},\ldots,X_{n+m})$, $P_{\sigma,\theta}^{(n)}$ is the conditional probability distribution of a PD(σ,θ) process given K_n and, by virtue of proposition 1 in Lijoi *et al.* (2008b),

$$q_{\sigma,\theta}^{(n)}(k) = \frac{\sigma^{K_n}(\theta/\sigma+K_n)_k}{(\theta+n)_m}$$

for any integer $k\geqslant 1$ and $q_{\sigma,\theta}^{(n)}(0) := 1/(\theta+n)_m$. Hence $(M_{\sigma,\theta,m}^{(n)},\mathscr{F}_m^{(n)})_{m\geqslant 1}$ is a $P_{\sigma,0}^{(n)}$-martingale. By a martingale convergence theorem, $M_{\sigma,\theta,m}^{(n)}$ has a $P_{\sigma,0}^{(n)}$ almost sure limit, say $M_{\sigma,\theta}^{(n)}$, as $m\to\infty$. Convergence holds in the pth mean as well, for any $p>0$. We clearly have that $E_{\sigma,0}^{(n)}[M_{\sigma,\theta}^{(n)}]=1$, where $E_{\sigma,0}^{(n)}$ denotes the expected value with respect to $P_{\sigma,0}^{(n)}$. It can be easily seen that

$$M_{\sigma,\theta,m}^{(n)} \sim \frac{\Gamma(\theta+n)\,\Gamma(K_n)}{\Gamma(n)\,\Gamma(\theta/\sigma+K_n)}\left(\frac{K_m^{(n)}}{m^\sigma}\right)^{\theta/\sigma}$$

as $m\to\infty$. Hence $(K_m^{(n)}/m^\sigma)^{\theta/\sigma}$ converges $P_{\sigma,0}^{(n)}$ almost surely to a random variable, say $Z_{n,j}$, such that

$$E_{\sigma,0}^{(n)}[Z_{n,j}^{\theta/\sigma}] = \frac{\Gamma(n)\,\Gamma(\theta/\sigma+K_n)}{\Gamma(\theta+n)\,\Gamma(K_n)}.$$

To identify the distribution of the limiting random variable $Z_{n,j}$ with respect to $P_{\sigma,\theta}^{(n)}$, we consider the asymptotic behaviour of $E[(K_m^{(n)})^r|K_n]$ as $m\to\infty$, for any $r\geqslant 1$. Letting $m\to\infty$ in equation (5) of proposition 1, use the Stirling formula to obtain

$$\frac{1}{m^{r\sigma}}E[(K_m^{(n)})^r|K_n] \to \left(K_n+\frac{\theta}{\sigma}\right)_r \frac{\Gamma(\theta+n)}{\Gamma(\theta+n+r\sigma)} =: \mu_r^{(n)}. \tag{17}$$

Such a moment sequence clearly arises by taking $Z_{n,j} =^{\mathrm{d}} B_{j+\theta/\sigma,n/\sigma-j}Y_{(\theta+n)/\sigma}$, with the beta random variable $B_{j+\theta/\sigma,n/\sigma-j}$ independent from $Y_{(\theta+n)/\sigma}$, which has density (9). Hence, we are left with showing that the distribution of $Z_{n,j}$ is uniquely characterized by the moment sequence $\{\mu_r^{(n)}\}_r$. To establish this, we can evaluate the characteristic function of $Z_{n,j}$ which, at any $t\in\mathbb{R}$, coincides with

$$\begin{aligned}
\Phi(t) &= \frac{\Gamma\{(\theta+n)/\sigma\}}{\Gamma(K_n+\theta/\sigma)\Gamma(n/\sigma-K_n)}\frac{\Gamma\theta+n+1)}{\Gamma\{(\theta+n)/\sigma+1\}}\int_0^\infty \exp(\mathrm{i}tz)z^{K_n+\theta/\sigma-1}\int_z^\infty w(w-z)^{n/\sigma-K_n-1}\,g_\sigma(w)\,\mathrm{d}w\,\mathrm{d}z\\
&= \frac{\sigma\,\Gamma(\theta+n)}{\Gamma(K_n+\theta/\sigma)\,\Gamma(n/\sigma-K_n)}\int_0^\infty w\,g_\sigma(w)\int_0^w \exp(\mathrm{i}tz)z^{K_n+\theta/\sigma-1}(w-z)^{n/\sigma-K_n-1}\,\mathrm{d}z\,\mathrm{d}w\\
&= \frac{\Gamma(\theta+n+1)}{\Gamma\{(\theta+n)/\sigma+1\}}\sum_{r\geqslant 0}\frac{(\mathrm{i}t)^r}{r!}\frac{(K_n+\theta/\sigma)_r}{((\theta+n)/\sigma)_r}\int_0^\infty w^{(\theta+n)/\sigma+r}\,g_\sigma(w)\,\mathrm{d}w\\
&= \sum_{r\geqslant 0}\frac{(\mathrm{i}t^r)}{r!}\frac{(K_n+\theta/\sigma)_r}{((\theta+n)/\sigma))_r}\frac{\Gamma(\theta+n+1)}{\Gamma\{(\theta+n)/\sigma+1\}}\frac{\Gamma\{(\theta+n)/\sigma+r+1\}}{\Gamma(\theta+n+1+r\sigma)} = \sum_{r\geqslant 0}\frac{(\mathrm{i}t)^r}{r!}\mu_r^{(n)}
\end{aligned}$$

and the conclusion follows.

References

Chao, A. and Shen, T.-J. (2004) Non-parametric prediction in species sampling. *J. Agric. Biol. Environ. Statist.*, **9**, 253–269.

Charalambides, C. A. (2005) *Combinatorial Methods in Discrete Distributions*. Hoboken: Wiley.

Cont, R. and Tankov, P. (2004) *Financial Modelling with Jump Processes*. Boca Raton: Chapman and Hall–CRC.

Dunson, D. B. (2010) Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics* (eds N. L. Hjort, C. C. Holmes, P. Müller and S. G. Walker). Cambridge: Cambridge University Press. To be published.

Efron, B. and Thisted, R. (1976) Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika*, **63**, 435–447.

Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theoret. Popln Biol.*, **3**, 87–112.

Ferguson, T. S. and Klass, M. J. (1972) A representation of independent increments processes without Gaussian components. *Ann. Math. Statist.*, **43**, 1634–1643.

Ishwaran, H. and James, L. F. (2001) Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Ass.*, **96**, 161–173.

Jara, A., Lesaffre, E., De Iorio, M. and Quintana, F. (2008) Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Technical Report*. Katholieke Universiteit Leuven, Leuven.

Lijoi, A., Mena, R. H. and Prünster, I. (2007a) Bayesian nonparametric estimation of the probability of discovering 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 Bioinform.*, **8**, article 339.

Lijoi, A., Mena, R. H. and Prünster, I. (2008a) A Bayesian nonparametric approach for comparing clustering structures in EST libraries. *J. Computnl Biol.*, **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.

Mao, C. X. (2004) Prediction of the conditional probability of discovering a new class. *J. Am. Statist. Ass.*, **99**, 1108–1118.

Mao, C. X. and Lindsay, B. G. (2002) A Poisson model for the coverage problem with a genomic application. *Biometrika*, **89**, 669–682.

Müller, P. and Quintana, F. A. (2004) Nonparametric Bayesian data analysis. *Statist. Sci.*, **19**, 95–110.

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

Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Theory Reltd Flds*, **102**, 145–158.

Pitman, J. (1996a) Notes on the two parameter generalization of Ewens random partition structure. *Manuscript*. University of California, Berkeley. Unpublished.

Pitman, J. (1996b) Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (eds T. S. Ferguson, L. S. Shapley and J. B. MacQueen), pp. 245–267. Hayward: Institute of Mathematical Statistics.

Pitman, J. (1999) Brownian motion, bridge, excursion and meander characterized by sampling at independent uniform times. *Electron. J. Probab.*, **4**, 1–33.

Pitman, J. (2006) Combinatorial stochastic processes. *Lect. Notes Math.*, **1875**.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. (2000) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.

Rosiński, J. (2007) Tempering stable processes. *Stochast. Processes Appl.*, **117**, 677–707.

Solow, A. R. and Polasky, S. (1999) A quick estimator for taxonomic surveys. *Ecology*, **80**, 2799–2803.

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.

Wang, J.-P. Z. and Lindsay, B. G. (2005) A penalized nonparametric maximum likelihood approach to species richness estimation. *J. Am. Statist. Ass.*, **100**, 942–959.

Yamato, H. and Sibuya, M. (2000) Moments of some statistics of Pitman sampling formula. *Bull. Inform. Cybernet.*, **32**, 1–10.