Collegio Carlo Alberto

Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior

Stefano Favaro
Antonio Lijoi
Ramsés H. Mena
Igor Prünster

Carlo Alberto Notebooks

www.carloalberto.org/working_papers

Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior

S. Favaro[1], A. Lijoi[2], R.H. Mena[3] and I. Prünster[4]

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

[2] Università degli Studi di Pavia and CNR–IMATI, Milan, Italy.
E-mail: lijoi@unipv.it

[3] Universidad Nacional Autónoma de México, México, D.F.
E-mail: ramses@sigma.iimas.unam.mx

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

July 2009

Abstract

A Bayesian nonparametric methodology has been recently proposed in order 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 nonparametric estimators is cumbersome and prevents their implementation. In this paper 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 allows, combined with a suitable simulation scheme, 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 (EST) datasets.

Key words and phrases: Asymptotics; Bayesian nonparametrics; EST 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 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 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 achievable by means of a sample of size $n+m$ which, in other words, is the proportion of distinct species detectable in a sample of size $n + m$. Recently there has been a renewed interest in the area due to its importance in genomics as witnessed by the recent contributions of, e.g., Mao and Lindsay (2002), Mao (2004), Susko and Roger (2004) and Wang and Lindsay (2005). In such inferential problems one is 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 $\sum_{i=1}^{K_n} N_i = n$. 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) generated by sequencing 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 nonparametric 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 nonparametric approach has been set forth in Lijoi, Mena and Prünster (2007a). Assuming the data form an exchangeable sequence $(X_n)_{n \geq 1}$, by de Finetti's representation theorem $(X_n)_{n \geq 1}$ can be characterized by a hierarchical model, with the X_n's as a random sample from some distribution $\tilde{P}$ and a prior Π on $\tilde{P}$, that is

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

Their idea then consists in deriving estimators for quantities 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, Prünster and Walker (2008) for a theoretical study and Lijoi, Mena and Prünster (2007b) for a practitioner oriented illustration. Although the Bayesian nonparametric 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 considered in Lijoi et al. (2007a). Our primary aim is the achievement of a considerable simplification of the estimators 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, in order 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 to the point estimates. The results 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 not related to Bayesian Nonparametrics 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 analyzing five real EST datasets. Proofs are postponed to the Appendix.

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 $(V_k)_{k \geq 1}$ denote a sequence of independent random variables, with $V_k \sim \text{Beta}(\theta + k\sigma, 1 - \sigma)$. Define the stick–breaking weights as $\tilde{p}_1 = V_1$, $\tilde{p}_j = V_j \prod_{i=1}^{j-1}(1 - \tilde{p}_i)$ and suppose $(Y_n)_{n \geq 1}$ is a sequence of independent and identically distributed (i.i.d.) 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 $\tilde{P}_{\sigma,\theta} = \sum_{j \geq 1} \tilde{p}_j \, \delta_{Y_j}$ is a Poisson–Dirichlet process with parameters (σ, θ). For the sake of brevity we write $\text{PD}(\sigma, \theta)$. See Pitman (2006) for a detailed account on general theoretical aspects and, e.g., Ishwaran and James (2001), Navarrete, Quintana and Müller (2008), Jara, Lesaffre, De Iorio and Quintana (2008) for applications in Bayesian Nonparametrics.

3

Under model (1) with $\tilde{P}$ being a $\mathrm{PD}(\sigma, \theta)$ process, the sample coverage, 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

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

$$= \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{\prod_{i=j}^{j+k-1}(\theta + i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma, -n+j\sigma) \quad (2)$$

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 (16). Such an expression is the key for evaluating Bayesian estimators useful for inference with species sampling problems. In Lijoi et al. (2007a) it has been deduced, resorting to combinatorial arguments, as a particular case of a general class of priors. In the Appendix we provide an alternative proof of (2) since it introduces the way of reasoning we will resort to for proving Proposition 1.

Based on (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, 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

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

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 (2) and (3) becomes overwhelming. This happens, for instance, in genomic applications where one has to 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 (2) is still required for determining the corresponding highest posterior density (HPD) intervals, we will study the asymptotics of $K_m^{(n)}$, given K_n, as $m \to \infty$: this allows one to use the distribution of the limiting random quantity in order 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, one has

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

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

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

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 (6) admits an interesting probabilistic interpretation. Indeed, one has that

$$E[K_m^{(n)} \mid K_n = j] = P\left([X_{n+1} = \text{new} \mid K_n = j]\, E_{\sigma, \theta+n}[K_m]\right)$$

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 PD 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 (1996b) and Yamato and Sibuya (2000) and are recovered from (5) by setting $n = j = 0$.

The formulae obtained in Proposition 1 provide point estimators for quantities of interest in species sampling problems. Besides them, one 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. In order to overcome this drawback, we analyze 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 (1996b,1999). See also Pitman (2006). In order 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 $q \geq 0$, a positive random variable with density function

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

One, then, has 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 $K_n = j$ one has

$$\frac{K_m^{(n)}}{m^\sigma} \to Z_{n,j} \qquad \text{a.s.} \tag{10}$$

and in the p–th mean, for any $p > 0$, where $Z_{n,j} \stackrel{d}{=} B_{j+\theta/\sigma, \, n/\sigma-j} \, Y_{(\theta+n)/\sigma}$, $B_{a,b}$ 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,

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

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 induced by a beta random variable. The density of $Z_{n,j}$ in (10) can be formally represented as

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

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

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

Nonetheless, even in the latter case one cannot easily determine the quantiles of $Z_{n,j}$ we need to use in order 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 we are going to 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

$$f(w) = \frac{\sigma \Gamma(q\sigma)}{\Gamma(q)} w^{-q\sigma} f_\sigma(w) = \frac{\sigma}{\Gamma(q)} f_\sigma(w) \int_0^\infty u^{q\sigma-1} e^{-uw} du$$

Via augmentation, one then has

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

where $f(u)$ is the density function of a r.v. U_q such that $U_q^\sigma \sim \mathrm{Gamma}(q,1)$, and

$$f_\sigma(w|u) = f_\sigma(w) e^{-uw+u^\sigma}.$$

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

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

where $e_i \stackrel{iid}{\sim} \mathrm{Exp}(U_q)$, $v_i \stackrel{iid}{\sim} \mathrm{U}(0,1)$ and $a_1 > a_2 > \cdots$ 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 proposed in Cont and Tankov (2004).

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

1. Generate $B \sim \mathrm{Beta}(j + \theta/\sigma, n/\sigma - j)$.

2. In order to sample from $Y_{(\theta+n)/\sigma}$:

   2.a generate $X \sim \mathrm{Ga}((\theta+n)/\sigma, 1)$ and set $U = X^{1/\sigma}$;

   2.b for a given truncation N and U sampled in step 2.a, generate: $\{e_i\} \stackrel{iid}{\sim} \mathrm{Exp}(U)$, $\{v_i\} \stackrel{iid}{\sim} \mathrm{U}(0,1)$, $\xi_j \stackrel{iid}{\sim} \mathrm{Exp}(1)$ and take $a_i = \sum_{j=1}^i \xi_j$, for $i = 1, \ldots, N$;

   2.c compute W according to (12) and set $Y = W^{-\sigma}$.

3. Take $Z = BY$.

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

3   Applications to genomics

We now show how to use the results of the previous section by applying them to five real EST datasets. As briefly mentioned in the Introduction, 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, due 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. Based on such estimates, geneticists have to decide whether it is worth to proceed with sequencing and, in the affirmative case, also the size of the additional sample. Here, we consider: (a) a tomato-flower cDNA library (Quackenbush et al. 2000), previously analyzed 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 so to increase the discovery rate; (c) two *Naegleria gruberi* cDNA libraries prepared from cells grown under different culture conditions, aerobic and anaerobic (Susko and Roger, 2004).

In order to implement the $\mathrm{PD}(\sigma, \theta)$ 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

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

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

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

An alternative way of eliciting (σ, θ) is by placing a prior distribution on it. Such an approach is useful when one is interested in testing the compatibility of clustering structures among different populations (Lijoi, Mena and Prünster, 2008). However, in terms of estimates there are typically no relevant differences given the posterior distribution of (σ, θ) is highly concentrated. Hence, in order to keep the exposition as simple as possible, in the sequel we focus on $\mathrm{PD}(\sigma, \theta)$

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 datasets, the corresponding estimates for additional samples of size $m \in \{n, 10n, 100n\}$ are reported in Table 1 below together with the corresponding values n and j of the basic sample and the empirical Bayes specifications of (σ, θ).

<div align="center">TABLE 1 AROUND HERE</div>

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. In order to combine the point estimate for $K_m^{(n)}$ with an asymptotic 95% HPD interval, one 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, one obtains an asymptotic 95% HPD interval for $K_m^{(n)}$ as $(z_1 \, m^\sigma, z_2 \, m^\sigma)$. Table 2 below reports both exact and simulated mean and variance of the limiting random variable $Z_{n,j}$ associated to each of the five EST datasets as well as the simulated 95% and 99% HPD intervals. The sampled values are obtained by generating 2000 random variates according to the algorithm devised in Section 2 with truncation of the series in (12) given by $N = 3 \times 10^7$. In fact, it is important to get 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 in order 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 (11) with the sampled ones.

<div align="center">TABLE 2 AROUND HERE</div>

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 constructed through such a procedure is not centered on and, in most cases, does not even include the estimated number of new genes $E[K_m^{(n)} \,|\, K_n = j]$. Indeed, if one looks at the exact estimator for $K_m^{(n)}$ given in (6), it is clearly much smaller than its asymptotic approximation $m^\sigma E[Z_{n,j}]$. This is due to the fact that, 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 (17), one sees that an

<div align="center">9</div>

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, $E[K_m^{(n)} \mid K_n = j] \approx r_{\sigma,\theta,n}(m) \, E[Z_{n,j}]$ and the asymptotic HPD interval $(r_{\sigma,\theta,n}(m) \, z_1, r_{\sigma,\theta,n}(m) \, z_2)$ is approximately centered on the estimator $E[K_m^{(n)} \mid K_n = j]$, as desired. Table 3 displays, for the five datasets, the exact estimator for $K_m^{(n)}$, its asymptotic 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 $m \in \{n, 10n, 100n\}$.

<center>TABLE 3 AROUND HERE</center>

For the Tomato flower library we have that, even for $m = 100n = 258600$, 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 finite sample size approximation is definitely necessary to use $r_{\sigma,\theta,n}(m)$ as rate function.

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

<center>TABLE 4 AROUND HERE</center>

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 ones will become. Hence, if we approximate $K_m^{(n)}/r_{\sigma,\theta,n}(m)$ by its asymptotic random variable $Z_{n,j}$, we will necessarily underestimate its variability which is reflected on 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: (i) 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 have to be recomputed for any m of interest and cannot even be calculated for large m; (ii) the

<center>10</center>

fact that the length of the asymptotic HPD intervals is always shorter than the exact length (and not oscillating), allows to interpret it a "lower bound" on the length of the exact ones 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. Indeed, if one considers the asymptotic 99% HPD intervals, by Proposition 2, there exists a $\bar{m}$ 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 "upper bound" for the exact one. Although the determination of such a m suitable, for any choice of parameters and basic samples, is not possible one can proceed empirically. From Table 4, where the 99% asymptotic HPD intervals are reported as well, one sees 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 w.r.t. the other libraries is represented by a larger basic sample, such a covering is not yet 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. Figure 1 below 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 the two asymptotic HPD are quite close, we thus achieve a satisfactorily accurate estimate of the uncertainty.

<div align="center">FIGURE 1 AROUND HERE</div>

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 PD process estimator with other widely used estimators. Specifically, we consider the Tomato Flower library, which, among the considered datasets, has the largest observed sample ($n = 2586$) thus allowing an effective cross validation study. We take sub–samples 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 to compare the results also with estimators, which become unstable for larger sizes of the additional sample such as the popular estimator of Efron and Thisted (1976). The sub–samples 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 present in the additional sample, $K_m^{(n)}$, is then equal to $1825 - K_n$, since 1825 are the distinct genes present in the observed sample of size 2586. Predictions of $K_m^{(n)}$ are derived using, in addition to the PD process estimator, the following: (a) the estimator of Efron and Thisted (1976) which is based on gamma–mixed Poisson model; (b) the plug–in estimator of Solow and Polasky (1999); (c) the

nonparametric estimator of Chao and Shen (2004); (d) the penalized nonparametric MLE of Wang and Lindsay (2005). Estimators (a)–(c) are computed using SPADE software available from `http://chao.stat.nthu.edu.tw`, whereas estimator in (d) is calculated using the EST-stat Java program available at `http://bioinfo.stats.northwestern.edu/~jzwang`. In order to make the comparison on representative samples, we generated $10,000$ sub–samples 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 sub–sample: the corresponding empirical deciles are $839, 844, 847,$ $850, 852, 855, 858, 861, 865$. Samples with number of distinct genes belonging to the low (high) deciles correspond to situations of under–representation (over–representation) of distinct genes w.r.t. the distinct genes present in the whole sample. Table 5 below displays the results for 10 samples, where each sample corresponds to a different decile. The Chao–Shen estimator, which allows to tune a cut–off point (see Chao and Shen, 2004), is reported with cut–off point equal to the gene(s) with highest frequency. Lower cut–off points worsen the resulting estimates.

TABLE 5 AROUND HERE

Table 5 shows that the PD and Chao–Shen 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 PD estimator exhibits narrower uncertainty estimates: their average length is 92 genes, whereas for the CS estimator it is 165. In cases where the point estimate is accurate this represents an advantage but when this is not the case it may lead to miss the correct value as it happens for sample 10. If one prefers larger HPDs with the PD 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 to capture the true value. It also worth noting that the extreme situation with under or over–representation 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, 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 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.

5 Acknowledgements

6 Appendix

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

An important result proved in Pitman (1996a) 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 $\tilde{P}_{\sigma,\theta}$, i.i.d. from $\tilde{P}_{\sigma,\theta}$ and the sample $X_1, \ldots, X_n$ contains $j \leq n$ distinct values $X_1^*, \ldots, X_j^*$, then

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

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 $w_{j+1} = 1 - \sum_{i=1}^{j} w_i$.

In order to derive (2), we will make use of the posterior representation given in (15) and of the distributional properties of K_n. Indeed, from (15) one notes 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 observed in the basic sample with probability w. Consequently

$$\mathrm{pr}\left[K_m^{(n)} = k \mid K_n = j\right] = \frac{\Gamma(\theta + n)}{\Gamma(\theta + j\sigma)\Gamma(n - j\sigma)} \int_0^1 \mathrm{pr}\left[K_m^{(n)} = k \mid K_n = j, w\right] \times$$

$$\times \; w^{\theta+j\sigma-1}(1 - w)^{n-j\sigma-1} \, \mathrm{d}w$$

In order 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 not observed among the $K_n = j$ species of the basic sample. Hence we have

$$\mathrm{pr}\left[K_m^{(n)} = k \mid K_n = j, w\right] = \sum_{i=k}^{m} \binom{m}{i} w^i (1 - w)^{m-i} \, \mathrm{pr}[K_i = k]$$

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

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

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 (2) by noting that

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

where the second equality follows from (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. $\qquad \square$

A.2. *Proof of Proposition 1.*

Indeed, one has

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

14

where the unconditional moment $E[K_i^r]$ is evaluated w.r.t. $\tilde{P}_{\sigma,\theta+j\sigma}$ prior. Such an expression is already available from Pitman (1996b) 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, one has

$$E\left[(K_m^{(n)})^r \mid K_n = j\right]$$

$$= \frac{\Gamma(\theta+n)}{\Gamma(\theta+j\sigma)\Gamma(n-j\sigma)} \int_0^1 w^{\theta+j\sigma-1}(1-w)^{n-j\sigma-1} E\left[(K_m^{(n)})^r \mid K_n = j, w\right] dw$$

$$= \frac{\Gamma(\theta+n)}{\Gamma(\theta+j\sigma)\Gamma(n-j\sigma)} \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(1 + \frac{\theta+j\sigma}{\sigma}\right)_\nu S\left(r,\nu; \frac{\theta+j\sigma}{\sigma}\right) \times$$

$$\times \sum_{i=0}^{m} \binom{m}{i} \frac{(\theta+j\sigma+\nu\sigma+1)_{i-1}}{(\theta+1)_{i-1}} \int_0^1 w^{\theta+j\sigma+i-1}(1-w)^{n-j\sigma+m-i-1} dw$$

$$= \frac{1}{(\theta+n)_m} \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(1 + \frac{\theta+j\sigma}{\sigma}\right)_\nu S\left(r,\nu; \frac{\theta+j\sigma}{\sigma}\right) \frac{\theta+j\sigma}{\theta+j\sigma+\nu\sigma} \times$$

$$\times \sum_{i=0}^{m} \binom{m}{i} (\theta+j\sigma+\nu\sigma)_i (n-j\sigma)_{m-i}$$

$$= \frac{1}{(\theta+n)_m} \sum_{\nu=0}^{r} (-1)^{r-\nu} \left(\frac{\theta}{\sigma} + j\right)_\nu S\left(r,\nu; \frac{\theta+j\sigma}{\sigma}\right) (\theta+n+\nu\sigma)_m,$$

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

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

A.3. Proof of Proposition 2.

The proof strategy is as follows: we first adopt a technique similar to the one suggested in Pitman (2006, Theorem 3.8) for the unconditional case in order to establish that $K_m^{(n)}/m^\sigma$ converges a.s. and in the p–th 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

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

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 Proposition 1 in Lijoi, Prünster and Walker (2008), $q_{\sigma,\theta}^{(n)}(k) = \sigma^{K_n}(\frac{\theta}{\sigma} + K_n)_k/(\theta+n)_m$ for any integer $k \geq 1$ and $q_{\sigma,\theta}^{(n)}(0) := 1/(\theta+n)_m$. Hence $(M_{\sigma,\theta,m}^{(n)}, \mathscr{F}_m^{(n)})_{m \geq 1}$ 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 p–th mean as well, for any $p > 0$. One clearly has that $E_{\sigma,0}^{(n)}[M_{\sigma,\theta}^{(n)}] = 1$, where $E_{\sigma,0}^{(n)}$ denotes the expected value w.r.t. $P_{\sigma,0}^{(n)}$. It can be easily seen that

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

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

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

In order to identify the distribution of the limiting random variable $Z_{n,j}$ w.r.t. $P_{\sigma,\theta}^{(n)}$, we consider the asymptotic behaviour of $E[(K_m^{(n)})^r \mid K_n]$ as $m \to \infty$, for any $r \geq 1$. Letting $m \to \infty$ in (5) of Proposition 1, use the Stirling formula to obtain

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

Such a moment sequence clearly arises by taking $Z_{n,j} \stackrel{d}{=} B_{j+\theta/\sigma,\, n/\sigma-j} Y_{(\theta+n)/\sigma}$, 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 $\{\mu_r^{(n)}\}_r$. In order to establish this, one can evaluate the characteristic function of $Z_{n,j}$ which, at any $t \in \mathbb{R}$, coincides with

$$\Phi(t) = \frac{\Gamma\left(\frac{\theta+n}{\sigma}\right)}{\Gamma\left(K_n + \frac{\theta}{\sigma}\right)\Gamma\left(\frac{n}{\sigma} - K_n\right)} \frac{\Gamma(\theta+n+1)}{\Gamma\left(\frac{\theta+n}{\sigma} + 1\right)}$$
$$\times \int_0^\infty e^{itz} z^{K_n+\frac{\theta}{\sigma}-1} \int_z^\infty w\,(w-z)^{\frac{n}{\sigma}-K_n-1} g_\sigma(w)\,\mathrm{d}w\,\mathrm{d}z$$

$$= \frac{\sigma\,\Gamma(\theta+n)}{\Gamma\left(K_n + \frac{\theta}{\sigma}\right)\Gamma\left(\frac{n}{\sigma} - K_n\right)} \int_0^\infty w\, g_\sigma(w) \int_0^w e^{itz} z^{K_n+\frac{\theta}{\sigma}-1}(w-z)^{\frac{n}{\sigma}-K_n-1}\,\mathrm{d}z\,\mathrm{d}w$$

$$= \frac{\Gamma(\theta+n+1)}{\Gamma\left(\frac{\theta+n}{\sigma} + 1\right)} \sum_{r\geq 0} \frac{(\mathrm{i}t)^r}{r!} \frac{\left(K_n + \frac{\theta}{\sigma}\right)_r}{\left(\frac{\theta+n}{\sigma}\right)_r} \int_0^\infty w^{\frac{\theta+n}{\sigma}+r} g_\sigma(w)\,\mathrm{d}w$$

$$= \sum_{r\geq 0} \frac{(\mathrm{i}t)^r}{r!} \frac{\left(K_n + \frac{\theta}{\sigma}\right)_r}{\left(\frac{\theta+n}{\sigma}\right)_r} \frac{\Gamma(\theta+n+1)}{\Gamma\left(\frac{\theta+n}{\sigma} + 1\right)} \frac{\Gamma\left(\frac{\theta+n}{\sigma} + r + 1\right)}{\Gamma(\theta+n+1+r\sigma)} = \sum_{r\geq 0} \frac{(\mathrm{i}t)^r}{r!} \mu_r^{(n)}$$

and the conclusion follows. $\qquad\square$

References

Cont, R. and Tankov, P. (2004). *Financial modelling with jump processes.* Chapman & Hall/CRC, Boca Raton.

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

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

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. Population Biol.* **3**, 87-112.

Dunson, D.B. (2008). Nonparametric Bayes applications to biostatistics. *Duke Statistics Discussion Papers* 2008-06. To appear in *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C., Müller, P., Walker S.G. Eds.), Cambridge University Press, Cambridge.

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

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

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **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.

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

Lijoi A., Mena, R.H., Prünster, I. (2008). A Bayesian Nonparametric approach for comparing clustering structures in EST libraries. *J. Comput. Biol.* **15**, 1315–1327.

Lijoi, A., Prünster, I. and Walker, S.G. (2008). 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. Amer. Statist. Assoc.* **99**, 1108–18.

Mao, C. X. and Lindsay, B. G. (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669–82.

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. *Stat. Model.* **8**, 3-21.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102**, 145–58.

Pitman, J. (1996a). 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.), Hayward: Institute of Mathematical Statistics, 245–267.

Pitman, J. (1996b). Notes on the two parameter generalization of Ewens random partition structure. *Unpublished manuscript.*

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.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875. New York: Springer.

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 Research* **29**, 159–164.

Rosiński, J. (2007). Tempering stable processes. *Stochastic Process. 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.

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

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

Table 1: Analysis of the five EST datasets. 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 $\hat{E}_m^{(n,j)}$ 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 $m \in \{n, 2n, 3n\}$.

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-Norm.	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

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

Library	$\mathrm{E}[Z_{n,j}]$	$\mathrm{Var}[Z_{n,j}]$	$\bar{Z}_{n,j}$	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-Norm.	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)

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

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

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. All values are rounded to the nearest integer.

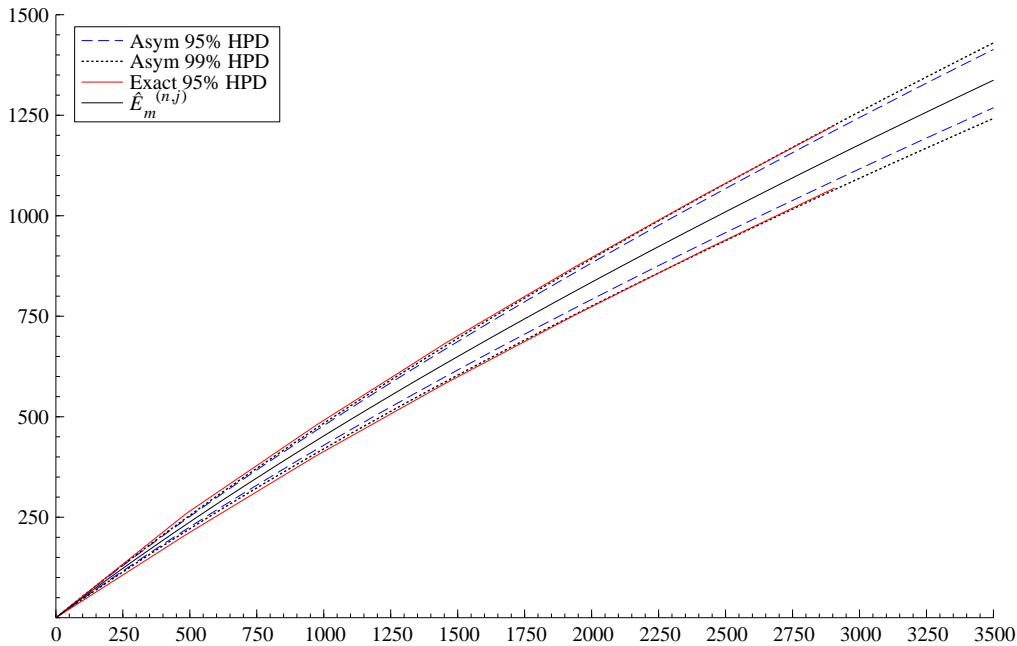| Library	m	$\hat{E}_m^{(n,j)}$	Exact 95% HPD	Asym. 95% HPD	Asym. 99% HPD
	n	1281	(1221 , 1341)	(1244 , 1318)	(1234 , 1329)
Tomato Flower	$2n$	2354	(2263 , 2449)	(2287 , 2422)	(2269 , 2442)
$n = 2586$	$3n$	3305	(3181 , 3434)	(3211 , 3400)	(3186 , 3430)
	n	346	(312 , 382)	(325 , 371)	(318 , 379)
Mastigamoeba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

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

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 ones). K_n reports the observed distinct genes in the sub–samples; the true $K_m^{(n)}$ is then given by $1825 - K_n$. Point and 95% uncertainty estimates are displayed for the PD estimator (PD), the Efron–Thisted estimator (ET), the Solow–Polasky estimator (SP), the Chao–Shen estimator (CS) and the Wang–Lindsay estimator (WL). All values are rounded to the nearest integer.

Est.\ Sample #	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)
ET	670	1000	900	790	800
	(360 , 980)	(760 , 1300)	(570 , 1200)	(510 , 1100)	(510 , 1100)
SP	899	926	907	928	940
	(818 , 980)	(844 , 1007)	(826 , 988)	(848 , 1008)	(860 , 1021)
CS	952	977	968	968	987
	(872 , 1033)	(897 , 1056)	(886 , 1049)	(885 , 1051)	(905 , 1069)
WL	909	927	918	933	948
	(834 , 955)	(857 , 983)	(842 , 966)	(849 , 983)	(881 , 1004)
Est.\ Sample #	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)
ET	650	640	1000	880	850
	(280 , 1000)	(58 , 1200)	(890 , 1200)	(690 , 1100)	(670 , 1000)
SP	927	941	939	956	950
	(846 , 1007)	(861 , 1021)	(859 , 1018)	(876 , 1036)	(869 , 1030)
CS	984	989	981	1000	1010
	(900 , 1068)	(901 , 1076)	(899 , 1062)	(916 , 1082)	(929 , 1092)
WL	933	950	939	964	960
	(867 , 1002)	(886 , 1013)	(866 , 1010)	(881 , 1026)	(893 , 1028)