Biometrika (2007), **94**, 4, *pp*. 769–786
doi: 10.1093/biomet/asm061
Advance Access publication 19 November 2007

Bayesian nonparametric estimation of the probability of discovering new species

BY ANTONIO LIJOI

Dipartimento di Economia Politica e Metodi Quantitativi, Università degli Studi di Pavia, 27100 Pavia, Italy
lijoi@unipv.it

RAMSÉS H. MENA

Departamento de Probabilidad y Estadística, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, México, 04510 México D.F., Mexico
ramses@sigma.iimas.unam.mx

AND IGOR PRÜNSTER

Dipartimento di Statistica e Matematica Applicata, Università degli Studi di Torino, 10122 Torino, Italy
igor@econ.unito.it

SUMMARY

We consider the problem of evaluating the probability of discovering a certain number of new species in a new sample of population units, conditional on the number of species recorded in a basic sample. We use a Bayesian nonparametric approach. The different species proportions are assumed to be random and the observations from the population exchangeable. We provide a Bayesian estimator, under quadratic loss, for the probability of discovering new species which can be compared with well-known frequentist estimators. The results we obtain are illustrated through a numerical example and an application to a genomic dataset concerning the discovery of new genes by sequencing additional single-read sequences of cDNA fragments.

Some key words: Bayesian nonparametrics; Gibbs-type random partition; Posterior probability of discovering a new species; Sample coverage; Species sampling.

1. INTRODUCTION

In biological and ecological studies, given that a sample of size n has been observed and j different species have been recorded, one is usually interested in the following statistical issues: (a) making inference about the number of unseen species; (b) estimating the probability that a further draw of m units from the population yields k new distinct species.

With reference to the problem of estimating the number of unobserved species, nonparametric approaches include those of Chao & Lee (1992), Shen et al. (2003) and Chao & Bunge (2002). A Bayesian approach is undertaken in Hill (1979), Boender & Rinnoy Kan (1987), Gandolfi & Sastri (2004) and Zhang & Stern (2005), whereas an empirical Bayes model is exploited in Efron & Thisted (1976). A sequential procedure for the determination of an optimal stopping of the sampling process is provided in Christen & Nakamura (2003). A very rich review, even though

now more than ten years old, can be found in Bunge & Fitzpatrick (1993). The subject is also closely connected to occupancy theory in probability and some nice examples can be found in Charalambides (2005).

As for issue (b), namely the estimation of the probability of discovering a new species, important references are Good (1953), Good & Toulmin (1956), Robbins (1968), Starr (1979), Chao (1981), Clayton & Frees (1987), Boneh et al. (1998) and Chao & Shen (2004). The determination of the discovery probability is intimately related to the classical problem of determining the optimal sample size in a species-sampling framework. The latter is typically faced by setting a threshold τ and making inference about the sample size m for which the probability of discovering a new species falls below τ. It is clear that the decay of the discovery probability as a function of m provides also a solution to the sample size problem, for which the Good–Toulmin estimator was designed. Recently, Mao (2004) has shown that the Good–Toulmin estimator can be seen as a nonparametric empirical Bayes estimator of the expected value of the discovery probability and can also be obtained as a moment-based estimator. Moreover, he proposes a likelihood-based estimator. In the present paper we focus on this problem and derive a closed-form expression for a nonparametric Bayes estimator of the probability of discovering a new species. A first attempt in this direction, based on the use of the Dirichlet process, is present in Tiwari & Tripathi (1989). The estimator we obtain is applied to a genomics dataset and is compared with previously known frequentist and empirical Bayes estimators. In this respect, it is worth mentioning that there have recently been contributions providing genuine nonparametric Bayesian counterparts to well-known empirical Bayes estimators for applied genetic problems; for example, see Do et al. (2005) who deal with the problem of estimating gene intensities in a mixture context.

The formal setting we deal with can be described as follows. Consider a population of individuals that can be grouped in different classes or species. If N is the total number of species, we denote by p_i the unknown proportion of individuals in the population belonging to species i. Suppose a sample of size n is drawn, and the number of distinct species being detected is equal to $j \in \{1, \ldots, N\}$. Moreover, $N_{i,n}$ represents the number of population units from the ith species. One might be then interested in making inference about $N - j$, i.e. the number of unseen species, or on

$$1 - U_n = \sum_{\{i:N_{i,n}=0\}} p_i, \tag{1}$$

which is the proportion of unobserved species, where U_n is known in the literature as the sample coverage. The interest in (1) can be motivated by concrete applied problems where the sampling procedure is expensive and further draws can only be motivated by the possibility of recording a new unobserved species. Hence, one can fix a possibly small threshold τ such that the sampling procedure takes place until the estimate of (1) becomes for the first time smaller than τ. This introduces a criterion for evaluating the effectiveness of further sampling. Moreover, it can provide a tool for assessing survey completeness. We will make these issues clearer when we consider applications in § 4.

The starting point of our approach is the randomization of the probabilities p_i. Next, we suppose that the observations X_i from the population are independent and identically distributed given a discrete random probability measure $\tilde{P} = \sum p_i \delta_{X_i}$. The law of $\tilde{P}$ plays the role of a nonparametric prior for Bayesian inference. When N is finite, the prior is a probability distribution on the $(N-1)$-dimensional simplex $\Delta_{N-1} = \{(p_1, \ldots, p_{N-1}) : p_i \geqslant 0, \sum_{i=1}^{N-1} p_i \leqslant 1\}$. A well-known example is the Dirichlet distribution. If N is large, it is reasonable to assume N to be infinite: this seems appropriate in many applications including those related to genomics. In such

a case a discrete nonparametric prior should be introduced. The most popular prior in Bayesian nonparametrics is the Dirichlet process, which has been introduced by Ferguson (1973); see Müller & Quintana (2004) for a review of the various uses of the Dirichlet process. Some of the drawbacks of the Dirichlet process have stimulated researchers to look for wide classes of random probabilities to be used as priors. Most of the proposals present in the literature, such as neutral to the right processes (Doksum, 1974), species-sampling models (Pitman, 1996) and normalized random measures with independent increments (Regazzini et al., 2003), contain the Dirichlet process as a special case and are almost surely discrete.

In the following sections we develop a Bayesian nonparametric treatment for species sampling problems by considering a class of priors which induce a random partition structure, for the observations, of Gibbs type. The notion of random exchangeable Gibbs partition is due to Pitman (2006) and is further considered, for different purposes, in recent papers by Gnedin & Pitman (2005), Berstycki & Pitman (2007) and in an International Center of Economic Research working paper by L.F. James, A. Lijoi and I. Prünster. The connection between random partition models and Bayesian nonparametric statistics is investigated in Quintana (2005).

2. Gibbs-type priors

Consider the following set-up. Let $(X_n)_{n \geqslant 1}$ be a sequence of exchangeable observations each taking values in some set $\mathbb{X}$. In other words, we suppose that there exists some random probability measure, $\tilde{P}$, whose probability distribution plays the role of a nonparametric prior and such that

$$\mathrm{pr}\,(X_1 \in A_1, \ldots, X_n \in A_n | \tilde{P}) = \prod_{i=1}^{n} \tilde{P}(A_i)$$

for any $n \geqslant 1$ and any subsets $A_1, \ldots, A_n$ of $\mathbb{X}$. We assume that $\tilde{P}$ is discrete with probability one and that $E\{\tilde{P}(\cdot)\} = P_0(\cdot)$, where P_0 is nonatomic, i.e. $P_0(\{x\}) = 0$ for any x in $\mathbb{X}$. Hence, the ties in the data $X_1, \ldots, X_n$ are explained by the discrete nature of $\tilde{P}$: the number of distinct observations, K_n, is an integer less than or equal to n. Such distinct observations identify the K_n different species being recorded. When $K_n = k$ different species are observed, we label them as $X_1^*, \ldots, X_k^*$, and $N_{j,n}$ represents the number of individuals in the n-sample $(X_1, \ldots, X_n)$ that belong to the jth species. The priors we will consider induce a joint distribution of K_n and of $(N_{1,n}, \ldots, N_{K_n,n})$ of the form

$$\mathrm{pr}\,[\{K_n = k\} \cap \{N_{j,n} = n_j,\ j = 1, \ldots, k\}] = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1} \tag{2}$$

for some $\sigma \in (0, 1)$, for some set of nonnegative weights $\{V_{n,k} : n \geqslant 1, 1 \leqslant k \leqslant n\}$ and where $(a)_n = a(a+1)\cdots(a+n-1)$ for any $n \geqslant 1$, and $(a)_0 = 1$. Note that the distribution is invariant with respect to permutations of $(n_1, \ldots, n_k)$. The random partitions of the observations identified by (2) are known as Gibbs-type random partitions; see Pitman (2006) and Gnedin & Pitman (2005). The distribution in (2) leads to predictive distributions for the observations that admit the representation

$$\mathrm{pr}\,(X_{n+1} \in A | X_1, \ldots, X_n) = \frac{V_{n+1,k+1}}{V_{n,k}} P_0(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma) \delta_{X_j^*}(A),$$

given that $(X_1, \ldots, X_n)$ is a sample of size n featuring $K_n = k$ different observations $X_1^*, \ldots, X_k^*$ with frequencies $n_1, \ldots, n_k$, respectively. The predictive distribution provides some insight

into the inferential implications associated with the specification in (2). The sampling scheme is such that the probability of sampling a new species depends solely on n and k: given that a new species will be observed, its label is generated according to P_0, whereas, if an 'old' species is observed, the probability of observing the jth depends on the frequency n_j and the parameter σ. It is worth recalling that each Gibbs-type partition uniquely determines a discrete nonparametric prior: for this reason, we also speak of Gibbs-type priors. Various noteworthy priors fall within this class, namely the Dirichlet process, the two-parameter Poisson–Dirichlet process and the normalized inverse Gaussian process.

Remark 1. Whenever the joint distribution of K_n and of $(N_{1,n}, \ldots, N_{K_n,n})$ computed at some point $(k, n_1, \ldots, n_k)$ is invariant with respect to permutations of $(n_1, \ldots, n_k)$, then such a distribution is known in the probability literature as an exchangeable partition probability function, extensively studied in Pitman (1995, 2006). An exchangeable partition probability function identifies the law of an exchangeable random partition $\tilde{\mathscr{P}}$ of the set of integers $\mathbb{N}$. The most celebrated example of an exchangeable partition probability function is the Ewens sampling formula (Ewens, 1972), a cornerstone of population genetics.

We shall assume that the species proportions p_i $(i = 1, 2, \ldots)$ within the population are random and give rise to a discrete random probability measure $\tilde{P} = \sum p_i \delta_{X_i}$ of Gibbs type. In the following section we assume that a sample $X^{(n)}$ of size n has been drawn from the population and, given the number of species recorded among the X_i's, we evaluate the probability that a certain number of new species will be observed in the sample $(X_{n+1}, \ldots, X_{n+m})$.

3. ESTIMATING THE PROBABILITY OF DISCOVERING A NEW SPECIES

Consider a population which is composed of an ideally infinite number of species. Let $X_1, \ldots, X_n$ be a sample of size n, also called the 'basic sample'. The distribution of the number of species K_n present in the sample, under the assumption that the X_i's are generated by a Gibbs-type prior,

$$\mathrm{pr}(K_n = k) = \frac{V_{n,k}}{\sigma^k} \mathscr{C}(n, k; \sigma), \tag{3}$$

(Gnedin & Pitman, 2005) where $\mathscr{C}(n, k; \sigma)$ is a generalized factorial coefficient. Such coefficients are easily computable and a short account is provided in the Appendix. The quantity in (3) is interpreted as the prior distribution on the number of species in the sample, of size n, to be observed. Next, a further sample of m individuals is selected thus giving rise to the 'enlarged sample' of size $n + m$. If one knows the number of species observed in the first n samples and the frequency with which each species has been recorded, it would be interesting to determine both (i) the probability distribution of the number of new species observed among the $X_{n+1}, \ldots, X_{n+m}$, and (ii) the probability of observing a new species at the $(n + m + 1)$th draw, without actually observing the intermediate m-sample $X_{n+1}, \ldots, X_{n+m}$: this automatically provides a solution to the important problem of determining the sample size such that the probability of discovering a new species falls below a given threshold. We first illustrate some notation we are going to use throughout. We denote by $X_j^{(1,n)} = (X_1, \ldots, X_n)$ a basic sample of size n containing j distinct species, with $j \in \{1, \ldots, n\}$. Analogously, $X^{(2,m)} = (X_{n+1}, \ldots, X_{n+m})$ is the second, unobserved, sample of size m. Moreover, let $K_m^{(n)} = K_{n+m} - K_n$ be the number of new species in $X^{(2,m)}$ and denote by $X_k^{(2,m)}$ the new m-sample featuring $K_m^{(n)} = k$. Evaluating the probability in (i) is equivalent to determining $\mathrm{pr}(K_m^{(n)} = k | X_j^{(1,n)})$, for any $k = 0, 1, \ldots, m$ and for any

$j = 1, \ldots, n$, which can be interpreted as the posterior probability distribution of the number of species to be observed in a sample of size m.

PROPOSITION 1. *Let $(X_n)_{n \geqslant 1}$ be a sequence of exchangeable observations governed by a Gibbs-type prior. Then, for any $k \in \{0, 1, \ldots, m\}$,*

$$\mathrm{pr}\big(K_m^{(n)} = k | X_j^{(1,n)}\big) = \frac{V_{n+m,j+k}}{V_{n,j}} \frac{1}{\sigma^k} \mathscr{C}(m, k; \sigma, -n + j\sigma) \tag{4}$$

for any $j \in \{1, \ldots, n\}$. This also implies that K_n is sufficient for predicting the number of new distinct observations.

The coefficient $\mathscr{C}(m, k; \sigma, -n + j\sigma)$ on the right-hand side of (4) is the so-called noncentral generalized factorial coefficient; see the Appendix and references therein. Note that the sufficiency of K_n leads to a simple interpretation of the Gibbs structure in statistical terms: Gibbs priors are priors which lead to prediction of the number of new species based only on the number of distinct observations in the sample. This structural assumption underlies various priors such as the two-parameter Poisson–Dirichlet process and the normalized inverse Gaussian prior.

Turning to problem (ii), we now derive a Bayesian estimator for the probability of discovering a new species at the $(n + m + 1)$th draw, given the basic sample $X_j^{(1,n)}$. If we suppose, for the moment, that we have observed both the basic sample and the second sample, the discovery probability is given by $\mathrm{pr}\,(K_1^{(n+m)} = 1 | X_j^{(1,n)}, X_k^{(2,m)})$. By virtue of the highlighted sufficiency of the number of distinct species, the discovery probability is also equal to $\mathrm{pr}\,(K_1^{(n+m)} = 1 | K_n = j, K_m^{(n)} = k)$. However, our estimate is obtained without observing the outcome of the second sample $X^{(2,m)}$ and, hence, we have to estimate the random probability

$$D_m^{(n:j)} := \mathrm{pr}\,\big(K_1^{(n+m)} = 1 \big| K_n = j, K_m^{(n)}\big), \tag{5}$$

where the randomness in the above expression is due to the randomness of $K_m^{(n)}$. Bayesian inference on (5) is based on the posterior distribution of $K_m^{(n)}$ given $K_n = j$ provided in Proposition 1. Thus, the Bayesian estimator of (5), with respect to a squared loss function, is given by its expected value with respect to the posterior distribution of the number of species. This represents a Bayesian counterpart of the celebrated Good–Toulmin estimator. In other words, we provide a Bayesian nonparametric estimator for $U_{n+m} = \sum_{i \geqslant 1} p_i \mathbb{I}_{\{0\}}(N_{i,n+m})$, where $\mathbb{I}_A$ denotes the indicator function of set A.

PROPOSITION 2. *Let $(X_n)_{n \geqslant 1}$ be a sequence of exchangeable random variables governed by a Gibbs-type prior. Then the Bayes estimate, under a squared loss function, of the probability of observing a new species at the $(n + m + 1)$th draw, conditional on the basic sample $X_j^{(1,n)}$ with j distinct species, is given by*

$$\hat{D}_m^{(n:j)} = \sum_{k=0}^{m} \frac{V_{n+m+1,j+k+1}}{V_{n,j}} \frac{1}{\sigma^k} \mathscr{C}(m, k; \sigma, -n + j\sigma). \tag{6}$$

Remark 2. An important feature of the distribution given in (4) and of the estimator given in (6) is that they can be computed exactly with little computational effort once a closed-form expression for the $V_{n,k}$'s is available. Hence, we now provide some specific examples of Gibbs-type priors where this is the case, allowing an immediate exact implementation of (4) and (6).

Example 1: *The Dirichlet process*. The Dirichlet process can be seen to be a Gibbs-type prior by letting $\sigma \to 0$. If $\tilde{P}$ is a Dirichlet process with parameter measure α such that $\alpha(\mathbb{X}) = \theta \in$

$(0, +\infty)$, then $V_{n,k} = \theta^k/(\theta)_n$. In this case, the prior distribution on the number K_n of distinct species in the sample $X^{(n)}$ has been derived in Ewens (1972) and Antoniak (1974). In the Gibbs setting such a distribution can be easily recovered by (3), giving rise to $\theta^k|s(n,k)|/(\theta)_n$ which is a version of the celebrated Ewens sampling formula. Here, $|s(n,k)|$ stands for the sign-less Stirling number of the first kind. The relationships between generalized factorial coefficients and Stirling numbers are recalled in the Appendix. In the specific case of the Dirichlet process, an attempt at deriving a closed-form expression for the discovery probability can be found in Tiwari & Tripathi (1989): their result is not of immediate application since the exact evaluation of the estimator they obtain requires a heavy computational burden. Here, the additional assumption of α being nonatomic together with Proposition 1 lead to an easy expression. The assumption of α being nonatomic does not cause a loss of generality since it just serves as a labelling procedure for the different species.

We now derive the quantities of interest. Proposition 1 yields an expression for the posterior distribution of the number of distinct species to be observed in the enlarged sample $X^{(2,n)}$, which coincides with

$$\text{pr}\,(K_m^{(n)} = k | X_j^{(1,n)}) = \frac{\theta^k(\theta)_n}{(\theta)_{n+m}} \sum_{l=k}^{m} \binom{m}{l} |s(l,k)|(n)_{m-l}$$

for any $k \in \{0, 1, \ldots, m\}$. Hence, the probability of discovering a certain number of new species does not depend on the number j of species recorded in the basic sample. This particular feature of the Dirichlet process is clearly undesirable from an inferential point of view since inference about the number of distinct species in a future sample would not depend on the number of distinct species present in the basic sample. This is reflected in the discovery estimator,

$$\hat{D}_m^{(n:j)} = \frac{\theta}{(\theta+n)_{m+1}} \sum_{k=0}^{m} \theta^k \sum_{l=k}^{m} \binom{m}{l} |s(l,k)|(n)_{m-l},$$

which does not depend on j. It is easy to see that this property characterizes the Dirichlet process within the class of Gibbs priors. Thus, any other Gibbs-type prior makes use of the information about the number of distinct species in the basic sample and is suitable for our purposes.

Example 2: *The two-parameter Poisson–Dirichlet process*. This family of random probability measures has been introduced in Pitman (1995). It is a very popular class of models which has found applications in various areas including excursion theory, combinatorics, Bayesian mixture models and population genetics, in particular fragmentations and coalescents; see Pitman (2006) and references therein. The joint distribution of K_n and of $(N_{1,n}, \ldots, N_{K_n,n})$ for a (σ, θ)-parameter Poisson–Dirichlet process is

$$\frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}, \tag{7}$$

which is known as Pitman's sampling formula. In the above we set $\prod_{i=1}^{0}(\theta + i\sigma) = 1$. The distribution of the number of distinct observations, within a sample of size n, coincides with

$$\text{pr}\,(K_n = k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma^k(\theta+1)_{n-1}} \mathscr{C}(n,k;\sigma).$$

As far as the posterior distribution is concerned, one applies Proposition 1 to obtain

$$\text{pr}\left(K_m^{(n)} = k \middle| X_j^{(1,n)}\right) = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{\prod_{i=j}^{j+k-1}(\theta+i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma) \tag{8}$$

for all $k \in \{0, 1, \ldots, m\}$. A straightforward application of Proposition 2 yields the corresponding discovery estimator,

$$\hat{D}_m^{(n:j)} = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \sum_{k=0}^{m} \frac{\prod_{i=j}^{j+k}(\theta+i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma). \tag{9}$$

Unlike the Dirichlet case, such an estimator depends on the number of distinct observations j present in the basic sample.

Example 3: *The normalized inverse Gaussian process*. The normalized inverse Gaussian process has been recently studied in Lijoi et al. (2005). The corresponding joint distribution of K_n and of $(N_{1,n}, \ldots, N_{K_n,n})$ is given in equation (A1) of the above paper and one immediately sees that the normalized inverse Gaussian process prior is a Gibbs-type random probability measure, with $\sigma = 1/2$ and

$$V_{n,k} = \frac{e^{\theta}(-\theta^2)^{n-1}}{\Gamma(n)2^{k-1}} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\theta^2)^{-i} \Gamma(k+2+2i-2n;\theta),$$

where θ is some positive constant and $\Gamma(\nu, x) = \int_x^{+\infty} t^{\nu-1}e^{-t}dt$ is the incomplete gamma function. The prior distribution for K_n is provided in Proposition 4 of Lijoi et al. (2005). As for the determination of the posterior distribution, application of Proposition 1 together with some algebra yields

$$\text{pr}\left(K_m^{(n)} = k \middle| X_j^{(1,n)}\right) = \frac{(-\theta^2)^m 2^k}{(n)_m \Gamma(k)} \frac{\sum_{i=0}^{n+m-1} \binom{n+m-1}{i} (-\theta^2)^{-i} \Gamma\{j+k+2+2i-2(m+n);\theta\}}{\sum_{i=0}^{n-1} \binom{n-1}{i} (-\theta^2)^{-i} \Gamma(j+2+2i-2n;\theta)}$$
$$\times \sum_{s=k}^{m} \binom{m}{s} \binom{2s-k-1}{s-1} \frac{\Gamma(s)}{2^{2s}} \left(n-\frac{j}{2}\right)_{m-s},$$

for each $k = 1, \ldots, m$. Finally, our Bayesian estimator for the discovery probability is

$$\hat{D}_m^{(n:j)} = \frac{(-\theta^2)^{m+1}}{(n)_{m+1}} \sum_{k=0}^{m} \frac{\sum_{i=0}^{n+m} \binom{n+m}{i} (-\theta^2)^{-i} \Gamma\{j+k+1+2i-2(m+n);\theta\}}{\sum_{i=0}^{n-1} \binom{n-1}{i} (-\theta^2)^{-i} \Gamma(j+2+2i-2n;\theta)}$$
$$\times \sum_{s=k}^{m} \binom{m}{s} \binom{2s-k-1}{s-1} \frac{2^{k-2s}\Gamma(s)}{\Gamma(k)} \left(n-\frac{j}{2}\right)_{m-s}.$$

4. Illustrations

4·1. *A simple numerical example*

Suppose a dataset of n observations is to be collected. Once the n observations are collected, the number of distinct ones j is recorded, and a prediction of the number of new distinct observations within another dataset of m observations has to be provided.

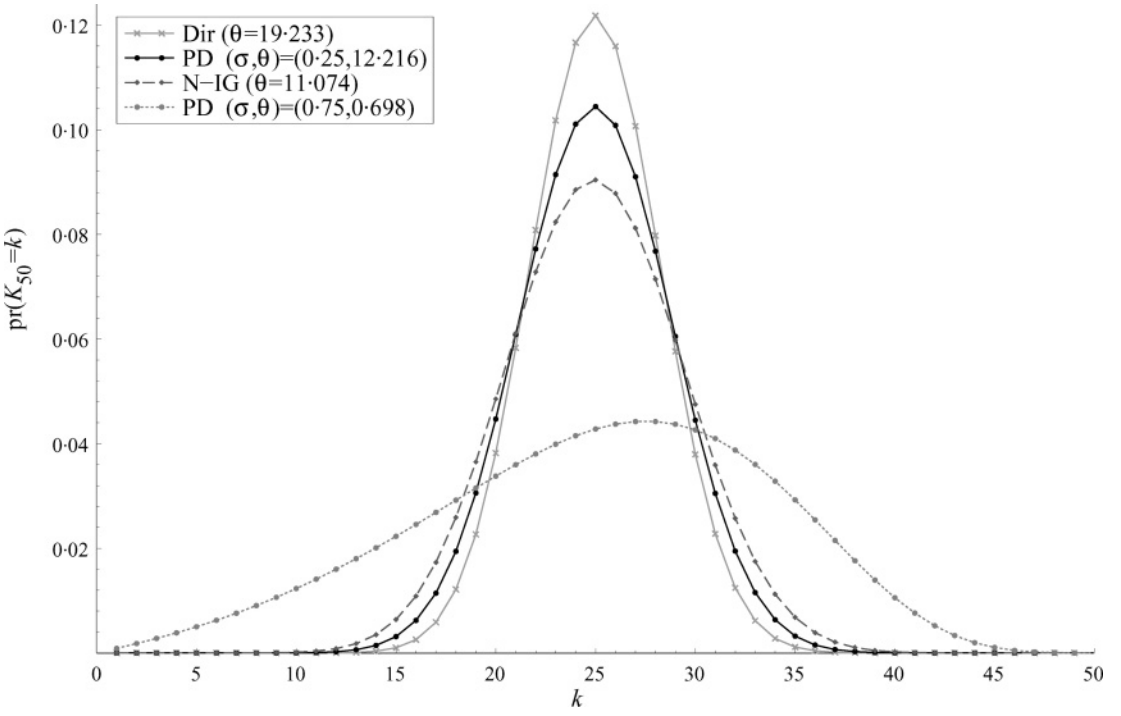

Fig. 1. Synthetic example. Prior probabilities for K_{50} corresponding to the Dirichlet process, the two choices of Poisson–Dirichlet process and the normalized inverse Gaussian process such that $E(K_{50}) = 25$.

We will compare the behaviour of the Dirichlet, the two-parameter Poisson–Dirichlet process and the normalized inverse Gaussian process. Suppose the number of observations to be collected at the first stage is $n = 50$ and the prior guess of the number of distinct ones is $j = 25$. Translating this prior guess into a prior specification results in a choice of parameters such that $E(K_{50}) = 25$. For the Dirichlet process this is achieved by setting $\theta = 19{\cdot}233$, while for the normalized inverse Gaussian process one needs to set $\theta = 11{\cdot}074$. For the Poisson–Dirichlet process, with two free parameters, there are many possible choices of (σ, θ) for which $E(K_{50}) = 25$. For comparison purposes, we choose $\sigma = 0{\cdot}25$ and $\sigma = 0{\cdot}75$, which lead to $(\sigma, \theta) = (0{\cdot}25, 12{\cdot}216)$ and $(\sigma, \theta) = (0{\cdot}75, 0{\cdot}698)$. Figure 1 displays the four corresponding prior distributions of K_{50}. The Dirichlet process is the one most concentrated around 25 and the Poisson–Dirichlet process with parameter $(\sigma, \theta) = (0{\cdot}75, 0{\cdot}698)$ represents the least informative prior. Note that θ controls the location of the prior distribution of K_n. Hence, a low value of σ, combined with a small θ, concentrates the distribution on small numbers of species, whereas large values for both σ and θ shift the mass towards large numbers of species.

Given that a sample of size $n = 50$ has been collected and the number j of distinct observations has been recorded, one can compute the posterior distribution of the number of new distinct observations in an additional dataset of size $m = 50$. The posterior distributions of $(K_{50}^{(50)} | K_{50} = j)$ corresponding to $j \in \{5, 25, 45\}$ are depicted in Fig. 2, and the corresponding Bayes estimates, together with their 95% highest posterior density intervals, are provided in Table 1.

The behaviour of the various random probability measures does not change significantly if the sample sizes n and m are modified and hence some structural conclusions can be derived. First, the inadequacy of the Dirichlet process is apparent, since the distribution of $(K_m^{(n)} | K_n = j)$ does not depend on j. Secondly, both the Poisson–Dirichlet process and the normalized inverse Gaussian

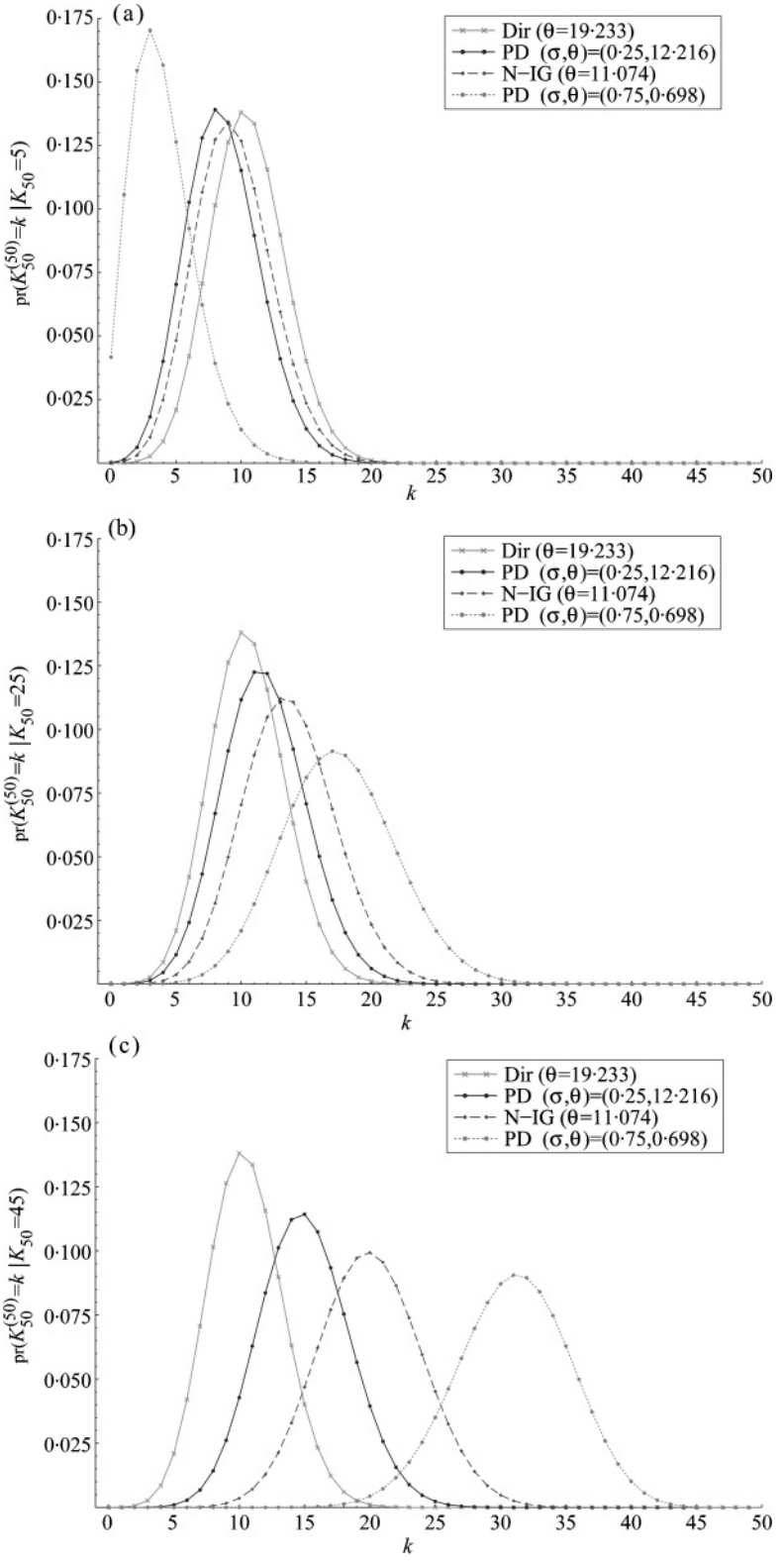

Fig. 2. Synthetic example. Posterior probability distributions of $(K_{50}^{(50)}|K_{50} = j)$ for (a) $j = 5$, (b) $j = 25$, (c) $j = 45$ corresponding to the Dirichlet process, the two choices of the Poisson–Dirichlet process and the normalized inverse Gaussian process.

Table 1. *Synthetic example. Posterior expected number of new observations* $E(K_{50}^{(50)}|K_{50} = j)$ *and 95% highest posterior density intervals corresponding to the Dirichlet process, the two choices of Poisson–Dirichlet process and the normalized inverse Gaussian process.*

$m = n = 50$	$j = 5$	$j = 25$	$j = 45$
Dirichlet process $\theta = 19{\cdot}233$	$10{\cdot}51 \in (4, 16)$	$10{\cdot}51 \in (4, 16)$	$10{\cdot}51 \in (4, 16)$
PD process $(\sigma, \theta) = (0{\cdot}25, 12{\cdot}216)$	$8{\cdot}6 \in (2, 14)$	$11{\cdot}79 \in (4, 18)$	$14{\cdot}99 \in (8, 22)$
N-IG process $\theta = 11{\cdot}074$	$9{\cdot}4 \in (3, 15)$	$13{\cdot}7 \in (6, 20)$	$20{\cdot}03 \in (12, 28)$
PD process $(\sigma, \theta) = (0{\cdot}75, 0{\cdot}698)$	$4 \in (0, 11)$	$17{\cdot}49 \in (8, 26)$	$30{\cdot}99 \in (22, 40)$

PD process, Poisson–Dirichlet process; N–IG process, normalized inverse Gaussian process.

process lead sensibly to posterior inferences monotone in j. With the fairly noninformative Poisson–Dirichlet process prior with parameters $(\sigma, \theta) = (0{\cdot}75, 0{\cdot}698)$ the posterior distribution unsurprisingly adheres very closely to the structure of the observed data.

4·2. *Analysis of a dataset from genomics*

An important area of application of the results in Propositions 1 and 2 concerns the analysis of expressed sequence tags in genomics. Expressed sequence tags are single-read sequences of cDNA fragments obtained by sequencing randomly selected cDNA clones from a cDNA library. Since a cDNA library consists of millions of cDNA clones, only a small fraction is usually sequenced because of cost constraints; see Mao (2004) for further references and details. This is a natural setting in which the estimation of the probability of discovering a new species is relevant: knowledge of the costs associated with sampling might suggest a threshold τ below which it is not convenient to proceed with sampling. Using the same expressed sequence tags dataset as in Mao (2004) allows us to draw a comparison with the frequentist estimators. The dataset concerns a cDNA library made from the 0 mm to 3 mm buds of tomato flowers (Mao & Lindsay, 2002; Mao, 2004). The basic sample consists of $n = 2586$ expressed sequence tags and this gives $j = 1825$ different cDNA fragments each of which represents a unique gene. If r_i denotes the number of clusters of size i, then the dataset gives $r_i = 1434, 253, 71, 33, 11, 6, 2, 3, 1, 2, 2, 1, 1, 1, 2, 1, 1$ with $i \in \{1, 2, \ldots, 14\} \cup \{16, 23, 27\}$. This means we are observing 1434 clusters of size 1, 253 clusters of size 2, and so on. In order to make direct comparison with the results summarized in Table 1 of Mao (2004), we choose $m \in \{517, 1034, 1552, 2069, 2586\}$, which correspond to 20%, 40%, 60%, 80% and 100% of the size of the basic sample.

As a prior distribution we use a two-parameter Poisson–Dirichlet process. A sensible strategy for selecting the values of θ and σ might rely upon empirical considerations. Indeed, we suggest a maximum likelihood procedure: consider the joint distribution, $\phi_{n_1,\ldots,n_j}(\sigma, \theta)$, of K_n and $(N_{1,n}, \ldots, N_{K_n,n})$, given in (7), evaluated at $(j, n_1, \ldots, n_j)$ as a function of σ and θ. In our case $j = 1825$ and the n_i's can be recovered from the r_i's, i.e. $n_1 = \cdots = n_{1434} = 1$, $n_{1435} = \cdots = n_{1687} = 2$ and so on. As a result of exchangeability, the ordering of the n_i's has no influence. Hence, we choose $\theta = \theta^*$ and $\sigma = \sigma^*$ such that

$$\phi_{n_1,\ldots,n_j}(\sigma^*, \theta^*) = \max_{\sigma,\theta} \phi_{n_1,\ldots,n_j}(\sigma, \theta).$$

In our case the likelihood is unimodal with maximum at $(\sigma^*, \theta^*) = (0{\cdot}612, 741)$. It is interesting to note that, if we fix $\sigma = 0{\cdot}612$, the value of θ which yields $E(K_{2586}) = 1825$, a common choice in Bayesian prior specification, is exactly $\theta = 741$.

We will also consider a less elaborate prior specification with an intermediate choice of $\sigma = 0{\cdot}5$ combined with θ such that $E(K_{2586}) = 1825$. This gives $(\sigma, \theta) = (0{\cdot}5, 1093{\cdot}313)$. Indeed, empirical investigations with simulated data seem to suggest that $\sigma = 0{\cdot}5$ is always a good

Table 2. *Genomics example. Posterior expected number of new genes* $E(K_m^{(2586)}|K_{2586} = 1825)$ *and 95% highest posterior density intervals for* $m \in \{517, 1034, 1552, 2069, 2586\}$.

m	Poisson–Dirichlet process $(\sigma, \theta) = (0{\cdot}612, 741)$	Poisson–Dirichlet process $(\sigma, \theta) = (0{\cdot}5, 1093{\cdot}313)$
517	280·59 ∈ (257, 305)	272·58 ∈ (249, 297)
1034	546·87 ∈ (512, 582)	528·83 ∈ (495, 563)
1552	801·62 ∈ (758, 846)	771·83 ∈ (729, 815)
2069	1045·6 ∈ (994, 1098)	1002·6 ∈ (953, 1053)
2586	1280·6 ∈ (1221, 1341)	1223·2 ∈ (1166, 1280)

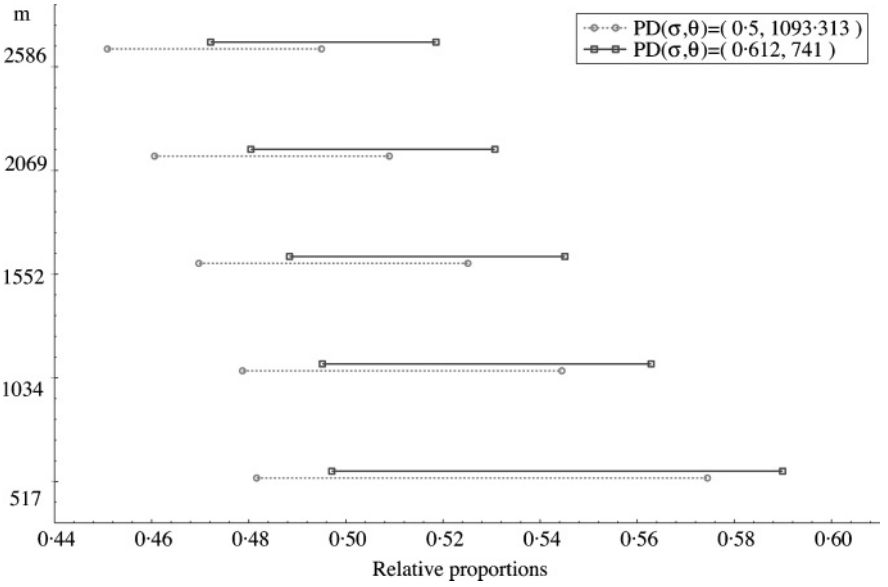

Fig. 3. Genomics example. The figure shows 95% highest posterior density intervals for the relative proportion of new genes in a sample of size $m = 517, 1034, 1552, 2069$ and 2586 based on the two choices of the Poisson–Dirichlet process.

choice when no precise prior information is available. Note, however, that this procedure does not incorporate information about the number of clusters of a given size, r_i.

Finally, one could also specify a prior for (σ, θ), implementation of which would be straightforward. However, the size of the dataset and the fact that the parameters directly describe the distribution of the observables suggested the use of an 'empirical Bayes' estimate of (σ, θ) for the present application.

First we consider the posterior distribution of K_n given the basic sample, namely the distribution of $(K_m^{(2586)}|K_{2586} = 1825)$ for $m \in \{517, 1034, 1552, 2069, 2586\}$. The expected number of new genes together with the corresponding 95% highest posterior density intervals are displayed in Table 2 for the two different parameter specifications. In Fig. 3, instead of the absolute numbers, the relative proportions of new genes for samples of size $m \in \{517, 1034, 1552, 2069, 2586\}$ are depicted.

Overall, the specification (0·612, 741) leads us to predict a slightly larger proportion of new genes, as can be explained as follows. In choosing $(\sigma, \theta) = (0{\cdot}612, 741)$ we made explicit use of the frequency of clusters of a given size. Indeed, the requirement $K_{2586} = 1825$ allows for a

Table 3. *Genomics example. Estimates in percentages for* $m \in \{517, 1034, 1552, 2069, 2586\}$ *obtained with the estimator* $\hat{D}_m^{(2586:1825)}$ *arising from the two choices of the Poisson–Dirichlet process, from the moment–based estimator* $\hat{U}_e$ *and from the likelihood–based estimator* $\tilde{U}_e$. *For the Poisson–Dirichlet processes, the 95% highest posterior density intervals are also shown.*

m	Poisson–Dirichlet process $(\sigma, \theta) = (0{\cdot}612, 741)$	Poisson–Dirichlet process $(\sigma, \theta) = (0{\cdot}5, 1093{\cdot}313)$	$\hat{U}_e$	$\tilde{U}_e$
0	55·84	54·52	55·46	55·45
517	52·80 ∈ (52·42, 53·19)	51·04 ∈ (50·76, 51·33)	51·86	51·83
1034	50·28 ∈ (49·79, 50·77)	48·17 ∈ (47·80, 48·53)	48·74	48·66
1552	48·14 ∈ (47·59, 48·69)	45·72 ∈ (45·31, 46·13)	45·99	45·74
2069	46·30 ∈ (45·70, 46·88)	43·62 ∈ (43·18, 44·05)	43·51	42·80
2586	44·68 ∈ (44·06, 45·30)	41·78 ∈ (41·32, 42·23)	41·24	39·98

number of clusters of size 1 in the range [1064, 1824], where the most realistic configurations are not close to the upper bound; for example $r_1 = 1824$ would imply that $r_{762} = 1$ and $r_i = 0$ for all other i. The actual number of clusters of size 1, namely 1434, is relatively high and this naturally leads to expect a similar configuration in the new sample which is tantamount of expecting a relatively high number of new genes. This is in accordance with the behaviour of the estimator $\hat{U}_n = 1 - (r_1/n)$ attributed to Turing: the higher the r_1 is the higher the probability of discovering a new species in further sampling; see Good (1953). Indeed, the choice (0·5, 1093·313) would correspond to the maximum likelihood choice for a basic sample less peaked in r_1. The maximum likelihood choice of (σ, θ) is to be rejected only if there is an expert opinion about the balancedness of the configuration of the enlarged sample $X^{(2,m)}$ that leads to a different pair (σ, θ).

We now consider the problem of estimating $D_m^{(2586:1825)}$, namely the probability of observing a new gene at the $(n + m + 1)$th draw, corresponding to sizes of the enlarged sample $m \in \{517, 1034, 1552, 2069, 2586\}$. This is the same set-up as in Mao (2004), where a moment-based estimator $\hat{U}_e$, which coincides with the Good–Toulmin estimator, and a likelihood-based estimator $\tilde{U}_e$ are considered. Table 3 illustrates the results provided by our Bayesian estimator $\hat{D}_m^{(2586:1825)}$, together with the results of Mao (2004).

We have also considered a measure of uncertainty about the evaluation of $D_m^{(2586:1825)}$. Let $[a, b]$ be a 95% highest posterior density interval for $(K_m^{(n)} \mid K_n = j)$ and note that the predictive probability $\mathrm{pr}\,(K_1^{(m+n)} = 1 \mid K_n = j, K_m^{(n)})$ is monotone increasing with respect to the number $K_m^{(n)}$ of new different species to be observed in $X^{(2,m)}$. Hence,

$$t_1 := \mathrm{pr}\left(K_1^{(m+n)} = 1 \,\middle|\, K_n = j, K_m^{(n)} = a\right) < \mathrm{pr}\left(K_1^{(m+n)} = 1 \,\middle|\, K_n, = j, K_m^{(n)} = b\right) =: t_2,$$

and we have $\mathrm{pr}\,(t_1 \leqslant D_m^{(n:j)} \leqslant t_2) \geqslant 0{\cdot}95$. Consequently, $[t_1,\ t_2]$ is the 95% highest posterior density interval for the probability of discovering a new species at the $(n + m + 1)$th draw, given that $K_n = j$ is the number of different species observed in the basic sample. These resulting intervals are reported in Table 3. Figure 4 displays the decay of the probability of discovering a new gene as m increases.

Note that the parameter choice of (0·5, 1093·313) for our estimator mimics very closely the frequentist estimates for the discovery probability, whereas the pair (0·612, 741) provides slightly larger estimates. As previously mentioned, unless an expert suggests a more balanced configuration of the m-sample, the Bayesian answer to the problem should be the one corresponding to (0·612, 741), thus predicting a somewhat higher discovery probability.

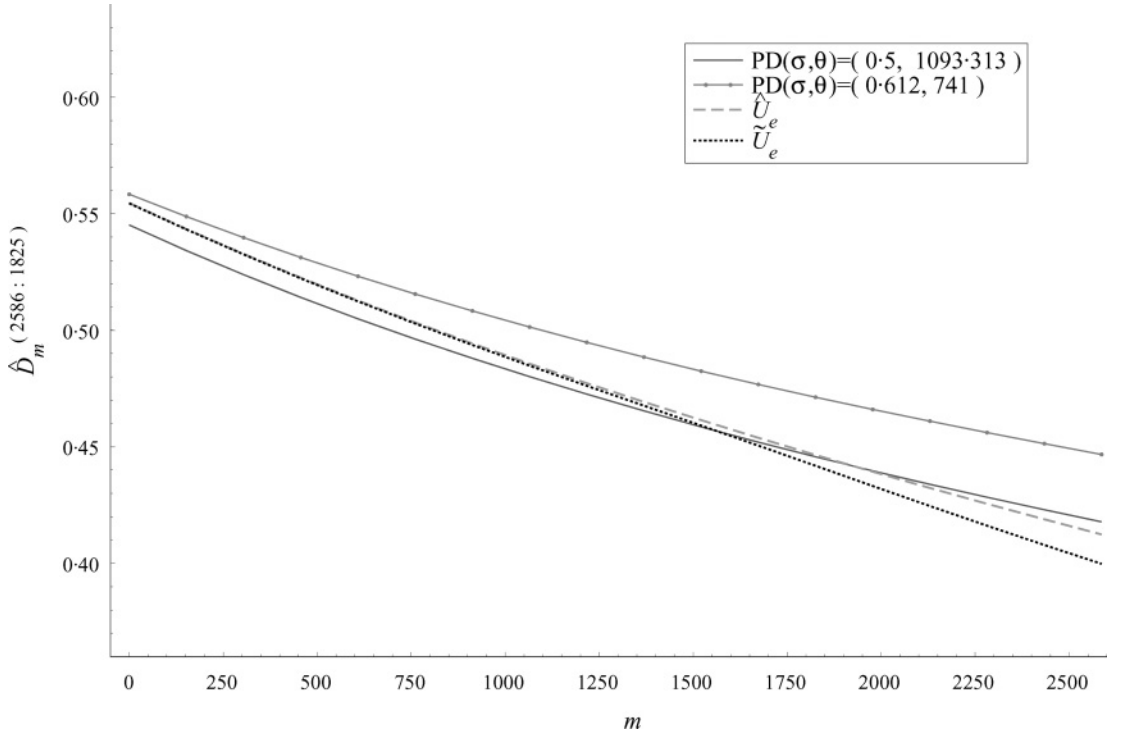

Fig. 4. Genomics example. Decay of the estimate $\hat{D}_m^{(2586:1825)}$ as m increases corresponding to the two choices of the Poisson–Dirichlet process, the moment-based estimator $\hat{U}_e$ and the likelihood-based estimator $\tilde{U}_e$.

Acknowledgement

The authors are grateful to the editor, an associate editor and two anonymous referees for their valuable comments. Antonio Lijoi is also affiliated to the Istituto di Matematica Applicata e Tecnologie Informatiche (C.N.R.), Milano, Italy. Igor Prünster is also affiliated to the Collegio Carlo Alberto and the International Center of Economic Research, Torino, Italy. The research of Antonio Lijoi and Igor Prünster was partially supported by grants of the Italian Ministry of University and Research. Ramsés Mena is grateful for the support of Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica, Universidad Nacional Autónoma de México, México.

Appendix

Technical details

Generalized factorial coefficients. The results in § 3 rely on the use of generalized factorial coefficients, both central and noncentral. Here we provide exact definitions for them and formulae for their evaluation; for further details and pointers to the literature, see Singh & Charalambides (1988) and Charalambides (2005). For any $n \geqslant 1$ and $k = 0, \ldots, n$, the generalized factorial coefficient $\mathscr{C}(n, k; \sigma)$ coincides with the coefficient of the kth-order factorial of t in the expansion of the nth-order generalized factorial of t with scale parameter σ, i.e.

$$(\sigma t)_n = \sum_{k=0}^{n} \mathscr{C}(n, k; \sigma)(t)_k.$$

In order to determine the distribution of the number K_n of different species appearing in a sample of size n, we use the representation

$$\mathscr{C}(n,k;\sigma)=\frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(-j\sigma)_n,$$

with the proviso that $\mathscr{C}(0,0;\sigma)=1$ and $\mathscr{C}(n,0;\sigma)=0$ for all $n\geqslant 1$. Note that $\mathscr{C}$ differs slightly from the definition of generalized factorial coefficient $C(n,k;\sigma)$ as given for example in Charalambides & Singh (1988) and Charalambides (2005). Indeed, $\mathscr{C}(n,k;\sigma)=(-1)^{n-k}C(n,k;\sigma)$. Moreover, for the special case $\sigma=1/2$, one has the simplification

$$\mathscr{C}(n,k,1/2)=2^{k-2n}\binom{2n-k-1}{n-1}\frac{\Gamma(n)}{\Gamma(k)}.$$

Along with $\mathscr{C}(n,k;\sigma)$ we consider the noncentral generalized factorial coefficient, $\mathscr{C}(n,k;\sigma,\gamma)$. It is defined as the coefficient of the kth-order factorial of t in the expansion of the nth-order noncentral generalized factorial of t, with scale parameter σ and noncentrality parameter γ, i.e.

$$(\sigma t-\gamma)_n=\sum_{k=0}^{n}\mathscr{C}(n,k;\sigma,\gamma)\,(t)_k.$$

From equation (2·60) in Charalambides (2005),

$$\mathscr{C}(n,k;\sigma,\gamma)=\frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(-\sigma j-\gamma)_n \tag{A1}$$

and this representation can be used to evaluate the probability of discovering a new species. Moreover, from equation (2·56) in Charalambides (2005) it is possible to relate noncentral and central generalized factorial coefficients, through

$$\mathscr{C}(n,k;\sigma,\gamma)=\sum_{s=k}^{n}\binom{n}{s}\mathscr{C}(s,k;\sigma)(-\gamma)_{n-s}. \tag{A2}$$

Finally we briefly recall the relationship to Stirling numbers, namely that

$$\lim_{\sigma\to 0}\frac{\mathscr{C}(n,k;\sigma)}{\sigma^k}=|s(n,k)|,$$

where, as before, $|s(n,k)|$ is the sign-less Stirling number of the first kind. Moreover, we have

$$\lim_{\sigma\to 0}\frac{\mathscr{C}(n,k;\sigma,\gamma)}{\sigma^k}=\sum_{i=k}^{n}\binom{n}{i}|s(i,k)|(-\gamma)_{n-i}.$$

Proof of Proposition 1. The proof will consist of two steps: first we derive a combinatorial result, which appears to be the key to the results of the present paper and then we exploit it in the context of Proposition 1. An alternative proof, based on Bayes' theorem, uses the prior distribution of K_n and the expression for $\mathrm{pr}\,(K_n=k|K_{n+m}=j)$ as given in Gnedin & Pitman (2005).

LEMMA A1. *For each* $\nu\geqslant 1$ *and* $j\geqslant 1$, *let* $A_{j,\nu}=\{(\nu_1,\ldots,\nu_j):\nu_i\geqslant 0,\sum_{i=1}^{j}\nu_i=\nu\}$. *Then*

$$\sum_{(\nu_1,\ldots,\nu_j)\in A_{j,\nu}}\binom{\nu}{\nu_1\cdots\nu_j}\prod_{i=1}^{j}(1-\sigma)_{n_i+\nu_i-1}=(n-j\sigma)_\nu\prod_{i=1}^{j}(1-\sigma)_{n_i-1},$$

where $(n_1, \ldots, n_k)$ *is such that* $n_i > 0$, *for* $i = 1, \ldots k$, *and* $\sum_{i=1}^{k} n_i = n$.

Proof of Lemma A1. First recall that

$$\sum_{(\nu_1,\ldots,\nu_j)\in A_{j,\nu}} \binom{\nu}{\nu_1 \cdots \nu_j} \prod_{i=1}^{j} (1-\sigma)_{n_i+\nu_i-1} = \sum_{(\nu_1,\ldots,\nu_j)\in A_{j,\nu}} \binom{\nu}{\nu_1 \cdots \nu_j} \frac{\prod_{i=1}^{j} \Gamma(n_i+\nu_i-\sigma)}{\Gamma(1-\sigma)^j}.$$

Rewrite $\prod_{i=1}^{j} \Gamma(n_i + \nu_i - \sigma)$ as a multiple gamma integral and exploit the multinomial formula, obtaining

$$\frac{1}{\{\Gamma(1-\sigma)\}^j} \sum_{(\nu_1,\ldots,\nu_j)\in A_{j,\nu}} \binom{\nu}{\nu_1 \cdots \nu_j} \int_{(\mathbb{R}^+)^j} e^{-\sum_{i=1}^{j} u_i} \left\{ \prod_{i=1}^{j} u_i^{n_i+\nu_i-\sigma-1} \right\} du_1 \cdots du_j$$

$$= \frac{1}{\{\Gamma(1-\sigma)\}^j} \int_{(\mathbb{R}^+)^j} e^{-\sum_{i=1}^{j} u_i} (u_1 + \cdots + u_j)^{\nu} \left\{ \prod_{i=1}^{j} u_i^{n_i-\sigma-1} \right\} du_1 \cdots du_j.$$

We change the variables to $y_i = u_i$ for $i \in \{1, \ldots, j-1\}$ and $y_j = \sum_{i=1}^{j} u_i$ to obtain

$$\frac{1}{\{\Gamma(1-\sigma)\}^j} \int_0^{+\infty} e^{-y_j} y_j^{\nu} \left\{ \int_{B(y_j)} y_1^{n_1-\sigma-1} \cdots y_{j-1}^{n_{j-1}-\sigma-1} \left(\sum_{i=1}^{j} y_i \right)^{n_j-\sigma-1} dy_1 \ldots dy_{j-1} \right\} dy_j,$$

where $B(y_j) = \{(y_1, \ldots, y_{j-1}) : y_i \geqslant 0, \sum_{i=1}^{j-1} y_i \leqslant y_j\}$. A further change of variables to $(z_1, \ldots, z_{j-1}, z_j) = (y_1/y_j, \ldots, y_{j-1}/y_j, y_j)$ yields

$$\frac{1}{\{\Gamma(1-\sigma)\}^j} \int_0^{+\infty} e^{-z_j} z_j^{\nu+n-j\sigma}$$

$$\times \left\{ \int_{\Delta_{j-1}} z_1^{n_1-\sigma-1} \cdots z_{j-1}^{n_{j-1}-\sigma-1} \left(1 - \sum_{i=1}^{j-1} z_i \right)^{n_j-\sigma-1} dz_1 \ldots dz_{j-1} \right\} dz_j,$$

where $\Delta_{j-1} := \{(z_1, \ldots, z_{j-1}) : z_i \geqslant 0, \sum_{i=1}^{j-1} z_i \leqslant 1\}$ is the $(j-1)$-dimensional simplex. Then the above integral reduces to

$$\frac{\prod_{i=1}^{j} \Gamma(n_i - \sigma)}{\{\Gamma(1-\sigma)\}^j \Gamma(n-j\sigma)} \int_0^{+\infty} e^{-z_j} z_j^{\nu+n-j\sigma-1} \, dz_j = \left\{ \frac{\prod_{i=1}^{j} \Gamma(n_i-\sigma)}{(\Gamma(1-\sigma))^j} \right\} (n-j\sigma)_{\nu}$$

and the result follows. The lemma can also be proved by induction from the classical Chu–Vandermonde identity, but the direct proof outlined above seems neater. □

We can now proceed with the proof of Proposition 1. In order to determine the conditional distribution of $K_m^{(n)}$ given a sample $X_j^{(n)}$, we make use of the exchangeable partition probability function as follows:

$$\text{pr}\left(K_m^{(n)} = k \,\middle|\, X_j^{(1,n)} \right)$$

$$= \frac{\sum_{\pi\in\mathscr{P}_{m,j+k}} \Pi_{j+k}^{(n+m)}\{n_1 + m_1(\pi), \ldots, n_j + m_j(\pi), m_{j+1}(\pi), \ldots, m_{j+k}(\pi)\}}{\Pi_j^{(n)}(n_1, \ldots, n_j)}, \tag{A3}$$

where $\mathscr{P}_{m,j+k}$ denotes the set of all partitions of m observations into $q \leqslant m$ classes, with $q \in \{k, \ldots, k+j\}$. Of the q classes into which the observations $X_{n+1}, \ldots, X_{n+m}$ are partitioned, k are new and $q - k \leqslant j$ coincide with some of those already observed in the conditioning sample $X_j^{(1,n)}$. If we make use of the

hypothesized Gibbs structure (2), the numerator in (A3) becomes

$$\sum_{\pi\in\mathscr{P}_{m,j+k}} \Pi_{j+k}^{(n+m)}\{n_1+m_1(\pi),\ldots,n_j+m_j(\pi),m_{j+1}(\pi),\ldots,m_{j+k}(\pi)\}$$
$$= V_{n+m,j+k}\sum_{\pi\in\mathscr{P}_{m,j+k}}\prod_{i=1}^{j}(1-\sigma)_{n_i+m_i(\pi)-1}\prod_{r=1}^{k}(1-\sigma)_{m_{j+r}(\pi)-1}.$$

In order to evaluate this sum, we split the new m observations into two groups: s of them will generate the new k classes and the remaining $m-s$ will be spread among the j old groups of distinct observations. Hence, considering just the sum, we have

$$\sum_{\pi\in\mathscr{P}_{m,j+k}}\prod_{i=1}^{j}(1-\sigma)_{n_i+m_i(\pi)-1}\prod_{r=1}^{k}(1-\sigma)_{m_{j+r}(\pi)-1}$$
$$=\sum_{s=k}^{m}\binom{m}{s}\left\{\sum_{(**)}\binom{m-s}{m_1\cdots m_j}\prod_{i=1}^{j}(1-\sigma)_{n_i+m_i-1}\right\}\left\{\frac{1}{k!}\sum_{(***)}\binom{s}{m_{j+1}\cdots m_{j+k}}\prod_{r=1}^{k}(1-\sigma)_{m_{j+r}-1}\right\},$$

where $(**)$ means that the sum runs through the set of nonnegative integers

$$\left\{(m_1,\ldots,m_j):\ m_i\geqslant 0 \text{ for } i=1,\ldots,j,\sum_{i=1}^{j}m_i=m-s\right\},$$

and $(***)=\{(m_{j+1},\ldots,m_{j+k}):\ m_{j+i}\geqslant 1 \text{ for } i=1,\ldots,k,\ \sum_{i=1}^{k}m_{j+i}=s\}$. Considering the second factor within the sum, as has been already observed, we have

$$\frac{1}{k!}\sum_{(***)}\binom{s}{m_{j+1}\cdots m_{j+k}}\prod_{r=1}^{k}(1-\sigma)_{m_{j+r}-1}=\frac{\mathscr{C}(s,k,\sigma)}{\sigma^k}.$$

This, combined with a straightforward application of Lemma A1 with $\nu=m-s$, gives

$$\sum_{\pi\in\mathscr{P}_{m,j+k}}\prod_{i=1}^{j}(1-\sigma)_{n_i+m_i(\pi)-1}\prod_{r=1}^{k}(1-\sigma)_{m_{j+r}(\pi)-1}$$
$$=\frac{\prod_{i=1}^{j}(1-\sigma)_{n_i-1}}{\sigma^k}\sum_{s=k}^{m}\binom{m}{s}\mathscr{C}(s,k,\sigma)(n-j\sigma)_{m-s},$$

and the last sum is seen to be a noncentral generalized factorial coefficient because of (A2). Now, if we specify the form of the denominator in (A3) according to (2), the expression in (4) follows. Finally note that, when $k=0$, the expression in (4) reduces to

$$\operatorname{pr}\left(K_m^{(n)}=j\,\middle|\,X_j^{(1,n)}\right)=\frac{V_{n+m,j}}{V_{n,j}}(n-j\sigma)_m$$

because of the definition of the generalized factorial coefficient. □

Proof of Proposition 2. In order to obtain the estimator, one can make use of (4) in Proposition 1. First note that the Bayes estimate of $D_m^{(n:j)}=\operatorname{pr}(K_1^{(n+m)}=1\mid X_j^{(n)},\ K_m^{(n)})$, with respect to a squared loss

function, is given by its expected value

$$\hat{D}_m^{(n:j)} = \sum_{k=0}^{m} \mathrm{pr}\left(K_1^{(n+m)} = 1 \,\middle|\, X_j^{(n)}, K_m^{(n)} = k\right) \mathrm{pr}\left(K_m^{(n)} = k \,\middle|\, K_n = j\right).$$

The second factor in each summand above is determined via Proposition 1, whereas

$$\mathrm{pr}\left(K_1^{(n+m)} = 1 \,\middle|\, X_j^{(n)}, K_m^{(n)} = k\right)$$

is just the one-step prediction, since it coincides with the probability of drawing a new species given that the first $(n + m)$ individuals observed come from $(j + k)$ distinct species. Hence

$$\mathrm{pr}\left(K_1^{(n+m)} = 1 \,\middle|\, X_j^{(n)}, K_m^{(n)} = k\right) = \frac{V_{n+m+1,\,j+k+1}}{V_{n+m,\,j+k}}$$

and the result is proved. □

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–74.

Berestycki, N. & Pitman, J. (2007). Gibbs distributions for random partitions generated by a fragmentation process. *J. Statist. Phys.* **127**, 381–418.

Boender, C. G. E. & Rinnooy Kan, A. H. G. (1987). A multinomial Bayesian approach to the estimation of population and vocabulary size. *Biometrika* **74**, 849–56.

Boneh, S., Boneh, A. & Caron, R. J. (1998). Estimating the prediction function of the number of unseen species in sampling with replacement. *J. Am. Statist. Assoc.* **93**, 372–9.

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

Chao, A. (1981). On estimating the probability of discovering a new species. *Ann. Statist.* **9**, 1339–42.

Chao, A. & Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics* **58**, 531–9.

Chao, A. & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Am. Statist. Assoc.* **87**, 210–7.

Chao, A. & Shen, T.-J. (2004). Nonparametric prediction in species sampling. *J. Agric. Biol. Envir. Statist.* **9**, 253–69.

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

Charalambides, C. A. & Singh, J. (1988). A review of the Stirling numbers, their generalisations and statistical applications. *Commun. Statist.* A **17**, 2533–95.

Christen, J. A. & Nakamura, M. (2003). Sequential stopping rules for species accumulation. *J. Agric. Biol. Envir. Statist.* **8**, 184–95.

Clayton, M. K. & Frees, E. W. (1987). Nonparametric estimation of the probability of discovering a new species. *J. Am. Statist. Assoc.* **82**, 305–11.

Do, K.-A., Müller, P. & Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Appl. Statist.* **54**, 627–44.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Prob.* **2**, 183–201.

Efron, B. & Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* **63**, 435–47.

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

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–30.

Gandolfi, A. & Sastri, C. C. A. (2004). Nonparametric estimations about species not observed in a random sample. *Milan J. Math.* **72**, 81–105.

Gnedin, A. & Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. POMI* **325**, 83–102.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–64.

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

Hill, B. M. (1979). Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *J. Am. Statist. Assoc.* **74**, 668–73.

Lijoi, A., Mena, R. H. & Prünster, I. (2005). Hierarchical mixture modeling with normalised inverse Gaussian priors. *J. Am. Statist. Assoc.* **100**, 1278–91.

Mao, C. X. (2004). Prediction of the conditional probability of discovering a new class. *J. Am. Statist. Assoc.* **99**, 1108–18.

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

MÜLLER, P. & QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19**, 95–110.

PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Prob. Theory Rel. Fields* **102**, 145–58.

PITMAN, J. (1996). Some developments of the Blackwell-MacQueen urn Scheme. In *Statistics, Probability and Game Theory. Papers in honor of David Blackwell* (Eds. Ferguson, T. S. et al.). Lecture Notes, Monograph Series, **30**, 245–67. Institute of Mathematical Statistics, Hayward.

PITMAN, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002. Lecture Notes in Mathematics N^o 1875. New York: Springer.

QUINTANA F. A. (2006). A predictive view of Bayesian clustering. *J. Statist. Plan. Infer.* **136**, 2407–29.

REGAZZINI, E., LIJOI, A. & PRÜNSTER, I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560–85.

ROBBINS, H. E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Statist.* **39**, 256–7.

STARR, N. (1979). Linear estimation of the probability of discovering a new species. *Ann. Statist.* **7**, 644–52.

SHEN, T.-J., CHAO, A. & LIN, C.-F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology* **84**, 798–804.

TIWARI, R. C. & TRIPATHI, R. C. (1989). Nonparametric Bayes estimation of the probability of discovering a new species. *Commun. Statist.* A **18**, 877–95.

ZHANG, H. & STERN, H. (2005). Investigation of a generalised multinomial model for species data. *J. Statist. Comp. Simul.* **75**, 347–62.

[*Received June* 2006. *Revised February* 2007]