SUPPLEMENTARY MATERIAL FOR "BAYESIAN NONPARAMETRIC INFERENCE FOR DISCOVERY PROBABILITIES: CREDIBLE INTERVALS AND LARGE SAMPLE ASYMPTOTICS"

Julyan Arbel[1,2], Stefano Favaro[1,3], Bernardo Nipoti[4] and Yee Whye Teh[5]

julyan.arbel@unibocconi.it, stefano.favaro@unito.it

nipotib@tcd.ie, y.w.teh@stats.ox.ac.uk

[1] *Collegio Carlo Alberto, Moncalieri, Italy,* [2] *Department of Decision Sciences, BIDSA and IGIER, Bocconi University, Milan, Italy*

[3] *Department of Economics and Statistics, University of Torino, Italy*

[4] *School of Computer Science and Statistics, Trinity College Dublin, Ireland*

[5] *Department of Statistics, University of Oxford, United Kingdom*

This supplementary material contains: i) the proofs of Theorem 1, Proposition 1, Proposition 2, Theorem 2, Proposition 3 and Proposition 4; ii) details on the derivation of the asymptotic equivalence between $\hat{\mathcal{D}}_n(l)$ and $\check{\mathcal{D}}_n(l; \mathscr{S}_{\mathrm{PD}})$; iii) additional application results.

Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ be a sample from a Gibbs-type RPM Q_h. Recall that, due to the discreteness of Q_h, the sample $\boldsymbol{X}_n$ features $K_n = k_n$ species, labelled by $X_1^*, \ldots, X_{K_n}^*$, with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k_n,n})$. Furthermore, let $M_{l,n} = m_{l,n}$ be the number of species with frequency l, namely $M_{l,n} = \sum_{1\leq i\leq K_n} \mathbb{1}_{\{N_{i,n}=l\}}$ such that $\sum_{1\leq i\leq n} M_{i,n} = K_n$ and $\sum_{1\leq i\leq n} iM_{i,n} = n$. For any $\sigma \in (0,1)$ let f_σ be the density function of a positive σ-stable random variable. According to Proposition 13 in Pitman (2003), as $n \to +\infty$

$$\frac{K_n}{n^\sigma} \xrightarrow{\text{a.s.}} S_{\sigma,h} \tag{S0.1}$$

and

$$\frac{M_{l,n}}{n^\sigma} \xrightarrow{\text{a.s.}} \frac{\sigma(1-\sigma)_{l-1}}{l!} S_{\sigma,h}, \tag{S0.2}$$

where $S_{\sigma,h}$ is a random variable with density function $f_{S_{\sigma,h}}(s) = \sigma^{-1} s^{-1/\sigma-1} h(s^{-1/\sigma}) f_\sigma(s^{-1/\sigma})$. Note that by the fluctuation limits displayed in (S0.1) and (S0.2), as n tends to infinity the number of species with frequency l in a sample of size n from Q_h becomes, almost surely, a proportion $\sigma(1-\sigma)_{l-1}/l!$ of the total number of species in the sample. All the random variables introduced in this web appendix are meant to be assigned on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

S1 Proofs

PROOF OF THEOREM 1. We proceed by induction. Note that the result holds for $r = 1$, and obviously for any sample size $n \geq 1$. Let us assume that it holds for a given $r \geq 1$, and also for any sample size $n \geq 1$. Then, the $(r+1)$-th moment of $Q_h(A) \mid \boldsymbol{X}_n$ can be written as follows

$$
\begin{aligned}
&\mathbb{E}[Q_h^r(A) \mid \boldsymbol{X}_n] \\
&\quad = \int_A \cdots \int_A \mathbb{P}[X_{n+r+1} \in A \mid \boldsymbol{X}_n, X_{n+1} = x_{n+1}, \ldots, X_{n+r} = x_{n+r}] \\
&\qquad \times \mathbb{P}[X_{n+r} \in \mathrm{d}x_{n+r} \mid \boldsymbol{X}_n, X_{n+1} = x_{n+1}, \ldots, X_{n+r-1} = x_{n+r-1}] \\
&\qquad\quad \times \cdots \times \mathbb{P}[X_{n+2} \in \mathrm{d}x_{n+2} \mid \boldsymbol{X}_n, X_{n+1} = x_{n+1}]\mathbb{P}[X_{n+1} \in \mathrm{d}x_{n+1} \mid \boldsymbol{X}_n] \\
&\quad = \int_A \mathbb{E}[Q_h^r(A) \mid \boldsymbol{X}_n, X_{n+1} = x_{n+1}] \\
&\qquad \times \left(\frac{V_{h,(n+1,k_n+1)}}{V_{h,(n,k_n)}} \nu_0(\mathrm{d}x_{n+1}) + \frac{V_{h,(n+1,k_n)}}{V_{h,(n,k_n)}} \sum_{i=1}^{k_n} (n_i - \sigma)\delta_{X_i^*}(\mathrm{d}x_{n+1}) \right).
\end{aligned}
$$

Further, by the assumption on the r-th moment and by dividing A into $(A \setminus \boldsymbol{X}_n) \cup (A \cap \boldsymbol{X}_n)$, one obtains

$$
\begin{aligned}
&\mathbb{E}[Q_h^{r+1}(A) \mid \boldsymbol{X}_n] \\
&\quad = \sum_{i=0}^{r} \frac{V_{n+r+1,k_n+r+1-i}}{V_{h,(n,k_n)}} [\nu_0(A)]^{r+1-i} R_{r,i}(\mu_{n,k_n}(A) + 1 - \sigma) \\
&\qquad + \sum_{i=1}^{r+1} \frac{V_{n+r+1,k_n+r+1-i}}{V_{h,(n,k_n)}} [\nu_0(A)]^{r+1-i} \mu_{n,k_n}(A) R_{r,i-1}(\mu_{n,k_n}(A) + 1),
\end{aligned}
$$

where we defined $R_{r,i}(\mu) := \sum_{0 \leq j_1 \leq \cdots \leq j_i \leq r-i} \prod_{1 \leq l \leq i} (\mu + j_l(1-\sigma) + l - 1)$. The proof is completed by noting that, by means of simple algebraic manipulations, $R_{r+1,i}(\mu) = R_{r,i}(\mu + 1 - \sigma) + \mu R_{r,i-1}(\mu + 1)$. Note that when $\nu_0(A) = 0$ and $i = r$, the convention $\nu_0(A)^{r-i} = 0^0 = 1$ is adopted. □

PROOF OF PROPOSITION 1. Let us consider the Borel sets $A_0 := \mathbb{X} \setminus \{X_1^*, \ldots, X_{K_n}^*\}$ and $A_l := \{X_i^* : N_{i,n} = l\}$, for any $l = 1, \ldots, n$. The two parameter PD prior is a Gibbs-type prior with $h(t) = p(t; \sigma, \theta) := \sigma\Gamma(\theta)t^{-\theta}/\Gamma(\theta/\sigma)$, for any $\sigma \in (0,1)$ and $\theta > -\sigma$. Therefore one has $V_{n,k_n} = V_{p,(n,k_n)} = [(\theta)_n]^{-1} \prod_{0 \leq i \leq k_n - 1} (\theta + i\sigma)$. By a direct application of Theorem 1 we can write

$$
\begin{aligned}
\mathbb{E}[Q_h^r(A_0) \mid \boldsymbol{X}_n] &= \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{(\theta)_n}{(\theta)_{n+i}} (n - \sigma k_n)_i \\
&= (\theta)_n \frac{(\theta + \sigma k_n)_r}{(\theta)_n (\theta + n)_r} \\
&= \frac{(\theta + \sigma k_n)_r}{(\theta + \sigma k_n + n - \sigma k_n)_r},
\end{aligned}
$$

which is r-th moment of a Beta random variable with parameter $(\theta+\sigma k, n-\sigma k)$. Let us define the random variable $Y = Z_p R_{\sigma, Z_p}$. Then, it can be easily verified that Y has density function

$$
\begin{aligned}
f_Y(y) &= \int_0^\infty \frac{1}{z} f_{R_{\sigma,z}}(y/z) f_{Z_p}(z) \mathrm{d}z \\
&= \frac{\sigma}{\Gamma(\theta/\sigma + k_n)} \int_0^\infty \mathrm{e}^{z^\sigma - y - z^\sigma} z^{\theta+\sigma k_n - 2} f_\sigma(y/z) \mathrm{d}z \\
&= \frac{\sigma}{\Gamma(\theta/\sigma + k_n)} y^{\theta+\sigma k_n - 1} \mathrm{e}^{-y} \int_0^\infty u^{-(\theta+\sigma k_n)} f_\sigma(u) \mathrm{d}u
\end{aligned}
$$

where, by Equation 60 in Pitman (2003), $\int_0^\infty u^{-(\theta+\sigma k_n)} f_\sigma(u)\mathrm{d}u = \Gamma(\theta/\sigma + k_n)/\sigma\Gamma(\theta+\sigma k_n)$. Hence Y is a Gamma random variable with parameter $(\theta+\sigma k_n, 1)$. Accordingly, we have $W_{n-\sigma k_n, Z_p} \stackrel{\mathrm{d}}{=} B_{\theta+\sigma k_n, n-\sigma k_n}$. Similarly, by a direct application of Theorem 1, for any $l > 1$ we can write

$$
\begin{aligned}
\mathbb{E}[Q_h^r(A_l) \,|\, \boldsymbol{X}_n] &= \frac{(\theta)_n}{(\theta)_{n+r}} ((l-\sigma)m_{l,n})_r \\
&= \frac{((l-\sigma)m_{l,n})_r}{((l-\sigma)m_{l,n})_r + \theta + n - (l-\sigma)m_{l,n}},
\end{aligned}
$$

which is the r-th moment of a Beta random variable with parameter $((l-\sigma)m_{l,n}, \theta+n-(l-\sigma)m_{l,n})$. Finally, the decomposition $B_{(l-\sigma)m_{l,n}, \theta+n-(l-\sigma)m_{l,n}} \stackrel{\mathrm{d}}{=} B_{(l-\sigma)m_{l,n}, n-\sigma k_n-(l-\sigma)m_{l,n}}(1-W_{n-\sigma k_n, Z_p})$ follows from a characterization of Beta random variables in Theorem 1 in Jambunathan (1954). It can be also easily verified by using the moments of Beta random variables. □

Proof of Proposition 2. Let us consider the Borel sets $A_0 := \mathbb{X} \setminus \{X_1^*, \ldots, X_{K_n}^*\}$ and $A_l := \{X_i^* : N_{i,n} = l\}$, for any $l = 1, \ldots, n$. The two parameter PD prior is a Gibbs-type prior with $h(t) = g(t; \sigma, \tau) := \exp\{\tau^\sigma - \tau t\}$, for any $\tau > 0$. By a direct application of Theorem 1 we can write

$$
\begin{aligned}
&\mathbb{E}[Q_g^r(A_0) \,|\, \boldsymbol{X}_n] \qquad \text{(S1.1)}\\
&\quad = \frac{\sigma\Gamma(n)}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k_n)} \int_0^1 w^r (1-w)^{n-1-\sigma k_n} \int_0^{+\infty} t^{-\sigma k_n} \mathrm{e}^{-\tau t} f_\sigma(wt) \mathrm{d}t\mathrm{d}w,
\end{aligned}
$$

where

$$
\begin{aligned}
C_{\sigma,\tau,n,k_n} &:= \frac{\sigma\Gamma(n)}{\Gamma(n-\sigma k_n)} \int_0^{+\infty} t^{-\sigma k_n} \mathrm{e}^{-\tau t} \int_0^1 (1-w)^{n-1-\sigma k_n} f_\sigma(wt) \mathrm{d}w\mathrm{d}t \\
&= \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau)^i \Gamma(k - i/\sigma; \tau^\sigma).
\end{aligned}
$$

Hereafter we show that (S1.1) coincides with the r-th moment of the random variable $W_{n-\sigma k_n, Z_g}$. Given $Z_g = z$ it is easy to find that the distribution of $W_{n-\sigma k_n, z}$ has the following density function

$$
f_{W_{n-\sigma k_n, z}}(w) = \frac{\exp\{z^\sigma\}}{z\Gamma(n-k_n\sigma)} (1-w)^{n-k_n\sigma-1} \int_0^{+\infty} u^{n-k_n\sigma} \mathrm{e}^{-u} f_\sigma\left(\frac{uw}{z}\right) \mathrm{d}u.
$$

By randomizing over z with respect to the distribution of Z_g provides the distribution of $W_{n-\sigma k_n, Z_g}$. Specifically,

$$
\begin{aligned}
f_{W_{n-\sigma k_n, Z_g}}(w) &= \frac{\sigma}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k_n)}(1-w)^{n-\sigma k_n -1} \\
&\quad \times \int_\tau^\infty z^{-n+\sigma k_n -1}(z-\tau)^{n-1}\int_0^\infty u^{n-\sigma k_n}\mathrm{e}^{-u} f_\sigma\left(\frac{uw}{z}\right)\mathrm{d}u\mathrm{d}z \\
&= \frac{\sigma}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k)}(1-w)^{n-\sigma k_n -1} \\
&\quad \times \int_\tau^\infty (z-\tau)^{n-1}\int_0^\infty t^{n-\sigma k_n}\mathrm{e}^{-tz} f_\sigma(wt)\,\mathrm{d}t\mathrm{d}z \\
&= \frac{\sigma\Gamma(n)}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k_n)}(1-w)^{n-\sigma k_n -1}\int_0^\infty t^{-\sigma k_n}\mathrm{e}^{-\tau t} f_\sigma(wt)\,\mathrm{d}t.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}[W^r_{n-\sigma k_n, Z_g}] \\
&\quad = \frac{\sigma\Gamma(n)}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k_n)}\int_0^1 w^r(1-w)^{n-\sigma k_n -1}\int_0^\infty t^{-\sigma k_n}\mathrm{e}^{-\tau t} f_\sigma(wt)\,\mathrm{d}t\mathrm{d}w
\end{aligned}
$$

which coincides with (S1.1). We complete the proof by determining the distribution of the random variable $Q_g(A_l)\,|\,\boldsymbol{X}_n$, for any $l > 1$. Again, by a direct application of Theorem 1 we can write

$$
\begin{aligned}
&\mathbb{E}[Q^r_g(A_l)\,|\,\boldsymbol{X}_n] \\
&= ((l-\sigma)m_{l,n})_r \frac{\frac{\sigma^{k_n}}{\Gamma(n-\sigma k_n + r)}}{\frac{\sigma^{k_n}}{\Gamma(n-\sigma k_n)}} \frac{\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 (1-z)^{n+r-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}{\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 (1-z)^{n-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z} \\
&= \frac{\Gamma(n-\sigma k_n)}{\Gamma((l-\sigma)m_{l,n})\,\Gamma(\sum_{1\le i\ne l\le n} i m_{i,n} - \sigma\sum_{1\le i\ne l\le n} m_{i,n})} \\
&\quad \times \int_0^1 x^{(l-\sigma)m_{l,n}+r-1}(1-x)^{\sum_{1\le i\ne l\le n} i m_{i,n} - \sigma\sum_{1\le i\ne l\le n} m_{i,n} -1} \\
&\quad\quad \times \frac{\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 (1-z)^{n+r-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}{\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 (1-z)^{n-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}\mathrm{d}x \\
&= \frac{\Gamma(n-\sigma k_n)}{\Gamma((l-\sigma)m_{l,n})\,\Gamma(\sum_{1\le i\ne l\le n} i m_{i,n} - \sigma\sum_{1\le i\ne l\le n} m_{i,n})} \\
&\quad \times \int_0^1 x^{(l-\sigma)m_{l,n}-1}(1-x)^{\sum_{1\le i\ne l\le n} i m_{i,n} - \sigma\sum_{1\le i\ne l\le n} m_{i,n} -1} \\
&\quad\quad \times \frac{\frac{\sigma\Gamma(n)}{\Gamma(n-\sigma k_n)}\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 x^r(1-z)^r(1-z)^{n-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}{\frac{\sigma^{k_n}}{\Gamma(n-\sigma k_n)}\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 (1-z)^{n-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}\mathrm{d}x,
\end{aligned}
$$

which is the r-th moment of the scale mixture $B_{(l-\sigma)m_{l,n}, n-\sigma k_n -(l-\sigma)m_{l,n}}(1-W_{n-\sigma k_n, Z_g})$, where $W_{n-\sigma k_n, Z_g}$ is the random variable characterized above, and where the Beta random variable $B_{(l-\sigma)m_{l,n}, n-\sigma k_n -(l-\sigma)m_{l,n}}$ is independent of the random variable $(1-W_{n-\sigma k_n, Z_g})$. The proof is completed. □

PROOF OF THEOREM 2. According to the fluctuation limit (S0.1) there exists a nonnegative and finite random variable $S_{\sigma,h}$ such that $n^{-\sigma}K_n \xrightarrow{\text{a.s.}} S_{\sigma,h}$ as $n \to +\infty$. Let $\Omega_0 := \{\omega \in \Omega : \lim_{n\to+\infty} n^{-\sigma}K_n(w) = S_{\sigma,h}(\omega)\}$. Furthermore, let us define $g_{0,h}(n,k_n) = V_{h,(n+1,k_n+1)}/V_{h,(n,k_n)}$, where $V_{h,(n,k_n)} = \sigma^{k_n-1}\Gamma(k_n)\mathbb{E}[h(S_{\sigma,k_n}/B_{\sigma k_n,n-\sigma k_n})]/\Gamma(n)$. Then we can write the following expression

$$g_{0,h}(n,k_n) = \frac{\sigma k_n}{n}\frac{\mathbb{E}\left[h\left(\frac{S_{\sigma,k_n+1}}{B_{\sigma k_n+1,n+1-\sigma(k_n+1)}}\right)\right]}{\mathbb{E}\left[h\left(\frac{S_{\sigma,k_n}}{B_{\sigma k_n,n-\sigma k_n}}\right)\right]}. \tag{S1.2}$$

We have to show that the ratio of the expectations in (S1.2) converges to 1 as $n \to +\infty$. For this, it is sufficient to show that, as $n \to +\infty$, the random variable $T_{\sigma,n,k_n} = S_{\sigma,k_n}/B_{\sigma k_n,n-\sigma k_n}$ converges almost surely to a random variable $T_{\sigma,h}$. This is shown by computing the moment of order r of T_{σ,n,k_n}, i.e.,

$$\mathbb{E}(T^r_{\sigma,n,k_n}) = \frac{\Gamma(n)}{\Gamma(n-r)}\frac{\Gamma(k_n - r/\sigma)}{\Gamma(k_n)} \simeq \frac{n^r}{k_n^{r/\sigma}}.$$

For any $\omega \in \Omega_0$ the ratio $n/K_n^{1/\sigma}(\omega) = n/k_n^{1/\sigma}$ converges to $S^{-1/\sigma}_{\sigma,h}(\omega) = T_{\sigma,h}(\omega) = t$. Accordingly, $n^r/k_n^{r/\sigma}$ converges to $\mathbb{E}[T^r_\sigma(\omega)] = t^r$ for any $\omega \in \Omega_0$. Since $\mathbb{P}[\Omega_0] = 1$, the almost sure limit, as n tends to infinity, of the random variable T_{σ,n,K_n} is identified with the nonnegative random variable $T_{\sigma,h}$, which has density function $f_{T_{\sigma,h}}(t) = h(t)f_\sigma(t)$. The proof is completed.

PROOF OF PROPOSITION 3. Let $h(t) = p(t;\sigma,\theta) := \sigma\Gamma(\theta)t^{-\theta}/\Gamma(\theta/\sigma)$, for any $\sigma \in (0,1)$ and $\theta > -\sigma$. Furthermore, let us define $g_{0,p}(n,k_n) = V_{p,(n+1,k_n+1)}/V_{p,(n,k_n)}$ and $g_{1,p}(n,k_n) = 1 - V_{p,(n+1,k_n+1)}/V_{p,(n,k_n)}$, so that we have $g_0(n,k_n) = (\theta+\sigma k_n)/(\theta+n)$ and $g_1(n,k_n) = 1/(\theta+n)$. Then,

$$g_{0,p}(n,k_n) = \frac{\sigma k_n}{n} + \frac{\theta}{n} + o\left(\frac{1}{n}\right) \tag{S1.3}$$

and

$$g_{1,p}(n,k_n) = \frac{1}{n} - \frac{\theta}{n^2} + o\left(\frac{1}{n^2}\right) \tag{S1.4}$$

follow by a direct application of the Taylor series expansion to $g_0(n,k_n)$ and $g_1(n,k_n)$, respectively, and then truncating the series at the second order. The proof is completed by combining (S1.3) and (S1.4) with the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ under a two parameter PD prior. □

PROOF OF PROPOSITION 4. The proof is along lines similar to the proof of Proposition 3.2. in Ruggiero et al. (2015), which, however, considers a different parameterization for the normalized GG prior. Let $h(t) = g(t;\sigma,\tau) := \exp\{\tau^\sigma - \tau t\}$, for any $\sigma \in (0,1)$ and $\tau > 0$, and let $g_{0,g}(n,k_n) = V_{g,(n+1,k_n+1)}/V_{g,(n,k_n)}$ and $g_{1,p}(n,k_n) = 1 - V_{g,(n+1,k_n+1)}/V_{g,(n,k_n)}$, where we have

$$V_{g,(n,k_n)} = \frac{\sigma^{k_n}\exp\{\tau^\sigma\}}{\Gamma(n)}\int_0^{+\infty} x^{n-1}(\tau+x)^{-n+\sigma k_n}\mathrm{e}^{-(\tau+x)^\sigma}\mathrm{d}x.$$

Note that, by using the triangular relation characterizing the nonnegative weight $V_{g,(n,k_n)}$, we can write

$$g_{0,g}(n,k_n) = \frac{V_{g,(n,k_n)} - (n-\sigma k_n)V_{g,(n+1,k_n)}}{V_{g,(n,k_n)}} = 1 - \left(1 - \frac{\sigma k_n}{n}\right) w(n,k_n),$$

where

$$w(n,k_n) = \frac{\int_0^\infty x^n \exp\{-[(\tau+x)^\sigma - \tau^\sigma]\}(\tau+x)^{\sigma k_n - n - 1}\,\mathrm{d}x}{\int_0^\infty x^{n-1} \exp\{-[(\tau+x)^\sigma - \tau^\sigma]\}(\tau+x)^{\sigma k_n - n}\,\mathrm{d}x}.$$

Let us denote by $f(x)$ the integrand function of the denominator of $1 - w(n,k_n)$, and let $f_N(x) = \tau f(x)/(\tau + x)$. That is, $f_N(x)$ is the denominator of $1 - w(n,k_n)$. Therefore we can write

$$1 - w(n,k_n) = \frac{\int_0^\infty \tau f(x)/(\tau+x)\,\mathrm{d}x}{\int_0^\infty f(x)\,\mathrm{d}x}.$$

Since $f(x)$ is unimodal, by means of the Laplace approximation method it can be approximated with a Gaussian kernel with mean $x^* = \arg\max_{x>0} x^{n-1}\exp\{-[(\tau+x)^\sigma - \tau^\sigma]\}(\tau+x)^{\sigma k_n - n}$ and with variance $-[(\log\circ f)''(x^*)]^{-1}$. The same holds for $f_N(x)$. Then, we obtain the approximation

$$1 - w(n,k_n) \simeq \frac{f_N(x_N^*)C(x_N^*, -[(\log\circ f_N)''(x_N^*)]^{-1})}{f(x_D^*)C(x_D^*, -[(\log\circ f)''(x_D^*)]^{-1})},$$

where x_N^* and x_D^* denote the modes of f_N and f, respectively, and where $C(x,y)$ denotes the normalizing constant of a Gaussian kernel with mean x and variance y. Specifically, this yields to

$$1 - w(n,k_n) \simeq \frac{f_N(x_N^*)}{f(x_D^*)}\left(\frac{(\log\circ f_N)''(x_N^*)}{(\log\circ f)''(x_D^*)}\right)^{-1/2}. \tag{S1.5}$$

The mode x_D^* is the only positive real root of the function $G(x) = \sigma x(\tau+x)^\sigma - (n-1)\tau - (\sigma k_n - 1)x$. A study of G shows that x_D^* is bounded by below by a positive constant times $n^{1/(1+\sigma)}$, which implies that the terms involving τ are negligible in the following renormalization of $G(x_D^*)$

$$\sigma\frac{x_D^*}{n}\left(\frac{\tau}{n} + \frac{x_D^*}{n}\right)^\sigma - \frac{n-1}{n^{\sigma+1}}\tau - \frac{\sigma k_n - 1}{n^\sigma}\frac{x_D^*}{n}.$$

The same calculation holds for x_N^*. According to the fluctuation limit (S0.1) there exists a nonnegative and finite random variable $S_{\sigma,g}$ such that $n^{-\sigma}K_n \xrightarrow{\text{a.s.}} S_{\sigma,g}$ as $n \to +\infty$. Let $\Omega_0 := \{\omega \in \Omega : \lim_{n\to+\infty} n^{-\sigma}K_n(w) = S_{\sigma,h}(\omega)\}$, and let $S_{\sigma,g}(\omega) = s_\sigma$ for any $\omega \in \Omega_0$. Then, we have

$$\frac{x_N^*}{n} \simeq \frac{x_D^*}{n} \simeq s_\sigma^{1/\sigma}. \tag{S1.6}$$

In order to make use of (S1.5), we also need an asymptotic equivalence for $x_D^* - x_N^*$. Note that $G(x_D^*) = 0$ and $G(x_N^*) = -x_N^*$ allow us to resort to a first order Taylor bound on G at x_N^* and shows that $x_D^* - x_N^*$ has a lower bound equivalent to $s_\sigma^{(1-\sigma)/\sigma}n^{1-\sigma}/\sigma^2$. The same argument applied to $G(x) + x$ at x_D^* provides an upper bound with the same asymptotic equivalence, thus

$$\frac{x_D^* - x_N^*}{n^{1-\sigma}} \simeq \frac{s_\sigma^{(1-\sigma)/\sigma}}{\sigma^2}. \tag{S1.7}$$

By studying f and f_N, as well as the second derivative of their logarithm, together with asymptotic equivalences (S1.6) and (S1.7), we can write $f(x_D^*) \simeq f(x_N^*)$ and $(\log\circ f)''(x_D^*) \simeq (\log\circ f)''(x_N^*) \simeq (\log\circ f_N)''(x_N^*)$. Hence, from (S1.5) one obtains $1-w(n,k_n) \simeq \tau/(\tau+x_N^*) \simeq \tau s_\sigma^{-1/\sigma}/n$, which leads to

$$\begin{aligned} g_{0,g}(n,k_n) &= 1-\left(1-\frac{\sigma k_n}{n}\right)\left(1-\tau s_\sigma^{-1/\sigma}\frac{1}{n}+o\left(\frac{1}{n}\right)\right), \\ &= \frac{\sigma k_n}{n}+\tau s_\sigma^{-1/\sigma}\frac{1}{n}+o\left(\frac{1}{n}\right), \end{aligned} \tag{S1.8}$$

and

$$\begin{aligned} g_{1,g}(n,k_n) &= \frac{1-g_{0,g}(n,k_n)}{n-\sigma k_n} = \frac{1}{n}\left(1-\frac{\tau s_\sigma^{-1/\sigma}/n+o\left(\frac{1}{n}\right)}{1-\frac{\sigma k}{n}}\right), \\ &= \frac{1}{n}\left(1-\frac{\tau s_\sigma^{-1/\sigma}}{n}+o\left(\frac{1}{n}\right)\right). \end{aligned} \tag{S1.9}$$

Expressions (S1.8) and (S1.9) provide second order approximations of $g_{0,g}(n,k_n)$ and $g_{1,g}(n,k_n)$, respectively. Recall that for any ω in Ω_0 we have $n^{-\sigma}k_n \simeq s_\sigma$, namely we can replace s_σ with $n^{-\sigma}k_n$. This is because of the fluctuation limit displayed in (S0.1). The proof is completed by combining (S1.8) and (S1.9) with the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ under a normalized GG prior. □

S2 Details on the derivation of $\hat{\mathcal{D}}_n(l) \simeq \check{\mathcal{D}}_n(l;\mathscr{S}_{\mathbf{PD}})$

Let us define $c_{\sigma,l} = \sigma(1-\sigma)_{l-1}/l!$ and recall that $\hat{\mathcal{D}}_n(0) = V_{n+1,k_n+1}/V_{n,k_n}$ and $\hat{\mathcal{D}}_n(l) = (l-\sigma)m_{l,n}V_{n+1,k_n}/V_{n,k_n}$. The relationship between the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ and the smoothed Good-Turing estimator $\check{\mathcal{D}}_n(l;\mathscr{S}_{\text{PD}})$ follows by combining Theorem 2 with the fluctuation limits (S0.1) and (S0.2). For any $\omega \in \Omega$, a version of the predictive distributions of $Q_{\sigma,h}$ is

$$\frac{V_{n+1,K_n(\omega)+1}}{V_{n,K_n(\omega)}}\nu_0(\cdot)+\frac{V_{n+1,K_n(\omega)}}{V_{n,K_n(\omega)}}\sum_{i=1}^{K_n(\omega)}(N_{i,n}(\omega)-\sigma)\delta_{X_i^*(\omega)}(\cdot).$$

According to (S0.1) and (S0.2), $\lim_{n\to+\infty} c_{\sigma,l}M_{l,n}/K_n = 1$ almost surely. See Lemma 3.11 in Pitman (2006) for additional details. By Theorem 2 we have $V_{n+1,K_n+1}/V_{n,K_n} \overset{\text{a.s.}}{\simeq} \sigma K_n/n$, and $M_{1,n} \overset{\text{a.s.}}{\simeq} \sigma K_n$, as $n \to +\infty$. Then, a version of the Bayesian nonparametric estimator of the 0-discovery coincides with

$$\begin{aligned} \frac{V_{n+1,K_n(\omega)+1}}{V_{n,K_n(\omega)}} &\simeq \frac{\sigma K_n(\omega)}{n} \\ &\simeq \frac{M_{1,n}(\omega)}{n}, \end{aligned} \tag{S2.1}$$

as $n \to +\infty$. By Theorem 2 we have $V_{n+1,K_n}/V_{n,K_n} \overset{\text{a.s.}}{\simeq} 1/n$, and $M_{l,n} \overset{\text{a.s.}}{\simeq} c_{\sigma,l}K_n$, as $n \to +\infty$. Accordingly, a version of the Bayesian nonparametric estimator of the l-discovery coincides with

$$\begin{aligned}(l-\sigma)M_{l,n}(\omega)\frac{V_{n+1,K_n(\omega)}}{V_{n,K_n(\omega)}} &\simeq (l-\sigma)\frac{M_{l,n}(\omega)}{n} \\ &\simeq c_{\sigma,l}(l-\sigma)\frac{K_n(\omega)}{n} \\ &\simeq (l+1)\frac{M_{l+1,n}(\omega)}{n},\end{aligned} \tag{S2.2}$$

as $n \to +\infty$. Let $\Omega_0 := \{\omega \in \Omega : \lim_{n\to+\infty} n^{-\sigma}K_n(w) = Z_{\sigma,\theta/\sigma}(\omega), \lim_{n\to+\infty} n^{-\sigma}M_{l,n}(\omega) = c_{\sigma,l}Z_{\sigma,\theta/\sigma}(\omega)\}$. From (S0.1) and (S0.2) we have $\mathbb{P}[\Omega_0] = 1$. Fix $\omega \in \Omega_0$ and denote by $k_n = K_n(\omega)$ and $m_{l,n} = M_{l,n}(\omega)$ the number of species generated and the number of species with frequency l generated by the sample $\boldsymbol{X}_n(\omega)$. Accordingly, $\hat{\mathcal{D}}_n(l) \simeq \check{\mathcal{D}}_n(l; \mathscr{S}_{\text{PD}})$ follows from (S2.1) and (S2.2).

S3 Additional illustrations

In this Section we provide additional illustrations accompanying those of Section 4 in the main manuscript. Specifically, we consider a Zeta distribution with parameter $s = 1.5$. We draw 500 samples of size $n = 1000$ from such distribution, we order them according to the number of observed species k_n, and we split them in 5 groups: for $i = 1, 2, \ldots, 5$, the i-th group of samples will be composed by 100 samples featuring a total number of observed species k_n that stays between the quantiles of order $(i-1)/5$ and $i/5$ of the empirical distribution of k_n. Then we pick at random one sample for each group and label it with the corresponding index i. This procedure leads to five samples. As shown in Table S1, the choice of $s = 1.5$ leads to samples with a smaller number of distinct values if compared with the case $s = 1.1$ (see also Table 1 in the main manuscript). Table S2, under the two parameter PD prior and the normalized GG prior, shows the estimated l-discoveries, for $l = 0, 1, 5, 10$, and the corresponding 95% posterior credible intervals. Finally, Figure S1 shows how the average ratio $\bar{r}_{1,2,n}$ evolves as the sample size increases (see Section 4.2 in the main manuscript).

Table S1: Simulated data with $s = 1.5$. For each sample we report the sample size n, the number of species k_n and the maximum likelihood values $(\hat{\sigma}, \hat{\theta})$ and $(\hat{\sigma}, \hat{\tau})$.

				PD		GG	
	sample	n	k_n	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\tau}$
	1	1000	128	0.624	1.207	0.622	3.106
	2	1000	135	0.675	0.565	0.673	0.957
Simulated data	3	1000	138	0.684	0.487	0.682	0.795
	4	1000	146	0.656	1.072	0.655	2.302
	5	1000	149	0.706	0.377	0.704	0.592

Table S2: Simulated data with $s = 1.5$. We report the true value of the probability $D_n(l)$ and the Bayesian nonparametric estimates of $D_n(l)$ with 95% credible intervals.

			Good–Turing		PD		GG	
l	sample	$D_n(l)$	$\check{\mathcal{D}}_n(l)$	95%-c.i.	$\hat{\mathcal{D}}_n(l)$	95%-c.i.	$\hat{\mathcal{D}}_n(l)$	95%-c.i.
	1	0.099	0.080	(0.010, 0.150)	0.081	(0.065, 0.098)	0.081	(0.065, 0.098)
	2	0.103	0.092	(0.012, 0.172)	0.092	(0.075, 0.110)	0.091	(0.075, 0.110)
0	3	0.095	0.096	(0.014, 0.178)	0.095	(0.078, 0.114)	0.095	(0.076, 0.113)
	4	0.096	0.096	(0.015, 0.177)	0.097	(0.079, 0.116)	0.097	(0.080, 0.115)
	5	0.093	0.108	(0.019, 0.197)	0.106	(0.087, 0.126)	0.105	(0.087, 0.124)
	1	0.030	0.038	(0.031, 0.045)	0.030	(0.020, 0.042)	0.030	(0.021, 0.042)
	2	0.037	0.030	(0.024, 0.036)	0.030	(0.021, 0.041)	0.030	(0.020, 0.042)
1	3	0.034	0.034	(0.028, 0.040)	0.030	(0.021, 0.042)	0.031	(0.021, 0.042)
	4	0.029	0.040	(0.033, 0.047)	0.033	(0.023, 0.045)	0.033	(0.022, 0.044)
	5	0.040	0.026	(0.021, 0.031)	0.032	(0.022, 0.044)	0.032	(0.023, 0.043)
	1	0.013	0.012	(0.008, 0.016)	0.013	(0.007, 0.021)	0.013	(0.007, 0.021)
	2	0.011	0.006	(0.003, 0.009)	0.004	(0.001, 0.009)	0.004	(0.001, 0.009)
5	3	0.010	0.012	(0.007, 0.017)	0.009	(0.004, 0.015)	0.009	(0.004, 0.016)
	4	0.010	0.036	(0.024, 0.048)	0.009	(0.004, 0.015)	0.009	(0.004, 0.015)
	5	0.012	0	(0, 0)	0.013	(0.007, 0.021)	0.013	(0.006, 0.021)
	1	0.019	0	(0, 0)	0.019	(0.011, 0.028)	0.019	(0.011, 0.028)
	2	0	0.011	n.a.	0	(0, 0)	0	(0,0)
10	3	0.011	0.011	(0.006, 0.016)	0.009	(0.004, 0.016)	0.009	(0.004, 0.016)
	4	0	0	n.a.	0	(0,0)	0	(0,0)
	5	0.006	0	(0, 0)	0.009	(0.004, 0.016)	0.009	(0.004, 0.017)

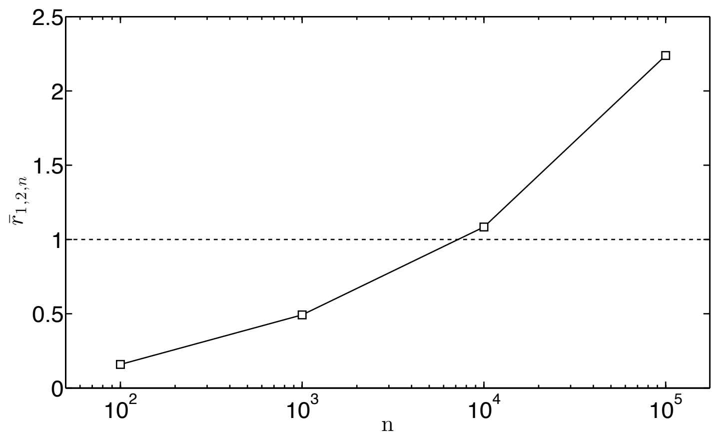

Figure S1: Average ratio $\bar{r}_{1,2,n}$ of sums of squared approximation errors for different sample sizes $n = 10^2, 10^3, 10^4, 10^5$. For the x-axis a logarithmic scale was used.

References

Jambunathan, M.V. (1954). Some Properties of Beta and Gamma Distributions. *Ann. Math. Statist.*. **25**, 401–405.

Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: A Festschrift for Terry Speed* (Goldstein, D.R. Eds.). Lecture notes monograph series. **40**, Institute of Mathematical Statistics.

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

Ruggiero, M., Walker, S.G., and Favaro, S. (2013). Alpha-diversity processes and normalized inverse Gaussian diffusions. *Ann. Appl. Probab.*. **23**, 386–425.