

Supplementary material for "Contaminated Gibbs-type priors"

Federico Camerlenghi^{*,†} and Riccardo Corradin[‡] and Andrea Ongaro^{*}

Abstract. This supplementary material contains additional proofs, figures, details on the simulation studies reported in the main paper.

Appendix A: Proofs

A.1 Proof of Proposition 1

The first assertion of Proposition 1 is immediate: indeed $E[\tilde{p}(A)] = \beta E[\tilde{q}(A)] + (1 - \beta)P_0(A)$, and $E[\tilde{q}(A)] = Q_0(Q)$ since \tilde{q} is a Gibbs-type prior. To prove the second assertion, note that

$$\text{cov}(\tilde{p}(A), \tilde{p}(B)) = \text{cov}(\beta\tilde{q}(A) + (1-\beta)P_0(A), \beta\tilde{q}(B) + (1-\beta)P_0(B)) = \beta^2 \text{cov}(\tilde{q}(A), \tilde{q}(B)),$$

with

$$\text{cov}(\tilde{q}(A), \tilde{q}(B)) = E[\tilde{q}(A)\tilde{q}(B)] - E[\tilde{q}(A)]E[\tilde{q}(B)]. \quad (\text{S1})$$

We focus on the evaluation of the first term in (S1)

$$E[\tilde{q}(A)\tilde{q}(B)] = P(X_1 \in A, X_2 \in B)$$

where $X_1, X_2 \mid \tilde{q} \stackrel{\text{iid}}{\sim} \tilde{q}$, thus, we get:

$$\begin{aligned} P(X_1 \in A, X_2 \in B) &= E[P(X_2 \in B \mid X_1) \cdot \mathbf{1}_A(X_1)] \\ &= E \left[\left((1 - \sigma) \frac{V_{2,1}}{V_{1,1}} \mathbf{1}_B(X_1) + \frac{V_{2,2}}{V_{1,1}} Q_0(B) \right) \mathbf{1}_A(X_1) \right], \end{aligned}$$

where we used the predictive distribution of a Gibbs-type prior. Then, the previous expression equals

$$E[\tilde{q}(A)\tilde{q}(B)] = (1 - \sigma) \frac{V_{2,1}}{V_{1,1}} Q_0(A \cap B) + \frac{V_{2,2}}{V_{1,1}} Q_0(A) Q_0(B).$$

We substitute the previous term in (S1), and we obtain

$$\begin{aligned} \text{cov}(\tilde{q}(A), \tilde{q}(B)) &= (1 - \sigma) \frac{V_{2,1}}{V_{1,1}} Q_0(A \cap B) + \frac{V_{2,2} - V_{1,1}}{V_{1,1}} Q_0(A) Q_0(B) \\ &= (1 - \sigma) \frac{V_{2,1}}{V_{1,1}} [Q_0(A \cap B) - Q_0(A) Q_0(B)] \end{aligned}$$

^{*}Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy federico.camerlenghi@unimib.it, andrea.ongaro@unimib.it

[†]Also affiliated to Collegio Carlo Alberto, Piazza V. Arbarello 8, Turin, and BIDS, Bocconi University, Milan, Italy.

[‡]School of Mathematical Sciences, University of Nottingham, Nottingham, United Kingdom riccardo.corradin@nottingham.ac.uk

where the last equality holds true in force of the recurrence relation of the weights $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$, which implies $V_{2,2} - V_{1,1} = -(1 - \sigma)V_{2,1}$.

Further, we remark that the correlation between $\tilde{p}(A)$ and $\tilde{p}(B)$ does not depend on β , since

$$\text{cor}(\tilde{p}(A), \tilde{p}(B)) = \frac{\beta^2 \text{cov}(\tilde{q}(A), \tilde{q}(B))}{\sqrt{\beta^2 \text{var}(\tilde{q}(A))} \sqrt{\beta^2 \text{var}(\tilde{q}(B))}} = \text{cor}(\tilde{q}(A), \tilde{q}(B)).$$

The inclusion of a contaminant term acts on the dispersion of \tilde{q} , but not on the dependency induced over its support.

A.2 Proof of Theorem 1

Let us denote by $\mu := 1/2P_0 + 1/2Q_0$ the diffuse probability measure on $(\mathbb{X}, \mathcal{X})$, with respect to which both P_0 and Q_0 are absolutely continuous measures. We would like to evaluate the EPPF using the definition (2) and focusing on the contaminated Gibbs-type prior case

$$\begin{aligned} \Pi_k^{(n)}(n_1, \dots, n_k) &= \int_{\mathbb{X}^k} \mathbb{E} \left[\prod_{i=1}^k (\beta \tilde{q}(\mathrm{d}x_i^*) + (1 - \beta)P_0(\mathrm{d}x_i^*))^{n_i} \right] \\ &= \int_{\mathbb{X}^k} \mathbb{E} \left[\prod_{i=1}^k \sum_{j_i=0}^{n_i} \binom{n_i}{j_i} \beta^{j_i} \tilde{q}^{j_i}(\mathrm{d}x_i^*) (1 - \beta)^{n_i - j_i} P_0^{n_i - j_i}(\mathrm{d}x_i^*) \right]. \quad (\text{S2}) \end{aligned}$$

We now concentrate on the evaluation of the expected value in (S2), that can be computed as follows

$$\begin{aligned} &\mathbb{E} \left[\prod_{i=1}^k \sum_{j_i=0}^{n_i} \binom{n_i}{j_i} \beta^{j_i} \tilde{q}^{j_i}(\mathrm{d}x_i^*) (1 - \beta)^{n_i - j_i} P_0^{n_i - j_i}(\mathrm{d}x_i^*) \right] \\ &= \sum_{j_1=0}^{n_1} \dots \sum_{j_k=0}^{n_k} \prod_{i=1}^k \binom{n_i}{j_i} \mathbb{E} \left[\prod_{i=1}^k \beta^{j_i} \tilde{q}^{j_i}(\mathrm{d}x_i^*) (1 - \beta)^{n_i - j_i} P_0^{n_i - j_i}(\mathrm{d}x_i^*) \right]. \quad (\text{S3}) \end{aligned}$$

We now recall that $m_1 = \#\{i : n_i = 1\}$ is the number of observations recorded only once out of the sample of size n . Without loss of generality we can assume that these observations are the first m_1 values $X_1^*, \dots, X_{m_1}^*$, which is tantamount to saying that $n_i = 1$ for any $i = 1, \dots, m_1$ and $n_i > 1$ if $i \geq m_1 + 1$. Neglecting the superior order

terms in (S3), we obtain

$$\begin{aligned} & \mathbb{E} \left[\prod_{i=1}^k \sum_{j_i=0}^{n_i} \binom{n_i}{j_i} \beta^{j_i} \tilde{q}^{j_i}(\mathrm{d}x_i^*) (1-\beta)^{n_i-j_i} P_0^{n_i-j_i}(\mathrm{d}x_i^*) \right] \\ &= \sum_{(j_1, \dots, j_{m_1}) \in \{0,1\}^{m_1}} \mathbb{E} \left[\prod_{i=1}^{m_1} \beta^{j_i} \tilde{q}^{j_i}(\mathrm{d}x_i^*) (1-\beta)^{1-j_i} P_0^{1-j_i}(\mathrm{d}x_i^*) \prod_{i=m_1+1}^k \beta^{n_i} \tilde{q}^{n_i}(\mathrm{d}x_i^*) \right] \\ & \quad + o \left(\prod_{i=1}^k \mu(\mathrm{d}x_i^*) \right). \end{aligned}$$

We now define $\bar{m}_1 = \#\{i : j_i = 0\}$ which represents the number of observations generated from P_0 , having frequency 1. Thus, by noticing that $\sum_{i=m_1+1}^k n_i = n - m_1$, we get

$$\begin{aligned} & \mathbb{E} \left[\prod_{i=1}^k \sum_{j_i=0}^{n_i} \binom{n_i}{j_i} \beta^{j_i} \tilde{q}^{j_i}(\mathrm{d}x_i^*) (1-\beta)^{n_i-j_i} P_0^{n_i-j_i}(\mathrm{d}x_i^*) \right] \\ &= \sum_{(j_1, \dots, j_{m_1}) \in \{0,1\}^{m_1}} \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} \mathbb{E} \left[\prod_{i=1}^{m_1} \tilde{q}^{j_i}(\mathrm{d}x_i^*) \prod_{i=m_1+1}^k \tilde{q}^{n_i}(\mathrm{d}x_i^*) \right] \prod_{i=1}^{m_1} P_0^{1-j_i}(\mathrm{d}x_i^*) \\ & \quad + o \left(\prod_{i=1}^k \mu(\mathrm{d}x_i^*) \right). \end{aligned}$$

By integrating the previous expression over \mathbb{X}^k , we get the expression of the EPPF

$$\begin{aligned} \Pi_k^{(n)}(n_1, \dots, n_k) &= \sum_{(j_1, \dots, j_{m_1}) \in \{0,1\}^{m_1}} \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} \\ & \quad \times \int_{\mathbb{X}^{k-\bar{m}_1}} \mathbb{E} \left[\prod_{i=1}^{m_1} \tilde{q}^{j_i}(\mathrm{d}x_i^*) \prod_{i=m_1+1}^k \tilde{q}^{n_i}(\mathrm{d}x_i^*) \right]. \end{aligned} \quad (\text{S4})$$

We recognize that the integral in (S4) is the EPPF of a Gibbs-type prior, therefore

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \sum_{(j_1, \dots, j_{m_1}) \in \{0,1\}^{m_1}} \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} V_{n-\bar{m}_1, k-\bar{m}_1} \prod_{i=m_1+1}^k (1-\sigma)_{n_i-1}, \quad (\text{S5})$$

where we have now to solve the summation over the j_i 's. We observe that each summand in (S5) depends on the vector (j_1, \dots, j_{m_1}) only through \bar{m}_1 . Moreover note that, fixed the value of \bar{m}_1 , there are $\binom{m_1}{\bar{m}_1}$ possible ways to choose (j_1, \dots, j_{m_1}) so that $\sum_{i=1}^{m_1} (1-j_i) = \bar{m}_1$, thus

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \sum_{\bar{m}_1=0}^{m_1} \binom{m_1}{\bar{m}_1} \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} V_{n-\bar{m}_1, k-\bar{m}_1} \prod_{i=m_1+1}^k (1-\sigma)_{n_i-1},$$

and the results easily follows having realized that the sum in the previous expression is an expected value w.r.t. the distribution of the Binomial random variable $\bar{M}_{m_1} \sim \text{Binom}(m_1, 1 - \beta)$.

A.3 Relative ratio between EPPFs: comments

Recall the probability ratio defined in the paper:

$$R(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k) := \frac{\Pi_k^{(n)}(n_1, \dots, n_k)}{\Pi_k^{(n)}(n'_1, \dots, n'_k)} \quad (\text{S6})$$

In this section, we want to compare the probability ratio when $\Pi_k^{(n)}$ is a Gibbs-type EPPF (3), denoted by R_G , and when it equals the EPPF of a contaminated Gibbs-type prior (5), denoted by R_{cG} .

Proposition 1. *Consider two compositions (n_1, \dots, n_k) and (n'_1, \dots, n'_k) deriving from two samples having the same size n and the same number of distinct values k , and denote by $m_1 = \#\{i : n_i = 1\}$ (resp. $m'_1 = \#\{i : n'_i = 1\}$) the number of singletons in the first (resp. second) composition. If $m_1 = m'_1$ then*

$$R_{cG}(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k) = R_G(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k),$$

whereas if $m_1 > m'_1$

$$R_{cG}(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k) \geq R_G(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k).$$

Proof. If $m_1 = m'_1$, we have that

$$\begin{aligned} R_{cG}(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k) &= \frac{E_{\bar{M}_{m_1}}[V_{n-\bar{M}_{m_1}, k-\bar{M}_{m_1}}] \beta^{n-m_1} \prod_{i=1}^k (1-\sigma)_{n_i-1}}{E_{\bar{M}_{m_1}}[V_{n-\bar{M}_{m_1}, k-\bar{M}_{m_1}}] \beta^{n-m_1} \prod_{i=1}^k (1-\sigma)_{n'_i-1}} \\ &= \frac{V_{n,k} \prod_{i=1}^k (1-\sigma)_{n_i-1}}{V_{n,k} \prod_{i=1}^k (1-\sigma)_{n'_i-1}} = R_G(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k) \end{aligned}$$

as stated.

For the second part of the proposition consider $m_1 > m'_1$. Firstly, observe that $V_{n-\bar{m}, k-\bar{m}}$ is an increasing function of \bar{m} , as a simple consequence of the recurrence relation which entails that

$$\frac{V_{n-\bar{m}+1, k-\bar{m}+1}}{V_{n-\bar{m}, k-\bar{m}}} \leq 1.$$

Moreover \bar{M}_{m_1} stochastically dominates $\bar{M}_{m'_1}$, since they are two binomial random variables with $m_1 > m'_1$, thus one has

$$E_{\bar{M}_{m_1}}[V_{n-\bar{M}_{m_1}, k-\bar{M}_{m_1}}] \geq E_{\bar{M}_{m'_1}}[V_{n-\bar{M}_{m'_1}, k-\bar{M}_{m'_1}}]. \quad (\text{S7})$$

One can exploit (S7) to conclude the proof:

$$\begin{aligned} R_{cG}(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k) &= \frac{E_{\bar{M}_{m_1}}[V_{n-\bar{M}_{m_1}, k-\bar{M}_{m_1}}] \beta^{n-m_1} \prod_{i=1}^k (1-\sigma)_{n_i-1}}{E_{\bar{M}_{m'_1}}[V_{n-\bar{M}_{m'_1}, k-\bar{M}_{m'_1}}] \beta^{n-m'_1} \prod_{i=1}^k (1-\sigma)_{n'_i-1}} \\ &\geq \beta^{m'_1-m_1} \frac{\prod_{i=1}^k (1-\sigma)_{n_i-1}}{\prod_{i=1}^k (1-\sigma)_{n'_i-1}} \geq \frac{\prod_{i=1}^k (1-\sigma)_{n_i-1}}{\prod_{i=1}^k (1-\sigma)_{n'_i-1}} \\ &= R_G(n_1, \dots, n_k; n'_1, \dots, n'_k; n, k) \end{aligned}$$

where we used the fact that $\beta^{m'_1-m_1} \geq 1$. \square

The previous proposition tells us that if the two configurations have the same number of singletons, the ratio is the same for the contaminated and non-contaminated model. On the other side, the ratio increases in the contaminated model if we increase the number of singletons in the composition at the numerator term. See Section A.3 for a proof of Proposition 1.

A.4 Proof of Proposition 2

The conditional predictive distribution (7) follows from the EPPF augmented with the introduction of the latent variables J_1, \dots, J_n , by considering all the possible scenarios: X_{n+1} is new from P_0 , X_{n+1} is new from Q_0 , X_{n+1} coincides with a previously observed value appearing with frequency one in the initial sample, and X_{n+1} coincides with a previously observed value having frequency $n_i \geq 2$ in the initial sample.

A.5 Proof of Proposition 3

Suppose now that $P_0 = Q_0$. We can integrate the distribution of the latent variables (J_1, \dots, J_n) in (7). We start from the law of the random partition dictated by the data, augmented with the introduction of the latent variables (J_1, \dots, J_{m_1}) . This can be recovered from the proof of Theorem 1, and assuming $P_0 = Q_0$ it amounts to be

$$\begin{aligned} \mathcal{L}(X_1, \dots, X_n, J_1, \dots, J_{m_1}) \\ = \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} V_{n-\bar{m}_1, k-\bar{m}_1} \prod_{i=\bar{m}_1+1}^k (1-\sigma)_{n_i-1} \prod_{i=1}^k P_0(dX_i^*) \end{aligned} \quad (S8)$$

where (j_1, \dots, j_{m_1}) are the observed values of (J_1, \dots, J_{m_1}) and $\bar{m}_1 = \#\{i : j_i = 0\}$ is the observed number of uniques generated from the diffuse component. From Equation (S8) we may apply the Bayes theorem to recover the conditional distribution of $(J_1, \dots, J_{m_1} | X_1, \dots, X_n)$ and this is proportional to

$$P(J_1 = j_1, \dots, J_{m_1} = j_{m_1} | X_1, \dots, X_n) \propto \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} V_{n-\bar{m}_1, k-\bar{m}_1} \quad (S9)$$

where the normalizing constant can be determined summing over all the vectors (J_1, \dots, J_{m_1}) belonging to the set $\{0, 1\}^{m_1}$. Indeed one can easily verify that

$$P(J_1 = j_1, \dots, J_{m_1} = j_{m_1} | X_1, \dots, X_n) = \frac{\beta^{m_1 - \bar{m}_1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1, k - \bar{m}_1}}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \quad (\text{S10})$$

with $\bar{M}_{m_1} \sim \text{Binom}(m_1, 1 - \beta)$. We can now integrate the expression in (7) w.r.t. the law (S10) to get the result. It is straightforward to integrate the first and the second terms on the r.h.s. of (7), obtaining

$$\sum_{(j_1, \dots, j_{m_1}) \in \{0, 1\}^{m_1}} \frac{\beta^{m_1 - \bar{m}_1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1, k - \bar{m}_1}}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} (1 - \beta) P_0(dx^*) = (1 - \beta) P_0(dx^*)$$

and

$$\begin{aligned} \sum_{(j_1, \dots, j_{m_1}) \in \{0, 1\}^{m_1}} \frac{\beta^{m_1 - \bar{m}_1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1, k - \bar{m}_1}}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \frac{V_{n - \bar{m}_1 + 1, k - \bar{m}_1 + 1}}{V_{n - \bar{m}_1, k - \bar{m}_1}} \beta P_0(dx^*) \\ = \frac{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1} + 1, k - \bar{M}_{m_1} + 1}]}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} (1 - \beta) P_0(dx^*) \end{aligned}$$

respectively. The last term can be decomposed in two sub-part, up to a multiplicative term β . For the first sub-part of the last term, fixing $i \in \{1, \dots, m_1\}$, we have to evaluate the following sum

$$\begin{aligned} \sum_{(j_1, \dots, j_{m_1}) \in \{0, 1\}^{m_1}} j_i (1 - \sigma) \frac{\beta^{m_1 - \bar{m}_1 + 1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1 + 1, k - \bar{m}_1}}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \\ = \sum_{\bar{m}_1=0}^{m_1} \sum_{(*)} j_i (1 - \sigma) \frac{\beta^{m_1 - \bar{m}_1 + 1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1 + 1, k - \bar{m}_1}}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \end{aligned} \quad (\text{S11})$$

where the sum $(*)$ is extended over all the possible vectors (j_1, \dots, j_{m_1}) such that $\sum_{h=1}^{m_1} j_h = m_1 - \bar{m}_1$. We note that if $j_i = 0$ the summand on the r.h.s. of (S11) is equal to 0, hence we can equivalently sum over all the vectors (j_1, \dots, j_{m_1}) such that $j_i = 1$. We further observe that, apart of j_i , the summand depends on (j_1, \dots, j_{m_1}) only through \bar{m}_1 . Thanks to these remarks, one has

$$\begin{aligned} \sum_{(j_1, \dots, j_{m_1}) \in \{0, 1\}^{m_1}} j_i (1 - \sigma) \frac{\beta^{m_1 - \bar{m}_1 + 1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1 + 1, k - \bar{m}_1}}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \\ = \sum_{\bar{m}_1=0}^{m_1-1} \#\{(j_1, \dots, j_{m_1}) : j_1 + \dots + j_{m_1} = m_1 - \bar{m}_1, j_i = 1\} \\ \times (1 - \sigma) \frac{\beta^{m_1 - \bar{m}_1 + 1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1 + 1, k - \bar{m}_1}}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \\ = \sum_{\bar{m}_1=0}^{m_1-1} \binom{m_1 - 1}{m_1 - \bar{m}_1 - 1} (1 - \sigma) \frac{\beta^{m_1 - \bar{m}_1 + 1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1 + 1, k - \bar{m}_1}}{E_{\bar{M}_{m_1}}[V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]}. \end{aligned}$$

By the fact that

$$\binom{m_1 - 1}{m_1 - \bar{m}_1 - 1} = \frac{(m_1 - \bar{m}_1)}{m_1} \binom{m_1}{\bar{m}_1}$$

the previous expression reduces to

$$\begin{aligned} & \sum_{(j_1, \dots, j_{m_1}) \in \{0,1\}^{m_1}} j_i (1 - \sigma) \frac{\beta^{m_1 - \bar{m}_1 + 1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1 + 1, k - \bar{m}_1}}{\mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \\ &= \sum_{\bar{m}_1=0}^{m_1} \frac{(m_1 - \bar{m}_1)}{m_1} \binom{m_1}{\bar{m}_1} (1 - \sigma) \frac{\beta^{m_1 - \bar{m}_1 + 1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1 + 1, k - \bar{m}_1}}{\mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \\ &= \frac{\beta (1 - \sigma) \mathbb{E}_{\bar{M}_{m_1}} [(m_1 - \bar{M}_{m_1}) V_{n - \bar{M}_{m_1} + 1, k - \bar{M}_{m_1}}]}{m_1 \mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \end{aligned}$$

and this provides the first sub-part of the last term on the r.h.s. of (8), after summing over $i = 1, \dots, m_1$. For the second sub-part of the last term of (8), integrated w.r.t. the law (S10), it is trivial to see that

$$\begin{aligned} & \sum_{(j_1, \dots, j_{m_1}) \in \{0,1\}^{m_1}} \frac{\beta^{m_1 - \bar{m}_1} (1 - \beta)^{\bar{m}_1} V_{n - \bar{m}_1, k - \bar{m}_1}}{\mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \\ & \quad \times \frac{V_{n - \bar{m}_1 + 1, k - \bar{m}_1}}{V_{n - \bar{m}_1, k - \bar{m}_1}} \sum_{i=m_1+1}^k (n_i - \sigma) \delta_{X_i^*}(\mathrm{d}x^*) \\ &= \sum_{i=m_1+1}^k (n_i - \sigma) \frac{\mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1} + 1, k - \bar{M}_{m_1}}]}{\mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} \delta_{X_i^*}(\mathrm{d}x^*) \end{aligned}$$

which conclude the proof.

A.6 Re-sampling mechanism: details

Assume that (X_1, \dots, X_n) is a sample from an exchangeable sequence of observations $X_i | \tilde{p} \stackrel{\text{iid}}{\sim} \tilde{p}$, where \tilde{p} is a random probability measure. In this section we study the ratio between the probability of re-sampling a distinct value observed n_1 times and the probability of re-sampling a value observed n_2 times out of the initial sample in two distinct cases: i) \tilde{p} is a Gibbs-type prior; ii) \tilde{p} is a contaminated Gibbs-type prior when $P_0 = Q_0$. From the predictive distribution (8) if $n_1, n_2 > 1$ the ratio in the two cases is the same. Indeed, under the contaminated Gibbs-type prior the ratio between the two probabilities is equal to

$$\frac{\beta(n_1 - \sigma) \mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1} + 1, k - \bar{M}_{m_1}}] \mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]}{\beta(n_2 - \sigma) \mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1} + 1, k - \bar{M}_{m_1}}] \mathbb{E}_{\bar{M}_{m_1}} [V_{n - \bar{M}_{m_1}, k - \bar{M}_{m_1}}]} = \frac{(n_1 - \sigma)}{(n_2 - \sigma)},$$

which is exactly the same for Gibbs-type prior. On the other hand, if $n_1 = 1$ and $n_2 > 1$, the ratio in the contaminated model decreases w.r.t. the same quantity under a

Gibbs-type prior specification. Indeed, the probability ratio when \tilde{p} is a contaminated Gibbs-type prior equals

$$\frac{\beta(1-\sigma)\mathbb{E}_{\tilde{M}_{m_1}}\left[\frac{(m_1-\tilde{M}_{m_1})}{m_1}V_{n-\tilde{M}_{m_1}+1,k-\tilde{M}_{m_1}}\right]\mathbb{E}_{\tilde{M}_{m_1}}[V_{n-\tilde{M}_{m_1},k-\tilde{M}_{m_1}}]}{\beta(n_2-\sigma)\mathbb{E}_{\tilde{M}_{m_1}}[V_{n-\tilde{M}_{m_1}+1,k-\tilde{M}_{m_1}}]\mathbb{E}_{\tilde{M}_{m_1}}[V_{n-\tilde{M}_{m_1},k-\tilde{M}_{m_1}}]}\leq\frac{(1-\sigma)}{(n_2-\sigma)},$$

where the last quantity corresponds to the probability ratio in the case of a Gibbs-type prior specification. The previous relations clarify that, conditionally on a sample X_1, \dots, X_n , the inclusion of a contaminant measure is preserving the same reinforcement as the discrete term of the model for observations with frequency larger than one, it mainly acts on singletons by decreasing the re-sampling probabilities w.r.t. observations with higher frequencies.

A.7 Proof of Proposition 4

We start from the augmented model

$$\begin{aligned} X_i \mid \tilde{p}, J_i &\stackrel{\text{iid}}{\sim} J_i \tilde{q} + (1 - J_i) P_0 \\ J_i &\stackrel{\text{iid}}{\sim} \text{Bern}(\beta). \end{aligned} \tag{S12}$$

We now use a slightly different notation to underline the dependence w.r.t. β of the cluster frequencies, in particular if V is a random variable depending on (X_1, \dots, X_n) in (S12), we write $V(\beta)$ to make explicit the dependence on β . If $\beta_1 < \beta_2$ we want to prove that $K_n(\beta_1)$ stochastically dominates $K_n(\beta_2)$, and the same for $M_{n,1}$. We recall that $\bar{M}_n(\beta) = \sum_{i=1}^n \{1 - J_i(\beta)\}$, which identifies the number of components coming from P_0 , therefore the following equality holds true a.s.

$$K_n(\beta) = \tilde{K}_{n-\bar{M}_n(\beta)} + \bar{M}_n(\beta) \tag{S13}$$

where $\tilde{K}_{n-\bar{M}_n(\beta)} \mid \bar{M}_n(\beta)$ is the number of distinct observations among the ones generated by the process \tilde{q} . Now let $\beta_1 < \beta_2$, under the contaminated model with parameter β_2 , $\sum_{i=1}^n J_i(\beta_2)$ represents the number of observations assigned to the discrete component \tilde{q} . Now we consider n independent Bernoulli random variables U_1, \dots, U_n with mean β_1/β_2 and we stochastically assign some of the $\sum_{i=1}^n J_i(\beta_2)$ observations to the diffuse component. The updated number of observations assigned to the diffuse component equals $\sum_{i=1}^n J_i(\beta_2)U_i$, which is a Binomial with parameters n and β_1 . As a consequence, the following chain holds true:

$$\begin{aligned} K_n(\beta_2) &= \tilde{K}_{\sum_{i=1}^n J_i(\beta_2)} + n - \sum_{i=1}^n J_i(\beta_2) \\ &\leq \tilde{K}_{\sum_{i=1}^n U_i J_i(\beta_2)} + \sum_{i=1}^n J_i(\beta_2) - \sum_{i=1}^n J_i(\beta_2)U_i + n - \sum_{i=1}^n J_i(\beta_2) \\ &= \tilde{K}_{\sum_{i=1}^n U_i J_i(\beta_2)} + n - \sum_{i=1}^n J_i(\beta_2)U_i \end{aligned}$$

where we used the fact that the distinct number of observations $\tilde{K}_{\sum_{i=1}^n J_i(\beta_2)}$ is always less than the number of observations still assigned to the discrete components plus the stochastic number of observations now assigned to P_0 . Note that the distribution of the random variable on the right hand side of the previous equation equals the distribution of

$$\tilde{K}_{n-\bar{M}_n(\beta_1)} + \bar{M}_n(\beta_1) = K_n(\beta_1)$$

as a consequence $P\{K_n(\beta_2) \geq x\} \leq P\{K_n(\beta_1) \geq x\}$, or in other words if $\beta_1 < \beta_2$, then $K_n(\beta_1)$ stochastically dominates $K_n(\beta_2)$. As for the observations with frequency 1, one can observe that

$$M_{n,1}(\beta) = \tilde{M}_{n-\bar{M}_n,1(\beta)} + \bar{M}_n(\beta)$$

where $\tilde{M}_{n-\bar{M}_n(\beta),1}$ denotes the number of observations with frequency 1 among those with $J_i(\beta) = 1$. Along similar lines as before, it is not difficult to see that if $\beta_1 < \beta_2$, then $M_{n,1}(\beta_1)$ stochastically dominates $M_{n,2}(\beta_2)$.

We now move to prove the asymptotic results, and we drop the explicit dependence on β of the sample statistics. We first focus on K_n , by exploiting the representation in (S13) one observe that $\tilde{K}_{n-\bar{M}_n}|\bar{M}_n$ is the number of distinct values in a sample of size $n - \bar{M}_n$ generated by a Gibbs-type prior \tilde{q} , then

$$P(\tilde{K}_m = k) = \frac{V_{m,k}}{\sigma^k} \mathcal{C}(m, k; \sigma), \quad \text{with } \mathcal{C}(m, k; \sigma) = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (-i\sigma)_m$$

see, e.g., (Gnedin and Pitman, 2005) and (Lijoi et al., 2007). Then, the distribution of K_n can be recovered by marginalizing out the distribution of \bar{M}_n as follows

$$\begin{aligned} P(K_n = k) &= \sum_{\ell=0}^k P(K_n = k | \bar{M}_n = \ell) P(\bar{M}_n = \ell) \\ &= \sum_{\ell=0}^k P(K_n = k | \bar{M}_n = \ell) \binom{n}{\ell} \beta^{n-\ell} (1-\beta)^\ell \\ &= \sum_{\ell=0}^k P(\tilde{K}_{n-\ell} = k - \ell) \binom{n}{\ell} \beta^{n-\ell} (1-\beta)^\ell \\ &= \sum_{\ell=0}^k \binom{n}{\ell} \beta^{n-\ell} (1-\beta)^\ell \frac{V_{n-\ell, k-\ell}}{\sigma^{k-\ell}} \mathcal{C}(n-\ell, k-\ell; \sigma). \end{aligned}$$

We can easily study the asymptotic distribution of K_n , by observing that

$$\frac{\bar{M}_n}{n} = \sum_{i=1}^n \frac{1 - J_i}{n} \xrightarrow{a.s.} (1 - \beta), \quad \text{as } n \rightarrow +\infty,$$

by the strong law of large numbers. Thanks to the previous equation we have that $n - \bar{M}_n = n(1 - \frac{\bar{M}_n}{n}) \xrightarrow{a.s.} +\infty$, since $1 - \frac{\bar{M}_n}{n}$ is almost surely positive and bounded as $n \rightarrow +\infty$. We further notice that

$$\frac{\tilde{K}_{n-\bar{M}_n}}{(n - \bar{M}_n)^\sigma} \rightarrow S_\sigma$$

where S_σ denotes a finite random variable termed σ -diversity (see, e.g., [Pitman, 2006](#); [De Blasi et al., 2015](#)). We can conclude that

$$\frac{K_n}{n} = \frac{\tilde{K}_{n-\bar{M}_n} + \bar{M}_n}{n} = \frac{\tilde{K}_{n-\bar{M}_n}}{(n - \bar{M}_n)^\sigma} \times \frac{(n - \bar{M}_n)^\sigma}{n} + \frac{\bar{M}_n}{n} \rightarrow 1 - \beta$$

almost surely as $n \rightarrow +\infty$.

We now study the convergence of $M_{n,r}$, the number of types having frequency r in the sample. We first recall that if $\tilde{M}_{n,r}$ is the number of unique elements with frequency r in a sample of size n generated from a Gibbs-type priors, then, thanks to ([Pitman, 2006](#), Lemma 3.11), one has

$$\frac{\tilde{M}_{n,r}}{n^\sigma} \rightarrow \frac{\sigma(1-\sigma)_{r-1}}{r!} S_\sigma, \quad \text{as } n \rightarrow +\infty. \quad (\text{S14})$$

See also ([Favaro et al., 2013](#)) for an explicit expression of the distribution of $\tilde{M}_{n,r}$. For the case $r = 1$, some observations are generated from P_0 and others from \tilde{q} , in formulas

$$M_{n,1} = \tilde{M}_{n-\bar{M}_n,1} + \bar{M}_n$$

where $\tilde{M}_{n-\bar{M}_n,1}$ denotes the number of observations with frequency 1 among those with $J_i = 1$. Thanks to ([S14](#)) we have that

$$\frac{\tilde{M}_{n-\bar{M}_n,1}}{(n - \bar{M}_n)^\sigma} \rightarrow \sigma S_\sigma,$$

since the random quantity $n - \bar{M}_n$ diverges as n grows to infinity. As a consequence we obtain

$$\frac{M_{n,1}}{n} = \frac{\tilde{M}_{n-\bar{M}_n,1} + \bar{M}_n}{n} = \frac{\tilde{M}_{n-\bar{M}_n,1}}{(n - \bar{M}_n)^\sigma} \times \frac{(n - \bar{M}_n)^\sigma}{n} + \frac{\bar{M}_n}{n} \rightarrow 1 - \beta, \quad \text{as } n \rightarrow +\infty.$$

If we now concentrate on the case $r \geq 2$, all the observations are generated from \tilde{q} , and we observe that

$$\frac{M_{n,r}}{n^\sigma} = \frac{\tilde{M}_{n-\bar{M}_n,r}}{n^\sigma} = \frac{\tilde{M}_{n-\bar{M}_n,r}}{(n - \bar{M}_n)^\sigma} \frac{(n - \bar{M}_n)^\sigma}{n^\sigma} \rightarrow \frac{\sigma(1-\sigma)_{r-1}}{r!} S_\sigma \beta^\sigma, \quad \text{as } n \rightarrow +\infty,$$

thanks to ([S14](#)) and the fact that $\frac{\bar{M}_n}{n} \rightarrow 1 - \beta$.

Appendix B: The contaminated Pitman-Yor processes case

B.1 Details for the determination of (10)

We resort to Theorem 1 and we specialize the expression of the EPPF in Equation (5) to the Pitman-Yor case:

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \sum_{\bar{m}_1=0}^{n_1} \binom{m_1}{\bar{m}_1} \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} \prod_{i=m_1+1}^k (1-\sigma)_{n_i-1} \frac{\prod_{i=1}^{k-\bar{m}_1-1} (\vartheta + i\sigma)}{(\vartheta + 1)_{n-\bar{m}_1-1}}$$

where we used in (5) the specification of the $V_{n,k}$'s of the Pitman-Yor process (9). It can be easily seen that

$$\begin{aligned} \Pi_k^{(n)}(n_1, \dots, n_k) &= \prod_{i=m_1+1}^k (1-\sigma)_{n_i-1} \sum_{\bar{m}_1=0}^{m_1} \binom{m_1}{\bar{m}_1} \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} \\ &\quad \times \frac{\sigma^{k-\bar{m}_1-1} \Gamma(\vartheta/\sigma + k - \bar{m}_1) \Gamma(\vartheta + 1)}{\Gamma(\vartheta/\sigma + 1) \Gamma(\vartheta + n - \bar{m}_1)} \\ &= \prod_{i=m_1+1}^k (1-\sigma)_{n_i-1} \beta^{n-m_1} \\ &\quad \times \sum_{\bar{m}_1=0}^{m_1} \binom{m_1}{\bar{m}_1} \beta^{m_1-\bar{m}_1} (1-\beta)^{\bar{m}_1} \frac{\sigma^{k-\bar{m}_1} (\vartheta/\sigma)_{k-\bar{m}_1}}{(\vartheta)_{n-\bar{m}_1}}. \end{aligned}$$

B.2 Monotonicity of the probability of sampling a new value

In this section we provide all the details to show that the probability of sampling a new value is monotone as a function of the number of distinct values m_1 for the contaminated Pitman-Yor process. First of all we prove the following.

Lemma 1. *Under the contaminated Pitman-Yor prior, the posterior distribution of \bar{M}_{m_1} , counting the number of values assigned to the diffuse component, satisfies the monotone likelihood ratio property, i.e., if $m_1 < m'_1$, then*

$$\frac{P(\bar{M}_{m'_1} = \bar{m} | X_1, \dots, X_n)}{P(\bar{M}_{m_1} = \bar{m} | X_1, \dots, X_n)}$$

increases as \bar{m} increases.

Proof. Recall that the posterior distribution of \bar{M}_{m_1} equals

$$P(\bar{M}_{m_1} = \bar{m} | X_1, \dots, X_n) \propto \binom{m_1}{\bar{m}} \beta^{n-\bar{m}} (1-\beta)^{\bar{m}} V_{n-\bar{m}, k-\bar{m}}$$

where the normalizing factor does not depend on \bar{m} , but only on m_1 and β . To prove the monotone likelihood ratio property, we consider $m_1 < m'_1$ and we focus on the ratio

$$\begin{aligned} \frac{P(\bar{M}_{m'_1} = \bar{m} | X_1, \dots, X_n)}{P(\bar{M}_{m_1} = \bar{m} | X_1, \dots, X_n)} &\propto \binom{m'_1}{\bar{m}} \cdot \binom{m_1}{\bar{m}}^{-1} = \frac{m'_1!}{m_1!} \cdot \frac{(m_1 - \bar{m})!}{(m'_1 - \bar{m})!} \\ &\propto \frac{(m_1 - \bar{m})!}{(m'_1 - \bar{m})!} = \frac{1}{(m'_1 - \bar{m}) \cdots (m_1 - \bar{m} + 1)}. \end{aligned}$$

The previous ratio increases as \bar{m} increases, as long as $m'_1 > m_1$. Thus, the result follows. \square

Thanks to Lemma 1 we are able to show the monotone property stated at the beginning of the section. From the predictive distribution (11), we observe that, conditionally on $\bar{M}_{m_1} = \bar{m}_1$, the probability of sampling a new value at the $(n+1)$ th stage equals

$$p_{new}(\bar{m}_1) := (1 - \beta) + \beta \frac{\vartheta + (k - \bar{m}_1)\sigma}{\vartheta + n - \bar{m}_1}$$

and it is immediate to see that such a probability is monotone in \bar{m}_1 , when n and k are fixed values. Moreover if $\sigma \rightarrow 0$, i.e. the Dirichlet case, $p_{new}(\bar{m}_1)$ is increasing, whereas if $\theta \rightarrow 0$ (stable process) the probability $p_{new}(\bar{m}_1)$ decreases in \bar{m}_1 . Note that the unconditional probability of sampling a new value may be recovered integrating $p_{new}(\bar{m}_1)$ with respect to the posterior distribution of \bar{M}_{m_1} . As a consequence of Lemma 1 such a distribution satisfies the monotone likelihood ratio property in \bar{m}_1 , thus the unconditional probability of sampling a new value is monotone as a function of the number of distinct values m_1 . This results in a richer predictive structure w.r.t. the Pitman-Yor case, where m_1 does not appear in the probability of sampling a new value.

B.3 Posterior inference

In the present section we face posterior inference for the contaminated Pitman-Yor process of Section 4. In particular, conditionally on a sample (X_1, \dots, X_n) of size n we derive closed-form expressions for the posterior expected value of the following statistics: i) $K_m^{(n)}$, i.e., the number of distinct values out a future sample X_{n+1}, \dots, X_{n+m} not yet observed in the initial sample; ii) $N_{m,r}^{(n)}$, which denotes the number of new and distinct observations with frequency r out of the additional sample, hitherto unobserved in the initial sample. By virtue of the posterior results, we also get formulas for the expected value of: i) K_n , the number of distinct values out of (X_1, \dots, X_n) ; ii) $M_{n,r}$, the number of clusters with frequency r out of (X_1, \dots, X_n) . Note that, in our framework, it is important to focus separately on the case $r = 1$ and $r \geq 2$, since the contaminated model acts in a different way on the number of observations with frequency one. Our results are based on the expressions derived by Favaro et al. (2009, 2013), who have faced posterior inference for the number of blocks with a certain frequency generated by Gibbs-type random partitions. It is possible to extend the results by Favaro et al. (2009, 2013) to contaminated Gibbs-type priors, here, for the easy of exposition, we discuss the contaminated Pitman-Yor case.

B.4 Posterior expected value of $N_{m,1}^{(n)}$

Conditionally on a sample X_1, \dots, X_n , we focus on predicting the number of new and distinct observations out of the additional sample X_{n+1}, \dots, X_{n+m} observed with frequency 1, denoted here as $N_{m,1}^{(n)}$. This is an important quantity in our framework, indeed the contaminated model mainly acts on the number of observations with frequency 1. We introduce the following random variables: i) $\bar{M}_n = \sum_{i=1}^n (1 - J_i)$ the number of observations generated by the diffuse component out of the sample of size n ; ii) $\bar{M}_m^{(n)} := \sum_{i=1}^m (1 - J_{i+n})$ the number of observations generated by the diffuse component out of the additional sample of size m . Note that these random variables have Binomial distributions and are independent.

We now focus on the evaluation of $E[N_{m,1}^{(n)} | X_1, \dots, X_n, \bar{M}_n]$, we first observe that

$$N_{m,1}^{(n)} = \tilde{N}_{m-\bar{M}_m^{(n)},1}^{(n-\bar{M}_n)} + \bar{M}_m^{(n)}$$

where $\tilde{N}_{m-\bar{M}_m^{(n)},1}^{(n-\bar{M}_n)}$ denotes the number of distinct values out of the additional sample observed with frequency 1 and coming from the discrete component, whose posterior expectation has been derived by (Favaro et al., 2013, Equation (30)). As a consequence we get:

$$\begin{aligned} E[N_{m,1}^{(n)} | X_1, \dots, X_n, \bar{M}_n] &= E[\tilde{N}_{m-\bar{M}_m^{(n)},1}^{(n-\bar{M}_n)} + \bar{M}_m^{(n)} | X_1, \dots, X_n, \bar{M}_n] \\ &= E[E[\tilde{N}_{m-\bar{M}_m^{(n)},1}^{(n-\bar{M}_n)} + \bar{M}_m^{(n)} | X_1, \dots, X_n, \bar{M}_n, \bar{M}_m^{(n)}] | X_1, \dots, X_n, \bar{M}_n] \\ &= E\left[(m - \bar{M}_m^{(n)})(\vartheta + (k - \bar{M}_n)\sigma) \cdot \frac{(\vartheta + n - \bar{M}_n + \sigma)_{m-\bar{M}_m^{(n)}-1}}{(\vartheta + n - \bar{M}_n)_{m-\bar{M}_m^{(n)}}} \right. \\ &\quad \left. + \bar{M}_m^{(n)} \middle| X_1, \dots, X_n, \bar{M}_n\right] \end{aligned}$$

where we used (Favaro et al., 2013, Equation (30)). By observing that $\bar{M}_m^{(n)}$ has a Binomial distribution with parameters $(m, 1 - \beta)$ and it is independent of X_1, \dots, X_n , we obtain

$$\begin{aligned} E[N_{m,1}^{(n)} | X_1, \dots, X_n, \bar{M}_n] &= \sum_{\ell=0}^m \left[(m - \ell)(\vartheta + (k - \bar{M}_n)\sigma) \cdot \frac{(\vartheta + n - \bar{M}_n + \sigma)_{m-\ell-1}}{(\vartheta + n - \bar{M}_n)_{m-\ell}} + \ell \right] \binom{m}{\ell} (1 - \beta)^\ell \beta^{m-\ell}. \end{aligned} \quad (\text{S15})$$

In order to find $E[N_{m,1}^{(n)} | X_1, \dots, X_n]$, we need to marginalize the previous expectation with respect to the conditional distribution of \bar{M}_n given X_1, \dots, X_n , which can be derived from the EPPF:

$$P(\bar{M}_n = \bar{m}_1 | X_1, \dots, X_n) = \frac{1}{C} \cdot \binom{m_1}{\bar{m}_1} \beta^{n-\bar{m}_1} (1 - \beta)^{\bar{m}_1} \sigma^{k-\bar{m}_1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta + n - \bar{m}_1)} \quad (\text{S16})$$

where C is the normalizing factor, i.e.,

$$C = \sum_{s=0}^{m_1} \binom{m_1}{s} \beta^{n-s} (1-\beta)^s \sigma^{k-s} \frac{\Gamma(\vartheta/\sigma + k - s)}{\Gamma(\vartheta + n - s)}.$$

Thus integrating (S15) with respect to the distribution (S16) we get

$$\begin{aligned} \mathbb{E} \left[N_{m,1}^{(n)} | X_1, \dots, X_n \right] &= \sum_{\bar{m}_1=0}^{m_1} \sum_{\ell=0}^m [(m-\ell)(\vartheta + (k - \bar{m}_1)\sigma) \cdot \frac{(\vartheta + n - \bar{m}_1 + \sigma)_{m-\ell-1}}{(\vartheta + n - \bar{m}_1)_{m-\ell}} \\ &\quad + \ell] \binom{m}{\ell} (1-\beta)^{\ell+\bar{m}_1} \beta^{m-\ell+n-\bar{m}_1} \frac{1}{C} \cdot \binom{m_1}{\bar{m}_1} \sigma^{k-\bar{m}_1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta + n - \bar{m}_1)}. \end{aligned} \quad (\text{S17})$$

By a close inspection of the conditional expected value in (S17), it is immediate to realize that it depends on the initial sample X_1, \dots, X_n through the sample size n , the number of distinct species k and also by m_1 , i.e., the number of species having frequency 1 in the initial sample. This is a remarkable difference with respect to Gibbs-type priors, in which the posterior expected value depends only on n and k (Bacallado et al., 2017), thus improving the flexibility of the model. Finally one may use (S17) to determine the distribution of $M_{n,1}$ for contaminated model by simply setting $n = 0$ and replacing m with n , thus obtaining:

$$\mathbb{E}[M_{n,1}] = n(1-\beta) + \sum_{\ell=0}^n (n-\ell) \frac{(\vartheta + \sigma)_{n-\ell-1}}{(\vartheta + 1)_{n-\ell-1}} \binom{n}{\ell} \beta^{n-\ell} (1-\beta)^\ell. \quad (\text{S18})$$

We finally provide the reader with another representation of the summation appearing in the previous equation (S18), by a change of variable $r = n - \ell$ we get

$$\begin{aligned} \sum_{\ell=0}^n (n-\ell) \frac{(\vartheta + \sigma)_{n-\ell-1}}{(\vartheta + 1)_{n-\ell-1}} \binom{n}{\ell} \beta^{n-\ell} (1-\beta)^\ell &= \sum_{r=1}^n r \frac{(\vartheta + \sigma)_{r-1}}{(\vartheta + 1)_{r-1}} \binom{n}{r} \beta^r (1-\beta)^{n-r} \\ &= \sum_{r=0}^{n-1} (r+1) \frac{(\vartheta + \sigma)_r}{(\vartheta + 1)_r} \binom{n}{r+1} \beta^{r+1} (1-\beta)^{n-1-r} \\ &= n\beta \sum_{r=0}^{n-1} \frac{(\vartheta + \sigma)_r}{(\vartheta + 1)_r} \binom{n-1}{r} \beta^r (1-\beta)^{n-1-r}. \end{aligned}$$

We now observe that the ratio between Pochhammer symbols is the expected value of B_1^r , where B_1 is a random variable having a Beta distribution with parameters $(\vartheta + \sigma, 1 - \sigma)$, then:

$$\begin{aligned} \sum_{\ell=0}^n (n-\ell) \frac{(\vartheta + \sigma)_{n-\ell-1}}{(\vartheta + 1)_{n-\ell-1}} \binom{n}{\ell} \beta^{n-\ell} (1-\beta)^\ell &= n\beta \sum_{r=0}^{n-1} \mathbb{E}[B_1^r] \binom{n-1}{r} \beta^r (1-\beta)^{n-1-r} \\ &= n\beta \mathbb{E}[(\beta B_1 + (1-\beta))^{n-1}]. \end{aligned}$$

As a consequence the expected value of $M_{n,1}$ boils down to

$$\mathbb{E}[M_{n,1}] = n(1-\beta) + n\beta \mathbb{E}[(\beta B_1 + (1-\beta))^{n-1}]. \quad (\text{S19})$$

B.5 Posterior expected value of $N_{m,r}^{(n)}$, with $r \geq 2$

Conditionally on a sample X_1, \dots, X_n , we focus on predicting the number of new and distinct observations out of the additional sample X_{n+1}, \dots, X_{n+m} observed with frequency $r \geq 2$, denoted here as $N_{m,r}^{(n)}$. We exploit the same notation introduced in Section B.4

We now focus on the evaluation of $E[N_{m,r}^{(n)} | X_1, \dots, X_n, \bar{M}_n]$, we first observe that

$$N_{m,r}^{(n)} = \tilde{N}_{m-\bar{M}_m^{(n)},r}^{(n-\bar{M}_n)}$$

where $\tilde{N}_{m-\bar{M}_m^{(n)},r}^{(n-\bar{M}_n)}$ takes into account the contribution of the discrete component \tilde{q} and it denotes the number of clusters containing r observations out of the additional sample and not observed in X_1, \dots, X_n . The posterior expected value of $\tilde{N}_{m-\bar{M}_m^{(n)},r}^{(n-\bar{M}_n)}$ has been found in (Favaro et al., 2013, Equation (30)), thus:

$$\begin{aligned} E[N_{m,r}^{(n)} | X_1, \dots, X_n, \bar{M}_n] &= E \left[E \left[\tilde{N}_{m-\bar{M}_m^{(n)},r}^{(n-\bar{M}_n)} | X_1, \dots, X_n, \bar{M}_n, \bar{M}_m^{(n)} \right] | X_1, \dots, X_n, \bar{M}_n \right] \\ &= E \left[\binom{m - \bar{M}_m^{(n)}}{r} (1 - \sigma)_{r-1} (\vartheta + (k - \bar{M}_n) \sigma) \right. \\ &\quad \times \left. \frac{(\vartheta + n - \bar{M}_n + \sigma)_{m - \bar{M}_m^{(n)} - r}}{(\vartheta + n - \bar{M}_n)_{m - \bar{M}_m^{(n)}}} \middle| X_1, \dots, X_n, \bar{M}_n \right]. \end{aligned}$$

By marginalizing the previous expression over $\bar{M}_m^{(n)}$, which is distributed as a Binomial with parameters m and $1 - \beta$, we get

$$\begin{aligned} E[N_{m,r}^{(n)} | X_1, \dots, X_n, \bar{M}_n] &= \sum_{\ell=0}^{m-r} \binom{m-\ell}{r} (1 - \sigma)_{r-1} (\vartheta + \sigma(k - \bar{M}_n)) \\ &\quad \times \frac{(\vartheta + n - \bar{M}_n + \sigma)_{m-r-\ell}}{(\vartheta + n - \bar{M}_n)_{m-\ell}} \binom{m}{\ell} (1 - \beta)^\ell \beta^{m-\ell}. \end{aligned} \quad (\text{S20})$$

The posterior expected value $E[N_{m,r}^{(n)} | X_1, \dots, X_n]$ can be easily evaluated by marginalizing (S20) with respect to the posterior distribution of \bar{M}_n , which appears in (S16). Thus, we obtain:

$$\begin{aligned} E[N_{m,r}^{(n)} | X_1, \dots, X_n] &= \sum_{\bar{m}_1=0}^{m_1} \sum_{\ell=0}^{m-r} \binom{m-\ell}{r} (1 - \sigma)_{r-1} (\vartheta + \sigma(k - \bar{m}_1)) \frac{(\vartheta + n - \bar{m}_1 + \sigma)_{m-r-\ell}}{(\vartheta + n - \bar{m}_1)_{m-\ell}} \\ &\quad \times \binom{m}{\ell} \beta^{m+n-\ell-\bar{m}_1} (1 - \beta)^{\ell+\bar{m}_1} \frac{\sigma^{k-\bar{m}_1}}{C} \binom{m_1}{\bar{m}_1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta + n - \bar{m}_1)}. \end{aligned} \quad (\text{S21})$$

As a consequence of (S21), one may derive an expression for $E[M_{n,r}]$, considering $n = 0$ and substituting n in place of m :

$$E[M_{n,r}] = \sum_{\ell=0}^{n-r} \binom{n-\ell}{r} (1 - \sigma)_{r-1} \vartheta \frac{(\vartheta + \sigma)_{n-r-\ell}}{(\vartheta)_{n-\ell}} \binom{n}{\ell} \beta^{n-\ell} (1 - \beta)^\ell \quad (\text{S22})$$

and, proceeding along similar lines as in Section B.4, one may easily see that

$$\mathbb{E}[M_{n,r}] = \frac{(1-\sigma)_{r-1}}{(\vartheta+1)_{r-1}} \binom{n}{r} \beta^r \mathbb{E}[(B_r \beta + 1 - \beta)^{n-r}] \quad (\text{S23})$$

where B_r is a Beta distribution with parameters $(\vartheta + \sigma, r - \sigma)$.

B.6 Posterior expected value of $K_m^{(n)}$

Here we focus on $K_m^{(n)}$, which represents the number of new and distinct observations out of an additional sample X_{n+1}, \dots, X_{n+m} , hitherto unobserved in X_1, \dots, X_n . More specifically, we are interested in the evaluation of its posterior expected value. Note that

$$K_m^{(n)} = \tilde{K}_{m-\bar{M}_m^{(n)}}^{(n-\bar{M}_n)} + \bar{M}_m^{(n)},$$

in which we have decomposed the new clusters generated by the discrete component ($\tilde{K}_{m-\bar{M}_m^{(n)}}^{(n-\bar{M}_n)}$) and the ones due to the diffuse component ($\bar{M}_m^{(n)}$). By resorting to (Favaro et al., 2009, Equation (6)), we have

$$\begin{aligned} \mathbb{E}[K_m^{(n)} | X_1, \dots, X_n, \bar{M}_n] &= \mathbb{E} \left[\mathbb{E} \left[\tilde{K}_{m-\bar{M}_m^{(n)}}^{(n-\bar{M}_n)} + \bar{M}_m^{(n)} | X_1, \dots, X_n, \bar{M}_n, \bar{M}_m^{(n)} \right] | X_1, \dots, X_n, \bar{M}_n \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\tilde{K}_{m-\bar{M}_m^{(n)}}^{(n-\bar{M}_n)} | X_1, \dots, X_n, \bar{M}_n, \bar{M}_m^{(n)} \right] + \bar{M}_m^{(n)} | X_1, \dots, X_n, \bar{M}_n \right] \\ &= \mathbb{E} \left[(k - \bar{M}_n + \vartheta/\sigma) \left(\frac{(\vartheta + n - \bar{M}_n + \sigma)_{m-\bar{M}_m^{(n)}}}{(\vartheta + n - \bar{M}_n)_{m-\bar{M}_m^{(n)}}} - 1 \right) + \bar{M}_m^{(n)} | X_1, \dots, X_n, \bar{M}_n \right]. \end{aligned}$$

The expected value in the r.h.s. of the previous expression is taken with respect to $\bar{M}_m^{(n)}$, having a Binomial distribution with parameters m and probability of success $1 - \beta$, therefore

$$\begin{aligned} \mathbb{E}[K_m^{(n)} | X_1, \dots, X_n, \bar{M}_n] &= \sum_{\ell=0}^m \left[(k - \bar{M}_n + \vartheta/\sigma) \left(\frac{(\vartheta + n - \bar{M}_n + \sigma)_{m-\ell}}{(\vartheta + n - \bar{M}_n)_{m-\ell}} - 1 \right) + \ell \right] \binom{m}{\ell} (1 - \beta)^\ell \beta^{m-\ell}. \end{aligned} \quad (\text{S24})$$

In order to obtain the posterior expected value of $K_m^{(n)}$ we need to marginalize the r.h.s. of (S24) with respect to the posterior distribution of \bar{M}_n , which has been found in (S16), as a consequence we get

$$\begin{aligned} \mathbb{E}[K_m^{(n)} | X_1, \dots, X_n] &= \sum_{\bar{m}_1=0}^{m_1} \sum_{\ell=0}^m \left[(k - \bar{m}_1 + \vartheta/\sigma) \left(\frac{(\vartheta + n - \bar{m}_1 + \sigma)_{m-\ell}}{(\vartheta + n - \bar{m}_1)_{m-\ell}} - 1 \right) + \ell \right] \\ &\quad \times \binom{m}{\ell} \beta^{m+n-\bar{m}_1-\ell} (1 - \beta)^{\ell+\bar{m}_1} \frac{\sigma^{k-\bar{m}_1}}{C} \binom{m_1}{\bar{m}_1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta + n - \bar{m}_1)}. \end{aligned} \quad (\text{S25})$$

We can now evaluate the expectation of the number of distinct values out of the initial sample, K_n , by setting $n = 0$ in (S25) and replacing m with n , more precisely we obtain

$$\mathbb{E}[K_n] = \sum_{\ell=0}^n \left[\frac{\vartheta}{\sigma} \left(\frac{(\vartheta + \sigma)_{n-\ell}}{(\vartheta)_{n-\ell}} - 1 \right) + \ell \right] \binom{n}{\ell} (1 - \beta)^\ell \beta^{n-\ell}. \quad (\text{S26})$$

By reasoning as in Section B.4, it is also possible to rewrite the expected value in (S26) in a simpler way

$$\mathbb{E}[K_n] = \frac{\vartheta}{\sigma} \mathbb{E}[(B_1 \beta + 1 - \beta)^n] + \frac{n\beta}{\sigma} \mathbb{E}[B_1 (B_1 \beta + 1 - \beta)^{n-1}] - \frac{\vartheta}{\sigma} + n(1 - \beta) \quad (\text{S27})$$

where B_1 is a Beta with parameters $\vartheta + \sigma$ and $1 - \sigma$.

Appendix C: Out-of-sample prediction

We perform a simulation study to investigate the difference between the contaminated and non-contaminated Pitman-Yor model in terms of prediction. More precisely we generate a sample X_1, \dots, X_n of size $n = 9000$ from the contaminated Pitman-Yor model with $\sigma = 0.2, \vartheta = 50$ and $\beta = 0.9$, and we first use the sample to estimate the parameters of the two models. Then, we predict the posterior expected values $\mathbb{E}[N_{m,1}^{(n)} | X_1, \dots, X_n]$, $\mathbb{E}[N_{m,2}^{(n)} | X_1, \dots, X_n]$, and $\mathbb{E}[K_m^{(n)} | X_1, \dots, X_n]$ for different additional sample sizes $m = 1, \dots, 1000$ for the two models, using (S20), (S17) and (S25) with the estimated parameters. The predicted quantities are compared with the *oracle* curves $n_{m,1}^{(n)}$, $n_{m,2}^{(n)}$ and $k_m^{(n)}$, which are obtained averaging over 1000 trajectories of $N_{m,r}^{(n)}$, as $r = 1, 2$, and $K_m^{(n)}$ from the generating contaminated Pitman-Yor process.

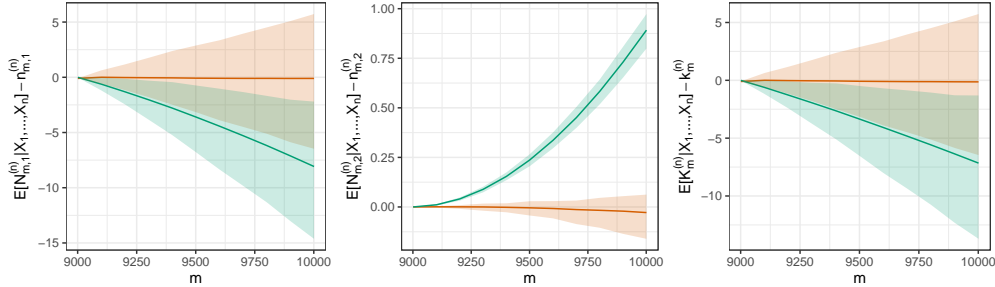


Figure 1: Simulation results of the difference between the number of new species with frequency one, the new number of species with frequency two and the number of new species in an additional sample of size m , discarded by their corresponding estimated true values. Two distinct models: the contaminated Pitman-Yor model (green) and the Pitman-Yor model (orange).

Each panel of Figure 1 refers to a different statistic ($N_{m,1}^{(n)}$, $N_{m,2}^{(n)}$, $K_m^{(n)}$). Each panel contains two curves for different values of the additional sample size m : the green curve

shows the difference between the oracle curve and the predicted value under the contaminated prior; the orange curve represents the difference between the oracle and the corresponding prediction under the Pitman-Yor model. All the experiments are averaged over 100 iterations. From Figure 1, it is appreciable how the contaminated Pitman-Yor process has an error which is averagely stable around zero for all the quantity considered, while the Pitman-Yor process, in presence of contamination, tends to underestimate the number of new elements with frequency one and the number of new distinct elements, and it overestimates the number of new elements with frequency two.

Appendix D: Posterior inference with the contaminated Pitman-Yor model

We can exploit the representation of the EPPF provided in (10) to sample realizations from the posterior distribution of the main quantities of interest. Let $\mathcal{L}(\sigma)$, $\mathcal{L}(\vartheta)$ and $\mathcal{L}(\beta)$ denote the prior distributions of σ , ϑ and β , respectively. We briefly describe an algorithm to perform posterior inference in the case of contaminated Pitman-Yor model, given a set of data $X_{1:n}$, by sampling R realizations from the posterior distributions of the quantities of interest. Such a sampling scheme can be used to sample posterior realizations of the parameters characterizing the urn scheme induced by a contaminated Pitman-Yor model.

Algorithm 1 Sampling scheme for contaminated Pitman-Yor model.

[0] Set initial values for ϑ , \bar{m}_1

for $r = 1, \dots, R$ **do**

[1] Update σ from

$$\mathcal{L}(\sigma \mid X_{1:n}, \vartheta, \bar{m}_1) \propto \mathcal{L}(\sigma) \sigma^{k-\bar{m}_1} \prod_{i=\bar{m}_1+1}^k (1-\sigma)_{n_i-1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)}$$

[2] Update ϑ from

$$\mathcal{L}(\vartheta \mid X_{1:n}, \sigma, \bar{m}_1) \propto \mathcal{L}(\vartheta) \frac{\Gamma(\vartheta) \Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma) \Gamma(\vartheta + n - \bar{m}_1)}$$

[3] Update β from $\mathcal{L}(\beta \mid X_{1:n}, \bar{m}_1) \propto \mathcal{L}(\beta) \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1}$

[4] Update \bar{m}_1 from

$$\mathcal{L}(\bar{m}_1 \mid X_{1:n}, \sigma, \vartheta, \beta) \propto \binom{m_1}{\bar{m}_1} \beta^{n-\bar{m}_1} (1-\beta)^{\bar{m}_1} \sigma^{k-\bar{m}_1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta + n - \bar{m}_1)}$$

We further notice that $\mathcal{L}(\beta) \stackrel{d}{=} \text{Beta}(a_\beta, b_\beta)$ is a conjugate prior distribution for β . We assume $\mathcal{L}(\sigma) \stackrel{d}{=} \text{Beta}(a_\sigma, b_\sigma)$ and $\mathcal{L}(\vartheta) \stackrel{d}{=} \text{Gamma}(a_\vartheta, b_\vartheta)$. We made the steps [1] and [2]

via Metropolis-Hastings with Gaussian proposal on a transformed scale $(\sigma, \vartheta) \rightarrow (\psi, \lambda)$, where $\psi = \log\left(\frac{\sigma}{1-\sigma}\right)$ and $\lambda = \log(\vartheta)$. The variances of the Gaussian proposals can be tuned to reach an optimal acceptance ratio for the Metropolis-Hastings steps (see [Roberts et al., 1997](#)). We can further initialize \bar{m}_1 with a vague strategy, by sampling uniformly on the integers from 0 to m_1 .

Appendix E: On the contaminated ZIPF distribution

In the simulation study, we have considered a contaminated ZIPF distribution. To obtain a sample from this distribution, we generate a fraction β of the labels of the species from a ZIPF distribution, with ζ total species and indexed by a parameter α , and we then contaminate the sample by a fraction $(1 - \beta)$ of singletons. The law of a ZIPF distribution with parameters α and total number of species ζ equals

$$P(X_i = t \mid \zeta, \alpha) = H_{\zeta, \alpha} \sum_{j=1}^{\zeta} \frac{1}{j^\alpha} \delta_j(t),$$

where $H_{\zeta, \alpha}$ is a normalization constant depending on ζ and α .

Appendix F: Simulation study with continuous data

We now move to the mixture scenario, by considering a set of continuous data taking values in \mathbb{R}^d . We simulate a set of data from a mixture of two Gaussian distributions with density $f(y) = 0.5\phi_d(y; -3, \text{diag}_d(1)) + 0.5\phi_d(y; 3, \text{diag}_d(1))$, where $\phi_d(\cdot; a, B)$ denotes the density of a d -dimensional Gaussian random variable with mean vector a and covariance matrix B , and $\text{diag}_d(b)$ denotes a diagonal matrix of dimension $d \times d$ with diagonal elements equal to b . After sampling m observations Y_1, \dots, Y_m , we then augment the sample with additional s outliers $cY_{m+1}, \dots, cY_{m+s}$ from an over-disperse truncated Gaussian distribution

$$\phi(y; 0, \text{diag}_d(3^2)) \mathbb{1}_{[(-\infty, -3\sqrt{\chi_d^2(0.9)}) \cup (3\sqrt{\chi_d^2(0.9)}, \infty)]} (\|y\|_2^2)$$

i.e. a multivariate Gaussian distribution with support \mathbb{R}^d minus the d -dimensional sphere of radius $3\sqrt{\chi_d^2(0.9)}$ centered at the origin, where $\chi_d^2(0.9)$ denotes the quantile of order 0.9 of a Chi-square distribution with d degrees of freedom, obtaining a sample of size $n = m + s$; the parameter c has the role to shrink or expand the nuisance observations towards the origin. We consider different simulated scenarios by selecting: $s = 10$, $m \in \{90, 240\}$, $d \in \{2, 4\}$ and $c \in \{1, 1.25, 1.5\}$ plus a case without outliers for each other setting. Figure 2 shows an example of simulate dataset for different values of the scaling constant c .

We face posterior inference with the mixture model described in Section 6, by considering a multivariate Gaussian kernel function with $y_i \mid \mu_i, \Sigma_i \sim N(\mu_i, \Sigma_i)$, and letting \tilde{p} distributed as a contaminated Pitman-Yor process. We assume P_0 and Q_0

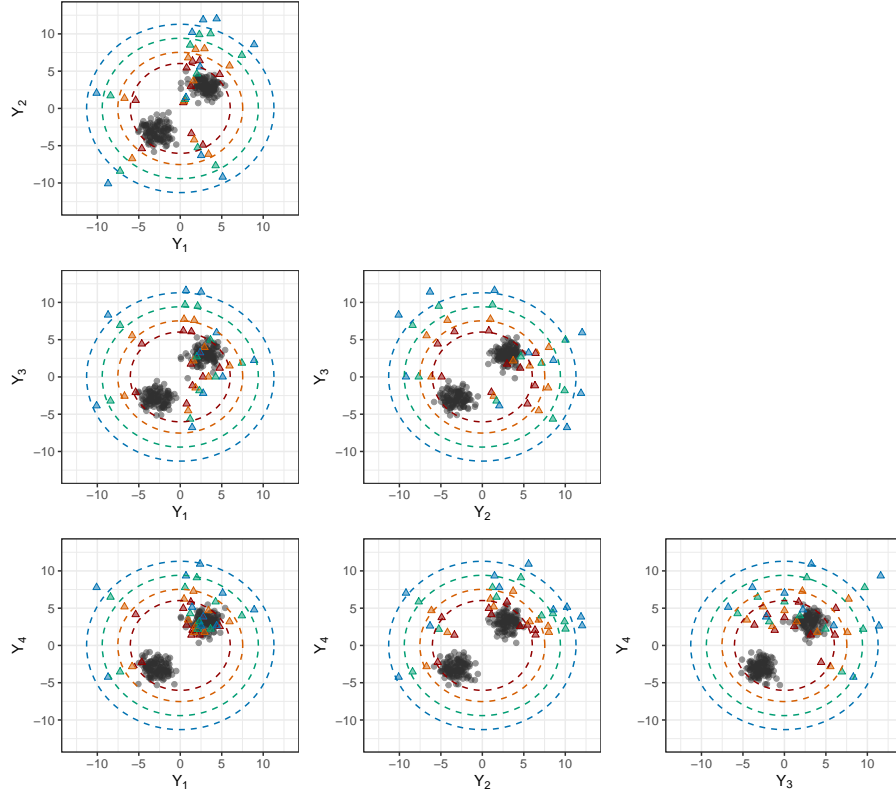


Figure 2: An example of simulated data set with dimension $d = 4$. Different colors correspond to different values of the scaling constant c , the dashed lines correspond to the hypersphere of radius equal to the square root of $3^2 \times \chi_4^2(0.9)$, the 0.9 quantile of a Chi-square distribution with 4 degrees of freedom, multiplied for different scaling constants.

distributed as Normal-Inverse-Wishart distributions, with $Q_0 \equiv \text{NIW}(\mu_0, \kappa_0, \nu_0, S_0)$ and $P_0 \equiv \text{NIW}(\mu_1, \kappa_1, \nu_1, S_1)$, where by $\text{NIW}(a, b, c, D)$ distribution for the parameters (μ, Σ) we mean that $\mu \mid \Sigma \sim N(a, \Sigma/b)$ and $\Sigma \sim \text{IW}(D, d)$. We set $\mu_0 = 0$, $\nu_0 = d + 3$, $\mu_1 = 0$, $\nu_1 = d + 3$, and $S_0 = S_1$ matching the diagonal of the sample variance of the data across all the scenarios. We consider two distinct specifications of the contaminated Pitman-Yor mixture model: i) CPY_1 , here the base measure is different from the contaminant measure, and we set $\kappa_0 = 1$, $\kappa_1 = 0.25$, so that we allow an over-disperse predictive distribution associated to the contaminant measure P_0 ; ii) CPY_2 , a scenario where the base measure and the diffuse component coincide, we assume $\kappa_0 = \kappa_1 = 0.5$. We further consider the Pitman-Yor mixture model (PY) with **base measure** $Q_0 \equiv \text{NIW}(\mu_0, \kappa_0, \nu_0, S_0)$ and $\mu_0 = 0$, $\kappa_0 = 1$, $\nu_0 = d + 3$, and S_0 equals the diagonal of the sample variance of the data. We finally consider the following vague prior specifications for the parameters: $\vartheta \sim \text{Gamma}(2, 0.02)$, $\sigma \sim \text{Unif}(0, 1)$, and for the

contaminated models $\beta \sim \text{Unif}(0, 1)$.

We applied the algorithm described in Section 6.1 to provide posterior estimates of the number of outliers out of the sample. In order to do this we have identified the optimal posterior point estimate of the latent partition of the data, which has been obtained minimizing the variation of information loss function among the sampled partitions (see Wade and Ghahramani, 2018; Rastelli and Friel, 2018). The MCMC procedure is based on 15 000 iterations, including a burn-in period of 5 000 iterations. The convergence of the chains was assessed by randomly checking some of the replications, without any evidence against it. Table 1 reports the number of singletons in the posterior point estimate of the latent partition of the data, while the true number of outliers equals 10.

| | | | $c = 1$ | $c = 1.25$ | $c = 1.5$ | no outliers |
|-----|-----|------------------|---------------|---------------|---------------|---------------|
| d | n | model | | | | |
| 2 | 100 | CPY ₁ | 9.10 (1.63) | 9.78 (1.34) | 9.63 (1.57) | 0.01 (0.10) |
| | | CPY ₂ | 4.58 (3.54) | 3.73 (3.81) | 2.50 (3.21) | 5.22 (20.47) |
| | | PY | 1.10 (9.99) | 0.09 (0.35) | 0.05 (0.29) | 0.00 (0.00) |
| | 250 | CPY ₁ | 9.67 (6.53) | 8.87 (2.59) | 8.36 (3.19) | 1.55 (2.99) |
| | | CPY ₂ | 8.16 (10.94) | 6.88 (10.29) | 10.32 (17.19) | 3.92 (8.60) |
| | | PY | 0.06 (0.23) | 0.03 (0.17) | 0.02 (0.14) | 0.01 (0.10) |
| | 4 | CPY ₁ | 10.68 (4.62) | 10.41 (2.12) | 10.89 (3.31) | 1.89 (1.76) |
| | | CPY ₂ | 8.46 (3.40) | 9.50 (4.99) | 9.09 (4.41) | 13.09 (21.50) |
| | | PY | 0.66 (1.87) | 0.54 (1.67) | 0.62 (1.67) | 0.00 (0.00) |
| | 250 | CPY ₁ | 14.26 (9.15) | 13.15 (5.71) | 13.07 (7.91) | 2.90 (7.55) |
| | | CPY ₂ | 17.36 (14.24) | 15.55 (12.93) | 16.65 (14.25) | 5.84 (12.19) |
| | | PY | 0.36 (1.31) | 0.55 (1.59) | 0.70 (1.94) | 0.00 (0.00) |

Table 1: Posterior mean (and standard deviation) of the number of singletons, averaged over 100 replications, detected with the two specifications of the contaminated Pitman-Yor mixture model and the Pitman-Yor mixture model.

We can appreciate how the Pitman-Yor mixture model lack in flexibility to estimate the number of outliers in a set of data. Among the contaminated models, the model with $P_0 \neq Q_0$, denote by CPY₁, shows overall estimates closer to the true number of contaminants compared to the case with $P_0 = Q_0$, with also a smaller uncertainty over different replications. On the counterpart, when the data does not contain any outlier, the pure Pitman-Yor mixture model shows better performance, but the contaminated Pitman-Yor mixture with inflated contaminant measure (i.e. CPY₁) performs really close to the pure case.

We can further study the accuracy of the density estimates and the predictive performance. To this ends, we consider symmetrised Kullback-Leibler divergence $S - KL(f_n, f_0) = KL(f_n | f_0) + KL(f_0 | f_n)$, where $KL(g | h) = \int_{\mathbf{x}} g(\mathbf{x}) \log(g(\mathbf{x})/h(\mathbf{x}))$, for the generic densities g, h , and the leave-one-out cross-validation criterion (LOO-CV) respectively. We consider as reference density f_0 the density function of the data generating process of the uncontaminated observations.

Table 2 reports the Symmetrized Kullback-Leibler divergence observed for the different settings of the simulation study. It is apparent that, in presence of outliers, the cPY

| d | n | model | $c = 1$ | $c = 1.25$ | $c = 1.5$ | no outliers |
|-----|-----|------------------|-----------------|-----------------|-----------------|-----------------|
| 2 | 100 | CPY ₁ | 0.0180 (0.0172) | 0.0202 (0.0213) | 0.0282 (0.0258) | 0.0097 (0.0051) |
| | | CPY ₂ | 0.0381 (0.0254) | 0.0435 (0.0309) | 0.0549 (0.0306) | 0.0342 (0.1068) |
| | | PY | 0.0348 (0.0178) | 0.0380 (0.0195) | 0.0421 (0.0245) | 0.0069 (0.0042) |
| | 250 | CPY ₁ | 0.0003 (0.0003) | 0.0002 (0.0003) | 0.0002 (0.0003) | 0.0013 (0.0018) |
| | | CPY ₂ | 0.0013 (0.0011) | 0.0011 (0.0008) | 0.0009 (0.0009) | 0.0118 (0.0194) |
| | | PY | 0.0010 (0.0008) | 0.0008 (0.0007) | 0.0008 (0.0008) | 0.0013 (0.0007) |
| 4 | 100 | CPY ₁ | 0.0127 (0.0051) | 0.0153 (0.0063) | 0.0183 (0.0073) | 0.0079 (0.0038) |
| | | CPY ₂ | 0.0205 (0.0049) | 0.0241 (0.0071) | 0.0271 (0.0079) | 0.0550 (0.0872) |
| | | PY | 0.0132 (0.0055) | 0.0149 (0.0065) | 0.0162 (0.0073) | 0.0028 (0.0015) |
| | 250 | CPY ₁ | 0.0004 (0.0002) | 0.0003 (0.0003) | 0.0003 (0.0002) | 0.0015 (0.0058) |
| | | CPY ₂ | 0.0007 (0.0004) | 0.0009 (0.0005) | 0.0009 (0.0005) | 0.0082 (0.0071) |
| | | PY | 0.0004 (0.0004) | 0.0003 (0.0002) | 0.0003 (0.0003) | 0.0006 (0.0003) |

Table 2: Symmetrized Kullback-Leibler divergence (lower is better) of the posterior density estimates and its standard deviation, averaged over 100 replications, computed with the two specifications of the contaminated Pitman-Yor mixture model and the Pitman-Yor mixture model.

mixture model with $P_0 \neq Q_0$ shows posterior density estimates closer to the density of the data generating process with respect to the PY mixture model, while in absence of outliers the PY mixture model performs slightly better. Regarding the cPY mixture model with $P_0 \equiv Q_0$, the accuracy of the estimates is always dominated by the other models, but still in the same decimal order while in presence of outliers.

| d | n | model | $c = 1$ | $c = 1.25$ | $c = 1.5$ | no outliers |
|-----|-----|------------------|-----------------|-----------------|-----------------|-----------------|
| 2 | 100 | CPY ₁ | 877.25 (15.60) | 894.09 (18.77) | 905.55 (18.24) | 805.67 (16.52) |
| | | CPY ₂ | 896.22 (19.99) | 912.09 (18.87) | 924.65 (19.09) | 792.64 (17.63) |
| | | PY | 844.93 (18.90) | 861.42 (19.53) | 871.25 (19.29) | 720.28 (18.15) |
| | 250 | CPY ₁ | 1627.21 (38.87) | 1649.82 (45.78) | 1661.41 (37.96) | 1426.92 (30.35) |
| | | CPY ₂ | 1572.73 (31.05) | 1604.61 (30.06) | 1626.91 (29.79) | 1412.75 (28.61) |
| | | PY | 1518.02 (30.58) | 1547.26 (31.35) | 1569.54 (30.86) | 1304.47 (30.12) |
| 4 | 100 | CPY ₁ | 772.82 (19.09) | 775.02 (17.98) | 773.46 (17.34) | 761.55 (16.98) |
| | | CPY ₂ | 779.93 (21.07) | 778.29 (18.55) | 779.90 (19.00) | 761.77 (18.71) |
| | | PY | 720.11 (17.51) | 720.89 (17.51) | 720.89 (17.57) | 712.18 (17.66) |
| | 250 | CPY ₁ | 1393.45 (38.53) | 1399.56 (39.63) | 1393.25 (38.07) | 1369.22 (26.83) |
| | | CPY ₂ | 1367.36 (30.88) | 1364.97 (28.24) | 1367.66 (27.44) | 1364.34 (28.03) |
| | | PY | 1293.55 (27.11) | 1293.60 (27.11) | 1293.56 (27.11) | 1283.71 (27.40) |

Table 3: Leave-one-out cross-validation criterion (lower is better) and its standard deviation, averaged over 100 replications, computed with the two specifications of the contaminated Pitman-Yor mixture model and the Pitman-Yor mixture model.

Table 3 reports the leave-one-out cross-validation criterion observed for the different

settings of the simulation study. In all the specifications considered in the study, the PY mixture model shows slightly better performance with respect to the cPY mixture models, but for all the models the LOO-CV ranges on similar values.

Appendix G: North America Ranidae dataset

To analyze the North America Ranidae dataset, we applied Algorithm 1 described in Section D. We have tuned the variances of the Metropolis-Hastings steps in order to attain optimal acceptance rates. We produced a raw chain of 35 000 iterations, which includes 10 000 burn-in iterations, and we thinned the chain every 10 realizations, obtaining a final sample of size 2 500. We report here the posterior summaries and the traceplots for the main parameters of the contaminated Pitman-Yor model and the Pitman-Yor model estimated with North America Ranidae data.

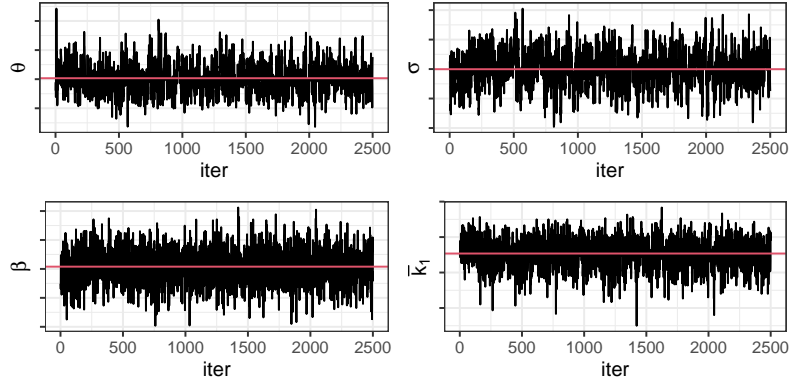


Figure 3: Traceplots for the contaminated Pitman-Yor model. Top-left panel: traceplot of ϑ . Top-right panel: traceplot of σ . Bottom-left panel: traceplot of β . Bottom-right panel: traceplot of \bar{m}_1 .

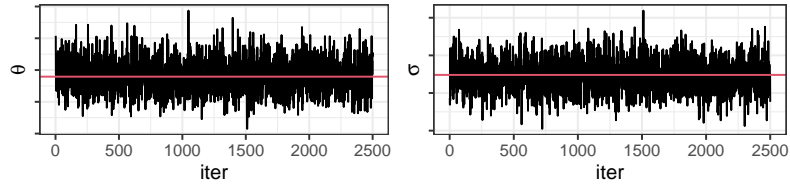


Figure 4: Traceplots for the Pitman-Yor model. Left panel: traceplot of ϑ . Right panel: traceplot of σ .

The traceplots reported in figure 4 and 3 show a good mixing of the chains, with random spikes and without systematic behavior. Tables 4 and 5 shows posterior summaries of the produced chains.

| | mean | SD | ESS | Geweke diagnostic |
|-------------|---------|---------|----------|-------------------|
| ϑ | 30.466 | 5.172 | 229.066 | 0.642 |
| σ | 0.099 | 0.031 | 254.958 | -0.612 |
| β | 0.998 | 0.00015 | 1186.791 | -0.520 |
| \bar{m}_1 | 226.196 | 12.373 | 524.699 | 1.115 |

Table 4: Posterior summaries of the chain sampled from the posterior distribution of the contaminated Pitman-Yor model with the north America ranidae dataset.

| | mean | SD | ESS | Geweke diagnostic |
|-------------|-------|--------|----------|-------------------|
| ϑ | 9.112 | 2.561 | 534.603 | 0.328 |
| σ | 0.339 | 0.0175 | 1141.642 | 0.872 |

Table 5: Posterior summaries of the chain sampled from the posterior distribution of the Pitman-Yor model with the north America ranidae dataset.

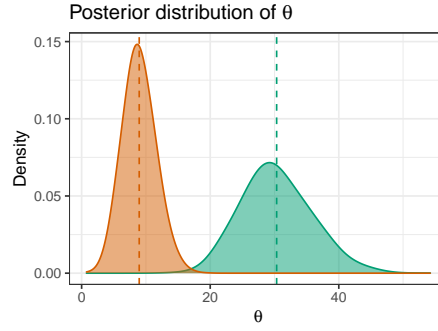


Figure 5: Posterior distributions of the parameter ϑ for the contaminated Pitman-Yor model (green) and the Pitman-Yor model (orange).

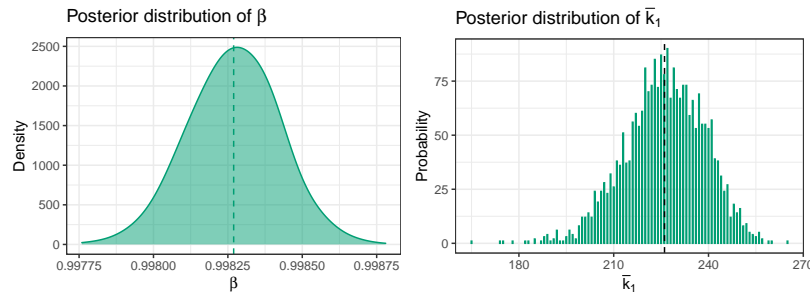


Figure 6: Traceplots for the Pitman-Yor model. Left panel: traceplot of ϑ . Right panel: traceplot of σ .

G.1 Inference on an additional sample

We validated the capability of the model of capturing an inflation of the number of species with frequency one in an additional sample. To this aim, we estimate the distribution of $E[N_{m,1}^{(n-m)}|X_1, \dots, X_{n-m}]$, $E[N_{m,2}^{(n-m)}|X_1, \dots, X_{n-m}]$ and $E[K_m^{(n-m)}|X_1, \dots, X_{n-m}]$ via cross-validation, by sampling without replacement the 80% of the observations $X_1^{\text{train}}, \dots, X_{n-m}^{\text{train}}$ to estimate the model, and then predicting the number of new species with frequency one in an additional sample, composed by the remaining 20% of the observations, $X_{n-m+1}^{\text{test}}, \dots, X_n^{\text{test}}$. We estimate both the contaminated Pitman-Yor model and the Pitman-Yor model with $X_1^{\text{train}}, \dots, X_{n-m}^{\text{train}}$, assuming the same model specification of Section 5.2, and by running the algorithm described in Section D for 10 000 iterations, of which 5 000 burn-in iterations, thinning the sampled chains every 5 realizations. We then evaluate $E[N_{m,1}^{(n-m)}|X_1, \dots, X_{n-m}]$, $E[N_{m,2}^{(n-m)}|X_1, \dots, X_{n-m}]$ and $E[K_m^{(n-m)}|X_1, \dots, X_{n-m}]$ for both the models. We further compute, for each sampled train and test data set, the observed number of new species with frequency equal to one, the observed number of new species with frequency equal to two, and the observed number of new species in the additional sample. We replicated the cross-validation for 1 000 times.

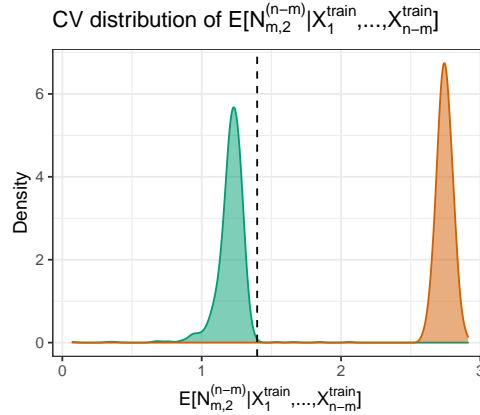


Figure 7: Cross-validated distribution of $E[N_{m,2}^{(n-m)}|X_1^{\text{train}}, \dots, X_{n-m}^{\text{train}}]$, for the contaminated Pitman-Yor model (green) and the Pitman-Yor model (orange). The black dashed line denotes the average of the observed number of species with frequency equal to two in the additional sample.

Figure 3 of the article shows the resulting distributions of $E[N_{m,1}^{(n-m)}|X_1^{\text{train}}, \dots, X_{n-m}^{\text{train}}]$ and $E[K_m^{(n-m)}|X_1^{\text{train}}, \dots, X_{n-m}^{\text{train}}]$, while Figure 7 show the resulting distributions of $E[N_{m,2}^{(n-m)}|X_1^{\text{train}}, \dots, X_{n-m}^{\text{train}}]$, for both the contaminated Pitman-Yor model and the Pitman-Yor model. We can appreciate that the distributions for the contaminated Pitman-Yor model shrink toward the observed values, while the Pitman-Yor model has a distortion in the produced estimates.

Appendix H: NGC 2419 data

In order to analyze the NGC 2419 dataset we applied the sampling strategy of Section 6.1. We have tuned the variances of the Metropolis-Hastings steps to update ϑ and σ to achieve optimal acceptance rates. We ran the model for 35 000 iterations, which include 10 000 burn-in iterations, thinning the produced chain every 10 realizations, and obtaining a final sample of size 2 500. We report here the posterior summaries and the traceplots for the main parameters of the contaminated Pitman-Yor mixture model estimated with the NGC 2419 data.

| | mean | SD | ESS | Geweke diagnostic |
|-------------|--------|-------|----------|-------------------|
| ϑ | 2.904 | 1.710 | 654.302 | -0.702 |
| σ | 0.116 | 0.032 | 302.821 | 1.373 |
| β | 0.925 | 0.034 | 1059.736 | -0.819 |
| k | 20.491 | 4.429 | 636.313 | 1.961 |

Table 6: Posterior summaries of the chain sampled for the contaminated Pitman-Yor mixture model. In evidence the main parameters of the mixing measure (ϑ , σ and β), and the number of different clusters k .

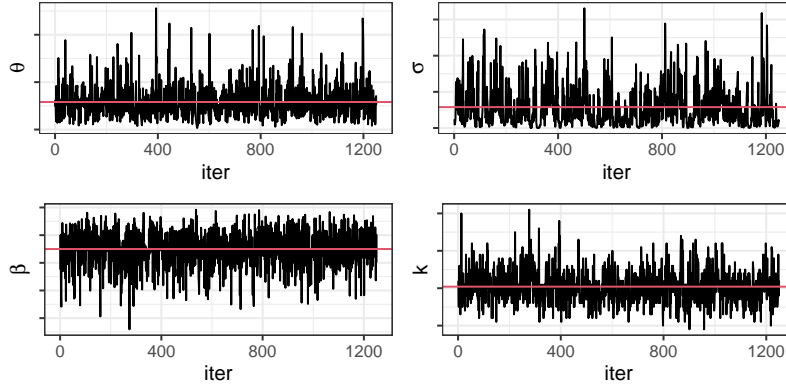


Figure 8: Traceplots for the contaminated Pitman-Yor mixture model. Top-left panel: traceplot of ϑ . Top-right panel: traceplot of σ . Bottom-left panel: traceplot of β . Bottom-right panel: traceplot of the number of distinct clusters k .

H.1 Comparison with Pitman-Yor mixture model

We compare the posterior inference faced for the contaminated Pitman-Yor mixture model, described in Section 6.2, with the standard Pitman-Yor mixture model case. We specify the Pitman-Yor mixture model preserving at most the commonalities with the contaminated model of Section 6.2, by considering the same multivariate Gaussian kernel function $\mathcal{K}(\cdot; (\boldsymbol{\mu}, \boldsymbol{\Sigma}))$, with expectation $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We further maintain the same specification of the base measure, i.e. $Q_0 \equiv NIW(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{S}_0)$ is a

Normal-Inverse-Wishart distribution. We specify the parameters of the base measure by setting μ_0 equals to the sample mean of the data, $\kappa_0 = 1$, $\nu_0 = d+3 = 7$ and S_0 equals to the diagonal of the sample variance of the data. We complete the model specification by choosing vague priors for the parameters of the mixing measure, $\vartheta \sim \text{Gamma}(2, 0.02)$, $\sigma \sim \text{Unif}(0, 1)$ and $\beta \sim \text{Unif}(0, 1)$. Posterior inference is carried out using a marginal sampling scheme (Escobar, 1988; Escobar and West, 1995), in the spirit of the algorithm described in section 6.1, but without the contaminant measure. We tune variances of the Metropolis-Hastings steps to update ϑ and σ to achieve optimal acceptance rates. We ran the model for 35 000 iterations, which include 10 000 burn-in iterations. The produced chain is thinned every 10 realizations, thus obtaining a final sample of size 2 500.

| | | CPY partition | | | | |
|--------------|--------------------------------|-------------------|-----------|------------|----------|----------|
| | | <i>Singletons</i> | <i>A</i> | <i>B</i> | <i>C</i> | <i>D</i> |
| | | <i>total</i> | <i>16</i> | <i>115</i> | <i>4</i> | <i>4</i> |
| PY partition | <i>Singletons_{PY}</i> | 7 | 5 | 2 | 0 | 0 |
| | <i>A_{PY}</i> | 113 | 3 | 109 | 0 | 1 |
| | <i>B_{PY}</i> | 15 | 7 | 1 | 4 | 3 |
| | <i>C_{PY}</i> | 2 | 0 | 2 | 0 | 0 |
| | <i>D_{PY}</i> | 2 | 1 | 1 | 0 | 0 |

Table 7: Comparison between the partition estimated using a contaminated Pitman-Yor mixture model and the optimal partition estimated using a Pitman-Yor mixture model.

Table 7 shows a comparison between the optimal partition estimated with a contaminated Pitman-Yor mixture model and the optimal partition estimated using a Pitman-Yor mixture model. Only 5 of the 16 stars identified as singletons in the contaminated mixture model are singletons also in the standard mixture model, while the optimal partition for the standard Pitman-Yor mixture model shows a total of 7 singletons. Out of 115 stars belonging to the main cluster in the contaminated model, 109 belong to the main cluster also in the standard model. The standard Pitman-Yor mixture model, with respect to the contaminated mixture model, is producing overall a partition with fewer singletons, and with the remaining stars less concentrated in the main cluster.

References

- Bacallado, S., Battiston, M., Favaro, S., and Trippa, L. (2017). “Sufficientness postulates for Gibbs-type priors and hierarchical generalizations.” *Statist. Sci.*, 32(4): 487–500. 14
- De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). “Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?” *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2): 212–229. 10
- Escobar, M. D. (1988). “Estimating the means of several normal populations by non-parametric estimation of the distribution of the means.” Ph.D. thesis, Department of Statistics, Yale University. 27

- Escobar, M. D. and West, M. (1995). “Bayesian density estimation and inference using mixtures.” *J. Amer. Statist. Assoc.*, 90(430): 577–588. [27](#)
- Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. (2009). “Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior.” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5): 993–1008. [12](#), [16](#)
- Favaro, S., Lijoi, A., and Prünster, I. (2013). “Conditional formulae for Gibbs-type exchangeable random partitions.” *Ann. Appl. Probab.*, 23(5): 1721–1754. [10](#), [12](#), [13](#), [15](#)
- Gnedin, A. and Pitman, J. (2005). “Exchangeable Gibbs partitions and Stirling triangles.” *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 12): 83–102, 244–245. [9](#)
- Lijoi, A., Mena, R. H., and Prünster, I. (2007). “Bayesian nonparametric estimation of the probability of discovering new species.” *Biometrika*, 94(4): 769–786. [9](#)
- Pitman, J. (2006). *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. [10](#)
- Rastelli, R. and Friel, N. (2018). “Optimal Bayesian estimators for latent variable cluster models.” *Stat. Comput.*, 28(6): 1169–1186. [21](#)
- Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). “Weak convergence and optimal scaling of random walk Metropolis algorithms.” *Ann. Appl. Probab.*, 7(1): 110–120. [19](#)
- Wade, S. and Ghahramani, Z. (2018). “Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion).” *Bayesian Analysis*, 13(2): 559 – 626. [21](#)