Electronic Journal of Statistics
Vol. 15 (2021) 5626–5651
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1933

Bayesian nonparametric disclosure risk assessment*

Stefano Favaro[†], **Francesca Panero and Tommaso Rigon**

Department of Economics and Statistics,
University of Torino and Collegio Carlo Alberto,
Corso Unione Sovietica 218/bis, 10134,
Torino, Italy
e-mail: stefano.favaro@unito.it

Department of Statistics,
University of Oxford,
24-29 St Giles', OX1 3LB,
Oxford, United Kingdom
e-mail: francesca.panero@stats.ox.ac.uk

Department of Economics, Management and Statistics,
University of Milano-Bicocca,
Piazza dell'Ateneo Nuovo 1, 20126,
Milano, Italy
e-mail: tommaso.rigon@unimib.it

Abstract: Any decision about the release of microdata for public use is supported by the estimation of measures of disclosure risk, the most popular being the number τ_1 of sample uniques that are also population uniques. In such a context, parametric and nonparametric partition-based models have been shown to have: i) the strength of leading to estimators of τ_1 with desirable features, including ease of implementation, computational efficiency and scalability to massive data; ii) the weakness of producing underestimates of τ_1 in realistic scenarios, with the underestimation getting worse as the tail behaviour of the empirical distribution of microdata gets heavier. To fix this underestimation phenomenon, we propose a Bayesian nonparametric partition-based model that can be tuned to the tail behaviour of the empirical distribution of microdata. Our model relies on the Pitman–Yor process prior, and it leads to a novel estimator of τ_1 with all the desirable features of partition-based estimators and that, in addition, allows to reduce underestimation by tuning a "discount" parameter. We show the effectiveness of our estimator through its application to synthetic data and real data.

MSC2020 subject classifications: Primary 62F15, 62G05.

Keywords and phrases: Bayesian nonparametrics, data confidentiality, Dirichlet process prior, disclosure risk assessment, empirical Bayes, Pitman–Yor process prior.

Received May 2021.

*Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

[†] Also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

Contents

1. Introduction

Releasing microdata for public use requires a careful assessment of the risk of disclosure (Willenborg and Waal [26]). Consider a microdata sample $(X_1, \dots, X_n)$ of units (individuals) from a finite population of size $N \geq n$, such that each X_i is a record containing identifying and sensitive information for the i-th unit. Identifying information consists of categorical variables which might match known units of the population. A threat of disclosure results from the possibility that an intruder, who could have personal or public information about the population (e.g. knowing who is included in the sample or using other available datasets), might succeed in identifying an individual through such a match, and hence be able to disclose sensitive information. To quantify disclosure risk, microdata units are partitioned according to a categorical variable that is defined by cross-classifying all identifying variables. That is, units X_i's are partitioned into non-empty cells, with each cell containing individuals with the same combination of values of identifying variables. A risk of disclosure arises from cells in which both sample and population frequencies are small, since the rarer the category the more likely the match is correct. Of special interest are cells with frequency 1 (uniques) since, assuming no errors in matching processes or data sources, for these cells the match is guaranteed to be correct (Bethlehem et al. [2], Skinner et al. [24]). This has motivated inferences on measures of disclosure risk that are functionals of the number of uniques, the most popular being the number τ_1 of sample uniques that are also population uniques. Once an estimate $\hat{\tau}_1$ of τ_1 is obtained, a criterion to understand if the data would incur an excessive risk in being published is to set a relative risk threshold C and check if the proportion of $\hat{\tau}_1$ with respect to the sample size does not exceed it, i.e. $\hat{\tau}_1/n \leq C$ (Bethlehem et al. [2]). If this is not the case, more care must be used before releasing data, possibly applying other privacy preserving methods.

Over the past three decades, a wide range of parametric and nonparametric approaches, both classical (frequentist) and Bayesian, have been proposed to estimate τ_1. One may identify two main streams in the disclosure risk literature: i) modeling the sole microdata partition by parametric and nonparametric partition-based models (Bethlehem et al. [2], Skinner et al. [24], Fienberg and Makov [11], Samuels [21], Skinner and Elliot [23], Camerlenghi et al. [6]); ii) modeling both the microdata partition and associations among identifying variables by parametric and semiparametric latent class models (Reiter [19], Skinner and Shlomo [25], Manrique-Vallier and Reiter [13, 14], Carota et al. [4, 5]). All these approaches have been applied to synthetic data and real data, showing the effectiveness of τ_1 as a sensible global measure for assessing the risk of disclosure. Partition-based models lead to estimators that are simple, linear in the sampling information, computationally efficient and scalable to massive data sets, though they typically show underestimation when the sampling fraction n/N becomes smaller than a certain threshold (Camerlenghi et al. [6]). Latent class models have typically a better empirical performance than partition-based models, especially for small sampling fractions, though this is achieved at the cost of an increased computational effort for the need of Markov chain Monte Carlo methods for posterior approximation (Reiter [19], Manrique-Vallier and Reiter [13]).

In this paper, we contribute to the partition-based literature from a Bayesian nonparametric perspective. Bayesian nonparametric ideas for estimating τ_1 date back to the seminal work of Samuels [21], where the Dirichlet process (Ferguson [10]) was applied as a prior model for the microdata partition. This approach leads to an estimator of τ_1 which is easy to implement, computationally efficient, and scalable to massive data. Despite these desirable features, empirical analyses in Samuels [21] show that such an approach underestimates τ_1 in many realistic scenarios, the issue being related to the tail behaviour of the empirical distribution of microdata. That is, the heavier the tail the worse the underestimation of τ_1. As heavy-tail scenarios occur when the number of sample uniques is large with respect to the population size, this phenomenon is a critical concern in disclosure risk assessment. A simulation study in Figure 1 shows analogous estimation issues for the most common partition-based estimators of τ_1 in such a heavy-tails setting. Our experiments use synthetic microdata from a power-law distribution of exponent $\sigma > 1$, samples being the 10% of the population of size 10^6, and they are averaged over 1000 iterations. It emerges that the smaller σ, namely the heavier the tail, the worse the underestimation of Bayesian parametric estimators (Bethlehem et al. [2], Skinner et al. [24]), and the worse the overestimation of a nonparametric empirical Bayes estimator (Camerlenghi et al. [6]).

To overcome the underestimation phenomenon of Samuels' approach, we propose a Bayesian nonparametric partition-based model that can be tuned to the

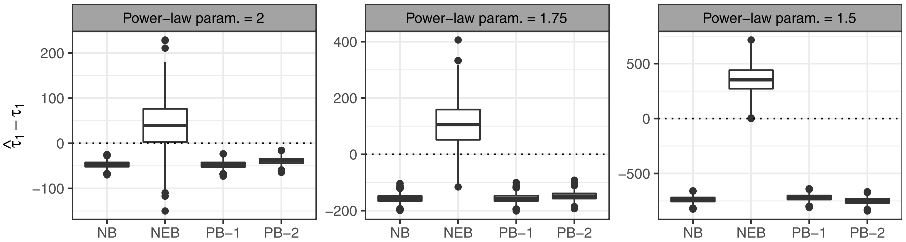

FIG 1. *Empirical performance, with respect to the true τ_1, of estimators $\hat{\tau}_1$: nonparametric Bayes (NB) of Samuels [21], nonparametric empirical Bayes (NEB) of Camerlenghi et al. [6], parametric Bayes (PB-1) of Bethlehem et al. [2], parametric Bayes (PB-2) of Skinner et al. [24].*

tail behaviour of the empirical distribution of microdata. In particular, as a prior model for the microdata partition, we assume the Pitman–Yor process (Perman et al. [15], Pitman [16], Pitman and Yor [18]). The Pitman–Yor process prior generalizes the Dirichlet process prior by means of an additional "discount" parameter that allows to control the tail behaviour of the prior, ranging from geometric tails to heavy power-law tails (Pitman and Yor [18]). Under the Pitman–Yor process prior, we present a simple characterization of the posterior distribution of τ_1, given the observed microdata, and we propose the posterior mean as a Bayesian nonparametric estimator of τ_1. Such an estimator has all the same desirable features as Samuels's estimator and, in addition, it allows to reduce its underestimation of τ_1 by tuning the "discount" parameter with respect to observable microdata. Our approach stands out for being the first partition-based approach to provide a closed-form posterior distribution of τ_1, which makes straightforward to quantify uncertainty of our Bayesian procedure through credible intervals. We investigate the empirical performance of our approach through synthetic data and real data from the 2018 American Community Survey, showing its effectiveness in reducing underestimation phenomenon of Samuels' approach.

The paper is structured as follows. In Section 2 we introduce the Pitman–Yor process prior and its sampling structure, and present our Bayesian nonparametric approach to infer τ_1. Section 3 contains an illustration of the proposed approach through synthetic data and real data. In Section 4 we conclude by discussing our results and directions for future work. Proofs are deferred to the Appendix.

2. Bayesian nonparametric inference for τ_1

We consider a super-population of units belonging to an (ideally) infinite number of distinct symbols $(z_j)_{j\geq 1}$, taking values in a measurable space $\mathbb{Z}$, with unknown proportions $(p_j)_{j\geq 1}$ such that $\sum_{j\geq 1} p_j = 1$. The partition of microdata into non-empty cells, both at the sample and population level, is modeled as a random partition induced by sampling from the unknown discrete distribution $P = \sum_{j \geq 1} p_j \delta_{z_j}$, where each symbol $z_j \in \mathbb{Z}$ takes the interpretation of a distinct combination of values of identifying variables. That is, a population of $N \geq 1$ of microdata units is assumed to be a random sample $(X_1, \ldots, X_N)$ from P, of which the first $n < N$ elements $(X_1, \ldots, X_n)$ are observable. These samples induce a random partition at population level consisting of K_N cells of frequencies $(N_{1,N}, \ldots, N_{K_N,N})$, and a random partition at the sample level consisting of K_n cells of frequencies $(N_{1,n}, \ldots, N_{K_n,n})$. If $I(\cdot)$ denotes the indicator function, then

$$\tau_1 = \sum_{i=1}^{K_n} I(N_{i,n} = 1) I(N_{i,N} = 1),$$

namely the number of sample uniques that are also population uniques (Bethlehem et al. [2], Skinner et al. [24]). Bayesian nonparametric inference for τ_1 relies on the specification of a (nonparametric) prior distribution on the discrete distribution P, which in turn leads to a prior model for the microdata partition.

2.1. The Pitman–Yor process prior

We assume the Pitman–Yor process as a prior model for the unknown discrete distribution P. A simple and intuitive definition of the Pitman–Yor process follows from its stick-breaking construction (Pitman [16]). For $\alpha \in [0, 1)$ and $\theta > -\alpha$ let: i) $(V_i)_{i \geq 1}$ be independent random variables such that V_i is distributed as a Beta distribution with parameter $(1 - \alpha, \theta + i\alpha)$; ii) $(Z_j)_{j \geq 1}$ be random variables, independent of the V_i's, and independent and identically distributed as a non-atomic distribution ν on $\mathbb{Z}$. If we set $p_1 = V_1$ and $p_j = V_j \prod_{1 \leq i \leq j-1}(1 - V_i)$ for $j \geq 2$, which ensures that $\sum_{j \geq 1} p_j = 1$ almost surely, then $P_{\alpha,\theta} = \sum_{j \geq 1} p_j \delta_{Z_j}$ is a Pitman–Yor process on $\mathbb{Z}$ with "discount" α and scale θ. The Dirichlet process arises as a special case by letting $\alpha = 0$. The Pitman–Yor process generalizes the Dirichlet process by means of the "discount" α, which controls the tail behaviour of $P_{\alpha,\theta}$, ranging from geometric tails to heavy power-law tails. In particular, for $\alpha \in (0, 1)$, let $(p_{(j)})_{j \geq 1}$ be the random probabilities p_j's of $P_{\alpha,\theta}$ in decreasing order. Then, as $j \to +\infty$ the $p_{(j)}$'s follow a power-law distribution of exponent $\sigma = \alpha^{-1}$ (Pitman and Yor [18]). This shows that $\alpha \in (0, 1)$ tunes the power-law tail behaviour of $P_{\alpha,\theta}$ through small probabilities $p_{(j)}$'s: the larger α the heavier the tail of $P_{\alpha,\theta}$, whereas a geometric tail arises as $\alpha \to 0$.

According to de Finetti's representation theorem, a random sample from $P_{\alpha,\theta}$ is part of an exchangeable sequence of $\mathbb{Z}$-valued random variables $(X_i)_{i \geq 1}$ whose directing measure Π is the law of $P_{\alpha,\theta}$. Let $(X_1, \ldots, X_n)$ be a random sample from $P_{\alpha,\theta}$, i.e.

$$\begin{aligned} X_i \mid P_{\alpha,\theta} &\overset{\text{iid}}{\sim} P_{\alpha,\theta} \qquad i = 1, \ldots, n, \\ P_{\alpha,\theta} &\sim \Pi. \end{aligned} \tag{1}$$

Because of the discreteness of $P_{\alpha,\theta}$, the sample $(X_1,\dots,X_n)$ induces a random partition of $\{1,\dots,n\}$ into $K_n \leq n$ blocks, labelled by distinct symbols $\{Z^*_1,\dots,Z^*_{K_n}\}$, with frequencies $(N_{1,n},\dots,N_{K_n,n}) = (n_1,\dots,n_k)$ such that $N_{i,n} \geq 1$ for $i = 1,\dots,K_n$ and $\sum_{1\leq i\leq K_n} N_{i,n} = n$ (Pitman [Chapter 3, 17]) for a detailed account. A generative model for the X_i's, and hence for the induced random partition, is provided by the predictive distribution of the Pitman–Yor process, namely

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1,\dots,X_n) = \frac{\theta + k\alpha}{\theta + n}\nu(\cdot) + \frac{1}{\theta+n}\sum_{i=1}^{k}(n_i - \alpha)\delta_{Z^*_j}(\cdot), \qquad (2)$$

for $n \geq 1$. That is, X_{n+1} is of a new symbol (block), namely a symbol not observed in the set $\{Z^*_1,\dots,Z^*_{K_n}\}$, with probability $(\theta + k\alpha)/(\theta + n)$, or X_{n+1} is of symbol (block) Z^*_i with probability $(n_i - \alpha)/(\theta + n)$, for $i = 1,\dots,k$. See Pitman [Chapter 3, 17] for a detailed account on the predictive distribution (2).

The predictive distribution of the Pitman–Yor process highlights the role of the "discount" parameter α in the sampling process: it drives a combined effect in terms of a reinforcement mechanism and the increase in the rate of generating new symbols. In particular, a new symbol z^* entering in the sample produces two effects: i) it is assigned a mass proportional to $(1-\alpha)$ to the z^*'s empirical component of (2); ii) it is assigned a mass proportional to α to the probability of generating new symbols in (2). That is, the probability mass assigned to the symbol z^*'s is less than proportional to 1, and the remaining probability mass is assigned to the probability of generating new symbols. The first effect gives rise to a reinforcement mechanism: the sampling procedure allocates more mass on symbols with higher frequencies. The second effect implies that the probability of generating new symbols, which overall still decreases as a function of n, is increased by $\alpha/(\theta + n + 1)$. The larger α the stronger the reinforcement mechanism and the higher is the probability of new symbols. For $\alpha = 0$, that is under the Dirichlet process prior, everything is proportional to symbols' frequencies, which do not alter the probability of discovering new symbols. We refer to Bacallado et al. [1] for a detailed account on the predictive distribution (2), as well a generalizations thereof, and for characterizations of (2) with respect to the use of the sampling information, i.e. "sufficientness postulate", and of Pólya like urn schemes.

Remark 1. The power-law tail behaviour of the Pitman–Yor process emerges from the large n asymptotic behaviour of the number K_n of distinct symbols and the number $M_{r,n}$ of distinct symbols with frequency $r \geq 1$ in n random samples from $P_{\alpha,\theta}$. From Pitman [17, Theorem 3.8], K_n behaves as n^α for large n; this is the behaviour of the number of distinct symbols in n random samples from a power-law distribution of exponent $\sigma = \alpha^{-1}$. Moreover, from Pitman [17, Lemma 3.11] it holds that the proportion $M_{r,n}/K_n$ of distinct symbols with frequency r behaves as $r^{-\alpha-1}$ for large n and large r; this is, up to a constant or proportionality, the distribution of the number of distinct symbols with frequency r in n random samples from a power-law distribution of exponent $\sigma = \alpha^{-1}$.

2.2. Posterior inference for τ_1

We consider microdata units to be modeled under the Bayesian nonparametric framework (1). That is, a population of $N \geq 1$ of microdata units is assumed to be a random sample $(X_1, \ldots, X_N)$ from a Pitman–Yor process, of which the first $n < N$ elements $(X_1, \ldots, X_n)$ are observable. We characterize the posterior distribution of τ_1, given $(X_1, \ldots, X_n)$. To introduce our main result, it is useful to recall the generalized factorial distribution (Charalambides [7, Chapter 2]). For a real a and $r \in \mathbb{N}$ let $(a)_{(r)}$ be the rising factorial, that is $(a)_{(0)} = 1$ and $(a)_{(r)} = \prod_{0 \leq i \leq r-1}(a+i)$ for $r \in \mathbb{N} \smallsetminus \{0\}$, and for $a > 0$ and $r, s \in \mathbb{N}$ with $r \leq s$ let $\mathscr{C}(r, s; a)$ be the generalized factorial coefficient (Charalambides [7]), that is $\mathscr{C}(r, s; a) = \sum_{0 \leq i \leq s}(-1)^i \{i!(s-i)!\}^{-1}(-ia)_{(r)}$. For $r \in \mathbb{N}$, $b \in [0,1]$ and $c > 0$, a random variable $U_{b,c,r}$ on $\{1, \ldots, r\}$ has a generalized factorial distribution if, for $x \in \{1, \ldots, r\}$

$$\mathbb{P}(U_{b,c,r} = x) = \frac{1}{(bc)_{(r)}} \mathscr{C}(r, x; b)(c)_{(x)}. \tag{3}$$

The next theorem provides the posterior distribution of τ_1, given $(X_1, \ldots, X_n)$, as a mixture of a (general) hypergeometric distribution (Johnson et al. [12, Chapter 6.2.5]) with respect to the generalized factorial distribution displayed in (3). Then, a Bayesian nonparametric estimator of τ_1 is given as the posterior mean.

Theorem 1. *For $N \geq 1$ let $(X_1, \ldots, X_N)$ be a random sample from $P_{\alpha,\theta}$, of which the first $n < N$ elements $(X_1, \ldots, X_n)$ are observable and featuring $M_{1,n} = m_1$ distinct symbols with frequency 1 (sample uniques). Then, for $x \in \{0, 1, \ldots, m_1\}$*

$$\mathbb{P}(\tau_1 = x \mid X_1, \ldots, X_n) = \sum_{u=1}^{N-n} \frac{\binom{\frac{\theta+n}{1-\alpha}-1}{x}\binom{u}{m_1-x}}{\binom{\frac{\theta+n}{1-\alpha}-1+u}{m_1}} \mathbb{P}(U_{1-\alpha, \frac{\theta+n}{1-\alpha}, N-n} = u), \tag{4}$$

and

$$\hat{\tau}_1 = \mathbb{E}(\tau_1 \mid X_1, \ldots, X_n) = m_1 \frac{(\theta + \alpha + n - 1)_{(N-n)}}{(\theta + n)_{(N-n)}}. \tag{5}$$

See Appendix A for the proof of Theorem 1. Theorem 1 is the first example in the literature to provide a closed-form posterior distribution of τ_1. This is critical to quantify, by means of Monte Carlo sampling, uncertainty of our Bayesian procedure through credible intervals; see Section 2.3 below. According to (4), for any fixed (α, θ), the number $M_{1,n} = m_1$ of sample uniques is sufficient for estimating τ_1. The estimator (5) is easy to implement, computationally efficient, and scalable to massive datasets. Moreover, it has a simple interpretation as the proportion

$$w_{n,N}(\alpha,\theta) = \frac{(\theta+n-1+\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}} \in (0,1),$$

of the number m_1 of sample uniques. The estimator (5) is somehow reminiscent of the "naive" nonparametric estimator (Bethlehem et al. [2], Skinner and Elliot [23]) of τ_1, namely

$$\bar{\tau}_1 = m_1 \frac{n}{N}.$$

In particular, $\hat{\tau}_1$ is a smoothed version of $\bar{\tau}_1$, where the smoothing acts by replacing the purely empirical proportion n/N with the parametric proportion $w_{n,N}(\alpha,\theta)$. For any fixed θ, n and N, the proportion $w_{n,N}(\alpha,\theta)$ increases in α, meaning that the larger α the higher $\hat{\tau}_1$. This behaviour, which agrees with the role of α discussed in Section 2.1, shows the effectiveness of the "discount" α in tuning the inference to the tail behaviour of the empirical distribution of microdata.

Remark 2. For $\alpha = 0$, namely under the Dirichlet process prior, Theorem 1 simplifies remarkably. In particular, the posterior distribution (4) reduces to a (general) hypergeometric distribution. That is, by setting $\alpha = 0$, Equation (4) reduces to

$$\mathbb{P}(\tau_1 = x \,|\, X_1,\dots,X_n) = \frac{\binom{\theta+n-1}{x}\binom{N-n}{m_1-x}}{\binom{\theta+N-1}{m_1}}. \tag{6}$$

for $x \in \{0,1,\dots,m_1\}$. Moreover, by setting $\alpha = 0$, Equation (5) reduces to the estimator of Samuels [21], namely $\hat{\tau}_1 = m_1(\theta+n-1)/(\theta+N-1)$. Equation (4) thus completes the work of Samuels [21], where only the estimator $\hat{\tau}_1$ was provided.

By assuming both the sample and population to be large, it emerges: i) the critical influence of the "discount" α in estimating τ_1, with respect to the scale θ; ii) the crucial limitation of the estimator proposed in Samuels [21]. In particular, let $f \approx g$ meaning $f/g \to 1$. As $n, N \to +\infty$ with $n < N$, for any $x \in \{0,1,\dots,m_1\}$

$$\mathbb{P}(\tau_1 = x \,|\, X_1,\dots,X_n) \approx \binom{m_1}{x}\left\{\left(\frac{n}{N}\right)^{1-\alpha}\right\}^{x}\left\{1-\left(\frac{n}{N}\right)^{1-\alpha}\right\}^{m_1-x}, \tag{7}$$

and hence

$$\hat{\tau}_1 \approx m_1\left(\frac{n}{N}\right)^{1-\alpha}. \tag{8}$$

That is, for large n and N with $n < N$, the posterior distribution (4) admits a first order (local) approximation in terms of a Binomial distribution with parameters $\{m_1, (n/N)^{1-\alpha}\}$. See Appendix B for the proof of (7). This result shows that, in realistic scenarios, the "discount" α is the sole tuning parameter of our Bayesian nonparametric model. In other terms, for $\alpha = 0$, namely under the Dirichlet process prior, the approximated estimator (8) reduces to the "naive" estimator $\bar{\tau}_1$. Equivalently, for large n and N, the "naive" estimator $\bar{\tau}_1$ approximates the estimator of Samuels [21]. Therefore, in realistic scenarios, Samuel's estimator is a purely empirical estimator, meaning that no tuning parameters are available.

2.3. Computations

For any fixed $\alpha \in (0,1)$ and $\theta > -\alpha$, the estimator (5) can be easily evaluated for arbitrary values of n and N. Instead, the evaluation of the posterior distribution (4) might be numerically unstable for large n and N, due to the overwhelming computational burden for evaluating generalized factorial coefficients. To address this issue, we rely on Monte Carlo sampling of the posterior distribution (4) to obtain credible intervals for the estimator (5). By the mixture representation of (4), Monte Carlo sampling requires to sample from a (general) hypergeometric distribution and from a generalized factorial distribution. The former is straightforward, for arbitrary values of n and N, and routines are available in standard software. The latter becomes easy upon noticing that it coincides with the distribution of the number K_{N-n} of distinct symbols in $N-n$ random samples from a Pitman–Yor process with "discount" $(1-\alpha)$ and scale $(\theta+n)$. See Appendix C for a detailed explanation. For arbitrary values of n and N, Monte Carlo sampling of the distribution of K_{N-n} is straightforward by Algorithm 1, which exploits the predictive distribution (2) of the Pitman–Yor process.

```
Set $k = 1$;
for $i = 1$ to $N - n - 1$ do
    Sample a binary variable $s$ with probability $\{\theta + n + (1-\alpha)k\}/(\theta + n + i)$;
    Set $k \leftarrow k + s$;
end
Return $k$.
```

Algorithm 1: Monte Carlo sampling of the mixing generalized factorial distribution.

To implement Theorem 1 we must specify the prior's parameters (α, θ), whose choice is critical for a correct estimation of τ_1. Two common approaches for estimating (α, θ) are: i) the hierarchical Bayes approach, which relies on Bayesian estimates obtained from the posterior distribution of (α, θ) with respect to suitable prior specification; ii) the empirical Bayes approach, which relies on estimates obtained by maximizing, with respect to (α, θ), the marginal likelihood of the observable sample. Here, we adopt the empirical Bayes approach. Let $(X_1, \ldots, X_n)$ feature $K_n = k$ distinct symbols with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$. Pitman [16, Proposition 9] provides the likelihood function of $(X_1, \ldots, X_n)$, and the empirical Bayes approach reduces to

Table 1

Estimates of τ_1 for synthetic data. The parameters are σ (Zipf data) and π (Geometric data). PB*-1 is parametric Bayes of Bethlehem et al. [2];* PB*-2 is parametric Bayes of Skinner et al. [24], and* NEB *is nonparametric empirical Bayes of Camerlenghi et al. [6].*

DATA	m_1	τ_1	PITMAN-YOR	DIRICHLET PR.	PB-1	PB-2	NEB
SCENARIO I: $N = 10^6, n = 10^5$							
Zipf 1.25	10818	6914	6818 [6689, 6947]	1123 [1042, 1203]	1543	946	8328
Zipf 1.50	2045	941	948 [890, 1006]	206 [171, 241]	224	194	1403
Zipf 1.75	557	205	203 [174, 232]	56 [38, 75]	58	66	283
Zipf 2.00	230	80	74 [56, 93]	23 [12, 35]	22	30	198
Geom. 10^{-4}	9938	1027	1113 [1034, 1195]	1113 [1034, 1195]	4666	2095	740
Geom. 10^{-3}	949	91	96 [73, 120]	96 [73, 120]	335	167	67
SCENARIO II: $N = 5000, n = 500$							
Zipf 1.25	139	76	82 [67, 96]	16 [7, 27]	34	20	120
Zipf 1.50	62	23	28 [18, 38]	7 [1, 13]	12	7	51
Zipf 1.75	28	7	10 [4, 17]	3 [0, 8]	5	3	22
Zipf 2.00	11	3	3 [0, 7]	1 [0, 4]	2	1	6
Geom. 10^{-4}	482	391	365 [341, 388]	365 [341, 388]	181	196	467
Geom. 10^{-3}	387	95	129 [106, 153]	129 [106, 153]	160	158	320

solve:

$$(\hat{\alpha}, \hat{\theta}) = \underset{(\alpha,\theta)}{\arg\max} \left\{ \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)}{(\theta)_{(n)}} \prod_{i=1}^{k} (1-\alpha)_{(n_i - 1)} \right\}. \tag{9}$$

The optimization problem (9) can be solved numerically and efficiently even for large values of n, by means of routines available in standard softwares. We refer to Favaro and Naulet [9] for provable guarantees of the estimator $\hat{\alpha}$. Alternatively, one could specify a prior distribution on (α, θ). However, we found no relevant differences between the fully Bayes and the empirical Bayes approach, given that the posterior distribution of (α, θ) is highly concentrated, when n is large.

3. Illustrations

3.1. Simulated data

We consider synthetic data from two super-populations P. For the first super-population, we let the "true" probability masses $(p_j)_{j\geq 1}$ to be those of a Zipf distribution with index $\sigma > 1$, so that data are generated from the discrete distribution $P = \zeta(\sigma)^{-1} \sum_{j\geq 1} j^{-\sigma}\delta_{z_j}$, with $\zeta(\sigma) = \sum_{j\geq 1} j^{-\sigma}$. As we discussed in Section 2, this is the scenario in which a Pitman–Yor specification is recommended. We considered different values of $\sigma = 1.25, 1.50, 1.75, 2$, and different combinations of n and N. The prior's parameter (α, θ) is estimated through maximum likelihood; see Section 2.3. Table 1 reports estimates of τ_1, together with 99% credible intervals (within brackets), and the "true" value of τ_1. Credible intervals are obtained via Monte Carlo sampling of the posterior distribution

Table 2

Maximum likelihood estimate for the parameter (α, θ) of the Pitman–Yor model.

Param.	Zipf 1.25	Zipf 1.50	Zipf 1.75	Zipf 2.00	Geom. 10^{-4}	Geom. 10^{-4}
SCENARIO I: $N = 10^6, n = 10^5$						
$\hat{\alpha}$	0.80	0.67	0.56	0.51	0	0
$\hat{\theta}$	1.48	0.82	0.70	0.34	13559.80	1141.16
SCENARIO II: $N = 5000, n = 500$						
$\hat{\alpha}$	0.77	0.66	0.57	0.39	0	0
$\hat{\theta}$	1.89	0.98	0.52	0.90	13529.12	1753.06

(1), by means of the scheme described in Section 2.3. Table 2 reports the corresponding estimates of (α, θ) for the Pitman–Yor model. In all these scenarios, the Bayesian nonparametric estimator (5) is much closer to the "true" value of τ_1, compared to its partition-based competitors. In particular, the approaches of Bethlehem et al. [2], Skinner et al. [24] and Samuels [21] underestimate the "true" τ_1, whereas the approach of Camerlenghi et al. [6] tends to overestimate it.

For the second super-population, we let $P = \sum_{j \geq 1} \pi(1 - \pi)^{j-1} \delta_{z_j}$, corresponding to a geometric distribution with parameter $\pi \in (0, 1)$. We consider two different values of $\pi = 10^{-3}, 10^{-4}$ and the same sample size n and a population size N as before. As we discussed in Section 2, this is the ideal setting for the Dirichlet process and this is indeed confirmed by Table 1. Moreover, the Pitman–Yor estimator reduces to the Dirichlet process since we obtain $\hat{\alpha} = 0$, as reported in Table 2.

3.2. The 2018 American Community Survey

We consider real data from the 2018 American Community Survey (Manrique-Vallier and Reiter [13], Carota et al. [4]). This dataset is a random sample of the American population (`usa.ipums.org/usa`). We regard the 2018 American Community data as a "population" of size $N = 2,432,323$, and we consider observable samples which are the 5% and 10% fractions of the population obtained by sampling at random $n = 121,616$ and $n = 243,232$ individual, respectively. We restricted the population to individuals older than 20, and we cross-classified the records according to the following variables: census region (9 levels), race (139 levels), and primary occupation (531 levels), obtaining $K_N = 60,215$ non empty classes.

As detailed in Section 2, the Pitman–Yor specification should be employed whenever the data follow a power-law behaviour. However, in real data problems such an assumption must be empirically validated. A simple approach is comparing the observed number m_r of distinct types with frequency $r = 1, \ldots, n$ against the model-based expected frequencies under a Pitman–Yor specification,

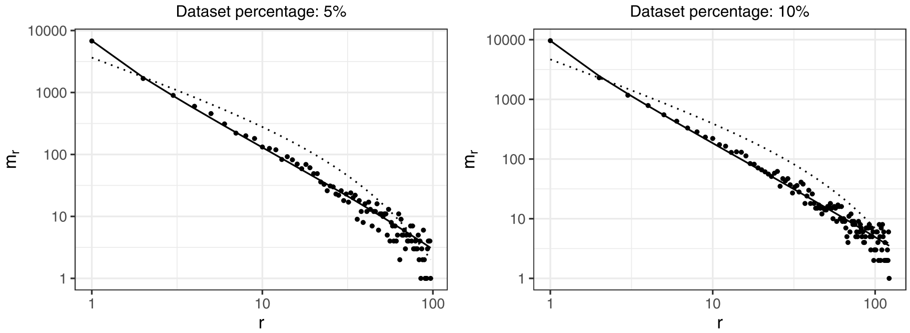

Fig 2. *Graphical representation in logarithmic scale of the number of distinct types m_r with frequency r (dots) and their expectations $\mathbb{E}(M_{r,n})$ under Pitman–Yor (solid line) and Dirichlet process (dotted line) models, relative to the 5% and 10% sample data from the American Community Survey.*

namely

$$\mathbb{E}(M_{r,n}) = \frac{\theta}{(\theta)_{(n)}}\binom{n}{r}(1-\alpha)_{(r-1)}(\theta+\alpha)_{(n-r)}, \qquad r = 1,\ldots,n,$$

where the parameters in the above formula are replaced by their maximum likelihood estimates; see also Favaro et al. [8] for further details. Poor in-sample fit strongly suggests that the corresponding disclosure risk assessment will be unreliable.

The observed values m_r for $r = 1,\ldots,n$ and their model-based estimates for the 5% and 10% fractions of the data from the American Community Survey presented in Section 2.2 are reported in Figure 2, both under a Pitman–Yor and Dirichlet process specification. These results confirm a very good in-sample fit for the Pitman–Yor. Conversely, the Dirichlet process seems unsuitable for this specific datasets. The prior's parameters α and θ are estimated through maximum likelihood; see Section 2.3. Results in Table 3 confirm what we observed for synthetic data, and in particular it is confirmed the superior empirical performance of our estimators, with respect to partition-based competitors. The approaches of Bethlehem et al. [2], Skinner et al. [24] and Samuels [21] underestimate the true τ_1, whereas the approach of Camerlenghi et al. [6] overestimates it.

Table 3

Estimates of τ_1 for real data the 2018 American Community Survey. The estimate PB*-1 refers to the parametric Bayes of Bethlehem et al. [2],* PB*-2 is the parametric Bayes of Skinner et al. [24], and* NEB *is the nonparametric empirical Bayes of Camerlenghi et al. [6].*

Data percentage	m_1	τ_1	Pitman–Yor	Dirichlet process	PB-1	PB-2	NEB
5%	6776	1447	1458 [1372, 1546]	349 [303, 397]	1427	425	3492
10%	9620	2852	2958 [2842, 3075]	979 [903, 1056]	1799	1059	4526

4. Discussion

In this paper, we considered the problem of Bayesian nonparametric estimation of τ_1, which is arguably the most popular measure of disclosure risk. Our study is motivated by an early work of Samuels [21], where empirical analyses showed that the use of Dirichlet process priors lead to underestimate τ_1 in many realistic scenario, with the underestimation getting worse as the tail behaviour of the empirical distribution of microdata gets heavier. Here, to overcome such an underestimation phenomenon, we proposed the use of the Pitman–Yor process prior, which generalizes the Dirichlet process prior through an additional "discount" parameter that allows to control the tail behaviour of the prior, ranging from geometric tails to heavy power-law tails. Under the Pitman–Yor process prior, we obtained a simple characterization of the posterior distribution of τ_1, in terms of a compound (general) hypergeometric distribution, and made use of the posterior mean as an estimator of τ_1. Such a novel estimator has all the desirable features as Samuels' estimator, including ease of implementation, computational efficiency and scalability to massive data, and, in addition, it allows to reduce its underestimation of τ_1 by tuning the "discount" parameter with respect to observable microdata. We presented an empirical analysis of our Bayesian nonparametric approach through synthetic data and real data, showing its effectiveness in reducing underestimation phenomenon of Samuels' approach.

While τ_1 is known to be the most popular measure of disclosure risk (Bethlehem et al. [2] and Skinner et al. [24]), one might consider alternative measures by broadening the definition of "uniqueness". For instance, Fienberg and Makov [11] considered a generalization of τ_1 which is defined in terms of the number of cells with frequency less or equal than 2. In general, one may consider the following measure

$$\tau_{p,q} = \sum_{i=1}^{K_n} I(N_{i,n} \leq p) I(N_{i,N} \leq p+q),$$

namely the number of cells with sample frequency less or equal than p which have population frequency less or equal than $p+q$. In particular, τ_1 corresponds to $\tau_{1,0}$. We refer to Appendix D for Bayesian nonparametric inference of $\tau_{1,q}$, which is arguably the most natural generalization of τ_1. It remains an open problem to adapt our Bayesian nonparametric approach to deal with structurally empty cells, i.e. structural zeros (Manrique-Vallier and Reiter [14]). In such a context, it may be useful to consider spike and slab generalizations of the Pitman-Yor process prior (Scarpa and Dunson [22], Canale et al. [3]). They consist in replacing the non-atomic distribution ν of the Pitman-Yor process prior with a distribution $\tilde{\nu}(\zeta) = \zeta\delta_0 + (1-\zeta)\nu$, with $\zeta \in [0,1]$ and ν being a non-atomic distribution. Then ζ may then be used to include the information on structural zeros, being interpretable as the proportion of structural zeros in the population.

Appendix A: Proof of Theorem 1

A.1. Generalized factorial coefficients

For $t \in \mathbb{R}$, $a > 0$ and $n \in \mathbb{N}_0$, let $(at)_{(n)}$ be the rising factorial of at of order n, i.e. $(at)_{(n)} = \prod_{0 \leq i \leq n-1}(at + i)$. The (n, k)-th generalized factorial coefficient, denoted by $\mathscr{C}(n, k; a)$, is the k-th coefficient in the expansion of $(at)_{(n)}$ into rising factorials, i.e.

$$(at)_{(n)} = \sum_{i=0}^{n} \mathscr{C}(n, i; a)(t)_{(i)}, \tag{10}$$

with $\mathscr{C}(0, 0; a) = 1$, $\mathscr{C}(n, 0; a) = 0$ for $n > 0$, $\mathscr{C}(n, i; a) = 0$ for $i > n$. For $b > 0$, let us consider the k-th coefficient in the expansion of $(at - b)_{(n)}$ into rising factorials, so that

$$(at - b)_{(n)} = \sum_{i=0}^{n} \mathscr{C}(n, i; a, b)(t)_{(i)}, \tag{11}$$

with $\mathscr{C}(0, 0; a, b) = 1$, $\mathscr{C}(n, 0; a, b) = (-b)_{(n)}$ for $n > 0$, $\mathscr{C}(n, i; a, b) = 0$ for $i > n$. The coefficient $\mathscr{C}(n, k; a, b)$ is referred to as the non-centered generalized factorial coefficient (Charalambides [7]). Here, it is useful to recall the following property

$$\mathscr{C}(n, i; b_1 b_2, b_1 r_2 + r_1) = \sum_{j=i}^{n} \mathscr{C}(n, j; b_1, r_1)\mathscr{C}(j, i; b_2, r_2), \tag{12}$$

for any $b_1, b_2 > 0$ and $r_1, r_2 > 0$. The convolutional identity (12) can be found in Charalambides [Chapter 2, 7] and plays a critical role in the proof of Theorem 1.

A.2. Generalized factorial and (general) hypergeometric distributions

The generalized factorial distribution (Charalambides [7, Chapter 2]) is defined by means of the identity (10), and it arises in the context of the classical coupon collector problem (Charalambides [7, Example 2.7]). For $r \in \mathbb{N}$ and $b, c > 0$, a random variable $U_{b,c,r}$ on the set $\{1, \ldots, r\}$ has a generalized factorial distribution if

$$\mathbb{P}(U_{b,c,r} = x) = \frac{1}{(bc)_{(r)}} \mathscr{C}(r, x; b)(c)_{(x)} I(x \in \{1, \ldots, r\}). \tag{13}$$

The (general) hypergeometric distribution (Johnson et al. [12, Chapter 6.2.5]) has the same form as the classical hypergeometric distribution, though with a more flexible parameterization. In particular, for $r, s \in \mathbb{N}$ and $a > 0$ such that $a > r$, a random variable $H_{a,r,s}$ on the set $\{0, 1, \ldots, r\}$ has a generalized factorial distribution if

$$\mathbb{P}(H_{a,r,s} = x) = \frac{\binom{a}{x}\binom{s}{r-x}}{\binom{a+s}{r}} I(x \in \{0, 1, \ldots, r\}). \tag{14}$$

Distributional properties, and moments, of the general hypergeometric distribution can be easily obtained from (14) (Johnson et al. [12, Chapter 6.3]). For $r, s \in \mathbb{N}$ with $s \leq r$ let $S(r, s)$ be the Stirling number of the second type (Charalambides [7, Chapter 2]), and let Γ denote the Gamma function. Then, for $z > 0$ it holds

$$\mathbb{E}\{(H_{a,r,s})^z\} = \sum_{i=1}^{z} S(z,i) i! \binom{r}{i} \frac{\Gamma(a+1+s-i)\,\Gamma(a+1)}{\Gamma(a+1-i)\,\Gamma(a+1+s)}. \tag{15}$$

We refer to Charalambides [7] and Johnson et al. [12] for a comprehensive account of the generalized factorial distribution and the (general) hypergeometric distribution.

A.3. Proof of Theorem 1

Let $(X_1, \ldots, X_n)$ be a random sample from the Pitman-Yor process $P_{\alpha,\theta}$, and let $(X_1, \ldots, X_n)$ feature $K_n = k$ distinct symbols, labelled by $\{Z_1^*, \ldots, Z_{K_n}^*\}$, with frequencies $\mathbf{N}_n = \mathbf{n}$, with $\mathbf{N}_n = (N_{1,n}, \ldots, N_{K_n,n})$, and $\mathbf{n} = (n_1, \ldots, n_k)$ be such that $N_{i,n} > 0$ and $\sum_{1 \leq i \leq K_n} N_{i,n} = n$. Moreover, for any $N > n$ let $(X_{n+1}, \ldots, X_N)$ be an additional random sample from $P_{\alpha,\theta}$, and let $N_{j,N-n} \geq 0$ be the number of records $X_{n+i}, i = 1 \ldots, N$ that coincide with the label Z_j^*, $j = 1, \ldots, K_n$. Moreover, let

$$V_{N-n} = N - n - \sum_{i=1}^{K_n} N_{i,N-n}$$

be the number of $X_{n+i}, i = 1, \ldots, N$ that do not coincide with any Z_j^*'s. To compute the posterior distribution of τ_1, we first determine its moment of order $z \geq 1$, i.e.,

$$\begin{aligned} &\mathbb{E}\{(\tau_1)^z \mid X_1, \ldots, X_n\} \qquad (16)\\ &\quad = \mathbb{E}\{(\tau_1)^z \mid \mathbf{N}_n = \mathbf{n}, K_n = k\} \\ &\quad = \mathbb{E}\left\{ \left(\sum_{i=1}^{K_n} I(N_{i,n} = 1) I(N_{i,N-n} = 0) \right)^z \mid \mathbf{N}_n = \mathbf{n}, K_n = k \right\}. \end{aligned}$$

For $s, t \in \mathbb{N}$ with $s \leq t$ recall that $S(s, t)$ denotes the Stirling number of the second type (Charalambides [7, Chapter 2]), and let $\mathcal{C}_{t,s}$ denote a set of combination defined as follows: $\mathcal{C}_{t,0} = \emptyset$ and $\mathcal{C}_{t,s} = \{(c_1, \ldots, c_s) : c_i \in \{1, \ldots, t\}, c_i \neq c_j, \text{ if } i \neq j\}$ for any $s \geq 1$. Accordingly, Equation (16) admits the following expansion

$$\begin{aligned} &\mathbb{E}\{(\tau_1)^z \mid X_1, \ldots, X_n\} \\ &= \sum_{x=1}^{k} \sum_{i_1=1}^{z} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x-1}=1}^{i_{x-2}-1} \binom{r}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x-2}}{i_{x-1}} \end{aligned}$$

$$
\begin{aligned}
&\times \sum_{(c_1,\dots,c_x)\in\mathcal{C}_{k,x}} \mathbb{E}\left\{\prod_{t=1}^{x} I(N_{c_t,n}=1)I(N_{c_t,N-n}=0)^{i_{x-t}-i_{x-t+1}} \mid \mathbf{N}_n=\mathbf{n}, K_n=k\right\}\\
&=\sum_{x=1}^{z} S(z,x)x! \sum_{(c_1,\dots,c_x)\in\mathcal{C}_{k,x}} \mathbb{E}\left\{\prod_{t=1}^{x} I(N_{c_t,n}=1)I(N_{c_t,N-n}=0) \mid \mathbf{N}_n=\mathbf{n}, K_n=k\right\}\\
&=\sum_{x=1}^{z} S(z,x)x! \sum_{(c_1,\dots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{E}\left\{\prod_{t=1}^{x} I(N_{c_t,N-n}=0) \mid \mathbf{N}_n=\mathbf{n}, K_n=k\right\}\\
&=\sum_{x=1}^{z} S(z,x)x! \qquad\qquad (17)\\
&\times \sum_{(c_1,\dots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{P}(N_{c_1,N-n}=0,\dots,N_{c_x,N-n}=0 \mid \mathbf{N}_n=\mathbf{n}, K_n=k).
\end{aligned}
$$

The conditional probability (17) can be computed by a direct application of Favaro et al. [8, Lemma 1]. In particular, from Favaro et al. [8, Equation 38 and Equation 40]

$$
\begin{aligned}
&\mathbb{P}(N_{c_1,N-n}=0,\dots,N_{c_x,N-n}=0 \mid \mathbf{N}_n=\mathbf{n}, K_n=k, V_{N-n}=v)\\
&\qquad = \frac{(n-\sum_{i=1}^{x} n_{c_i}-(k-x)\alpha)_{(N-n-v)}}{(n-k\alpha)_{(N-n-v)}};
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{P}(V_{N-n}=v \mid \mathbf{N}_n=\mathbf{n}, K_n=k)\\
&\qquad = \binom{N-n}{v}(n-k\alpha)_{(N-n-v)} \sum_{j=0}^{v} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \frac{\mathscr{C}(v,j;\alpha)}{\alpha^j}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\mathbb{P}(N_{c_1,N-n}=0,\dots,N_{c_x,N-n}=0 \mid \mathbf{N}_n=\mathbf{n}, K_n=k) \qquad\qquad (18)\\
&\quad = \sum_{v=0}^{N-n} \binom{N-n}{v}(n-k\alpha)_{(N-n-v)} \sum_{j=0}^{v} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \frac{\mathscr{C}(v,j;\alpha)}{\alpha^j}\\
&\qquad\qquad \times \frac{(n-\sum_{i=1}^{x} n_{c_i}-(k-x)\alpha)_{(N-n-v)}}{(n-k\alpha)_{(N-n-v)}}\\
&\quad = \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \sum_{v=j}^{N-n} \binom{N-n}{s}(n-k\alpha)_{(N-n-v)}\mathscr{C}(v,j;\alpha)\\
&\qquad\qquad \times \frac{(n-\sum_{i=1}^{x} n_{c_i}-(k-x)\alpha)_{(N-n-v)}}{(n-k\alpha)_{(N-n-v)}}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \sum_{v=j}^{N-n} \binom{N-n}{v} \mathscr{C}(v,j;\alpha) \\
&\qquad \times (n - \sum_{i=1}^{x} n_{c_i} - (k-x)\alpha)_{(N-n-v)} \qquad (19) \\
&= \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \sum_{v=j}^{N-n} \binom{N-n}{v} \mathscr{C}(v,j;\alpha) \\
&\qquad \times \mathscr{C}(N-n-v, 0, \alpha, -n + \sum_{i=1}^{x} n_{c_i} + (k-x)\alpha). \qquad (20)
\end{aligned}
$$

Then, by the application of the convolutional identity (12) to the sum over v, we get

$$
\begin{aligned}
&\mathbb{P}(N_{c_1,N-n} = 0, \ldots, N_{c_x,N-n} = 0 \mid \mathbf{N}_n = \mathbf{n}, K_n = k) \\
&\quad = \sum_{j=0}^{N-n} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \frac{\mathscr{C}(N-n, j; \alpha, -n + \sum_{i=1}^{x} n_{c_i} + (k-x)\alpha)}{\alpha^j} \\
&\quad = \frac{1}{(\theta+n)_{(N-n)}} \sum_{j=0}^{N-n} \left(\frac{\theta}{\alpha} + k\right)_{(j)} \mathscr{C}(N-n, j; \alpha, -n + \sum_{i=1}^{x} n_{c_i} + (k-x)\alpha).
\end{aligned}
$$

Therefore, by the application of the identity (11) to the sum over j, we obtain that

$$
\begin{aligned}
&\mathbb{P}(N_{c_1,N-n} = 0, \ldots N_{c_x,N-n} = 0 \mid \mathbf{N}_n = \mathbf{n}, K_n = k) \\
&\quad = \frac{(\theta + n - \sum_{i=1}^{x} n_{c_i} + x\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}}.
\end{aligned}
$$

By a direct combination of Equation (17) and Equation (18), the moment formula (16) is

$$
\begin{aligned}
&\mathbb{E}\{(\tau_1)^z \mid X_1, \ldots, X_n\} \\
&\quad = \sum_{x=1}^{z} S(z,x) x! \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n} = 1) \\
&\qquad \times \mathbb{P}(N_{c_1,N-n} = 0, \ldots, N_{c_x,N-n} = 0 \mid \mathbf{N}_n = \mathbf{n}, K_n = k) \\
&\quad = \sum_{x=1}^{z} S(z,x) x! \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n} = 1) \frac{(\theta + n - \sum_{i=1}^{x} n_{c_i} + x\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\frac{(\theta+n-x+x\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}} \\
&= \frac{1}{(\theta+n)_{(N-n)}}\sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\left\{(1-\alpha)\left(\frac{\theta+n}{1-\alpha}-x\right)\right\}_{(N-n)},
\end{aligned}
$$

and from (10)

$$
\begin{aligned}
&\mathbb{E}\{(\tau_1)^z \mid X_1,\ldots,X_n\} \\
&\quad = \frac{1}{(\theta+n)_{(N-n)}}\sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\sum_{i=1}^{N-n}\mathscr{C}(n,i;1-\alpha)\left(\frac{\theta+n}{1-\alpha}-x\right)_{(i)} \\
&\quad = \frac{1}{(\theta+n)_{(N-n)}}\sum_{i=1}^{N-n}\mathscr{C}(n,i;1-\alpha)\sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\frac{\Gamma\left(\frac{\theta+n}{1-\alpha}+i-x\right)}{\Gamma\left(\frac{\theta+n}{1-\alpha}-x\right)} \\
&\quad = \frac{1}{(\theta+n)_{(N-n)}}\sum_{i=1}^{N-n}\mathscr{C}(n,i;1-\alpha)\frac{\Gamma\left(\frac{\theta+n}{1-\alpha}+i\right)}{\Gamma\left(\frac{\theta+n}{1-\alpha}\right)} \\
&\qquad\qquad \times\sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\frac{\Gamma\left(\frac{\theta+n}{1-\alpha}+i-x\right)\Gamma\left(\frac{\theta+n}{1-\alpha}\right)}{\Gamma\left(\frac{\theta+n}{1-\alpha}-x\right)\Gamma\left(\frac{\theta+n}{1-\alpha}+i\right)}
\end{aligned}
$$

[by the application of (15)]

$$
= \frac{1}{(\theta+n)_{(N-n)}}\sum_{i=1}^{N-n}\mathscr{C}(n,i;1-\alpha)\left(\frac{\theta+n}{1-\alpha}\right)_{(i)}\mathbb{E}\{(H_{\frac{\theta+n}{1-\alpha}-1,m_1,i})^z\} \tag{21}
$$

[by the definition of generalized factorial distribution (13)]

$$
= \sum_{i=1}^{N-n}\mathbb{E}\{(H_{\frac{\theta+n}{1-\alpha}-1,m_1,i})^z\}\mathbb{P}(U_{1-\alpha,\frac{\theta+n}{1-\alpha},N-n}=i). \tag{22}
$$

According to the above expression for $\mathbb{E}\{(\tau_1)^z \mid X_1,\ldots,X_n\}$, the proof of Theorem 1 is completed by using the definition of (general) hypergeometric distribution (14).

Appendix B: Proofs of Equation (7) and Equation (8)

Let $(X_1,\ldots,X_n)$ be a random sample from $P_{\alpha,\theta}$, and let $(X_1,\ldots,X_n)$ feature $K_n = k$ distinct symbols with $\mathbf{N}_n = (N_{1,n},\ldots,N_{K_n,n})$ corresponding frequencies, $\mathbf{n} = (n_1,\ldots,n_k)$ such that $N_{i,n} > 0$ and $\sum_{1\le i\le K_n} N_{i,n} = n$. From the proof on Theorem 1,

$$
\mathbb{E}\{(\tau_1)^z \mid X_1,\ldots,X_n\} = \sum_{i=1}^{z} S(z,x)i!\binom{m_1}{i}\frac{(\theta+n-i+i\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}}. \tag{23}
$$

Recall that by means of Stirling formula $\Gamma(n+i)/\Gamma(n) \approx n^i$ as $n \to +\infty$ is a first order approximation of the Gamma function. By applying it to (23), as $n \to +\infty$ and $N \to +\infty$.

$$\begin{aligned}
&\mathbb{E}\{(\tau_1)^z \mid X_1, \ldots, X_n\} \\
&\quad = \sum_{i=1}^{z} S(z,i) i! \binom{m_1}{i} \frac{(\theta+n-i+i\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}} \\
&\quad = \sum_{i=1}^{z} S(z,i) i! \binom{m_1}{i} \frac{\frac{\Gamma(\theta+N-i+i\alpha)}{\Gamma(\theta+n-i+i\alpha)}}{\frac{\Gamma(\theta+N)}{\Gamma(\theta+n)}} \\
&\quad \approx \sum_{i=1}^{z} S(z,i) i! \binom{m_1}{i} \left\{ \left(\frac{n}{N}\right)^{1-\alpha} \right\}^i . \qquad (24)
\end{aligned}$$

Equation (24) is the moment of order z of a Binomial random variable with parameter $(m_1, (n/N)^{1-\alpha})$, with m_1 being the number of trials and $(n/N)^{1-\alpha}$ being the probability of success in a trial. This completes the proof of Equation (7) and Equation (8).

Appendix C: On the distribution of $U_{1-\alpha, \frac{\theta+n}{1-\alpha}, N-n}$

Let $(X_1, \ldots, X_n)$ be a random sample from $P_{\alpha,\theta}$, and let $(X_1, \ldots, X_n)$ feature $K_n = k$ distinct symbols with corresponding frequencies $\mathbf{N}_n = \mathbf{n}$, where $\mathbf{N}_n = (N_{1,n}, \ldots, N_{K_n,n})$ and $\mathbf{n} = (n_1, \ldots, n_k)$ such that $N_{i,n} > 0$ and $\sum_{1 \leq i \leq K_n} N_{i,n} = n$. The distribution of K_n is known from Pitman [17, Chapter 3]. In particular, for $x \in \{1, \ldots, n\}$

$$\mathbb{P}(K_n = x) = \frac{(\theta/\alpha)_{(x)}}{(\theta)_{(n)}} \mathscr{C}(n, x; \alpha). \qquad (25)$$

According to (25), the distribution of $U_{1-\alpha, \frac{\theta+n}{1-\alpha}, N-n}$ coincides with the distribution of the number K_{N-n} distinct symbols in $N-n$ random samples from $P_{1-\alpha, \theta+n}$.

Appendix D: Bayesian nonparametric inference for $\tau_{1,q}$

Under the Pitman-Yor process prior, we characterize the posterior distribution of $\tau_{1,q}$ through its moments; this leads to a Bayesian nonparametric estimator of $\tau_{1,q}$ in terms of the posterior mean. The proof is along lines similar to the proof of Theorem 1. Let $(X_1, \ldots, X_n)$ be a random sample from the Pitman-Yor process $P_{\alpha,\theta}$, and let $(X_1, \ldots, X_n)$ feature $K_n = k$ distinct symbols, labelled by $\{Z_1^*, \ldots, Z_{K_n}^*\}$, with frequencies $\mathbf{N}_n = \mathbf{n}$, with $\mathbf{N}_n = (N_{1,n}, \ldots, N_{K_n,n})$, and $\mathbf{n} = (n_1, \ldots, n_k)$ be such that $N_{i,n} > 0$ and $\sum_{1 \leq i \leq K_n} N_{i,n} = n$. Moreover, for any $N > n$ let $(X_{n+1}, \ldots, X_N)$ be an additional random sample from $P_{\alpha,\theta}$, and let $N_{j,N-n} \geq 0$ be the number of records $X_{n+i}, i = 1 \ldots, N$ that coincide with the label $Z_j^*, j = 1, \dots, K_n$. Moreover, let $V_{N-n} = N - n - \sum_{1 \le i \le K_n} N_{i,N-n}$ be the number of $X_{n+i}, i = 1, \dots, N$ that do not coincide with any Z_j^*'s. To compute the posterior distribution of $\tau_{1,q}$, we first determine its moment of order $z \ge 1$, i.e.,

$$\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})^z \mid X_1, \dots, X_n\} \qquad (26)\\
&\quad = \mathbb{E}\{(\tau_{1,q})^z \mid \mathbf{N}_n = \mathbf{n}, K_n = k\}\\
&\quad = \mathbb{E}\left\{ \left(\sum_{i=1}^{K_n} I(N_{i,n} = 1) I(N_{i,N-n} \le q) \right)^z \mid \mathbf{N}_n = \mathbf{n}, K_n = k \right\}.
\end{aligned}$$

For $s, t \in \mathbb{N}$ with $s \le t$ recall that $S(s,t)$ denotes the Stirling number of the second type (Charalambides [7, Chapter 2]), and let $\mathcal{C}_{t,s}$ denote a set of combination defined as follows: $\mathcal{C}_{t,0} = \emptyset$ and $\mathcal{C}_{t,s} = \{(c_1, \dots, c_s) : c_i \in \{1, \dots, t\}, c_i \neq c_j, \text{ if } i \neq j\}$ for any $s \ge 1$. Accordingly, Equation (26) admits the following expansion

$$\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})^z \mid X_1, \dots, X_n\}\\
&= \mathbb{E}\left\{ \left[\sum_{i=1}^{k} \left(\sum_{h=0}^{q} I(N_{i,n}=1) I(N_{i,N-n}=h) \right) \right]^z \mid \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\}\\
&= \sum_{x=1}^{k} \sum_{i_1=1}^{z} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x-1}=1}^{i_{x-2}-1} \binom{z}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x-2}}{i_{x-1}}\\
&\quad \times \sum_{(c_1,\dots,c_x) \in \mathcal{C}_{k,x}} \mathbb{E}\left\{ \prod_{t=1}^{x} \left(\sum_{h=0}^{q} I(N_{c_t,n}=1) I(N_{c_t,N-n}=h) \right)^{i_{x-t}-i_{x-t+1}} \mid \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\}\\
&= \sum_{x=1}^{k} \sum_{i_1=1}^{z} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x-1}=1}^{i_{x-2}-1} \binom{z}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x-2}}{i_{x-1}}\\
&\quad \times \sum_{(c_1,\dots,c_x) \in \mathcal{C}_{k,x}} \mathbb{E}\left\{ \prod_{t=1}^{x} \left(\sum_{h=0}^{q} I(N_{c_t,n}=1) I(N_{c_t,N-n}=h) \right) \mid \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\}\\
&= \sum_{x=1}^{z} S(z,x) x!\\
&\quad \times \sum_{(c_1,\dots,c_x) \in \mathcal{C}_{k,x}} \mathbb{E}\left\{ \prod_{t=1}^{x} \left(\sum_{h=0}^{q} I(N_{c_t,n}=1) I(N_{c_t,N-n}=h) \right) \mid \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\}.
\end{aligned}$$

Now, we define the (cartesian product) set $\mathcal{H}_{q,x} = \{0, \dots, q\}^x$, such that we can write

$$\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})^z \mid X_1, \dots, X_n\}\\
&= \sum_{x=1}^{z} S(z,x) x!
\end{aligned}$$

$$\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}} \mathbb{E}\left\{\prod_{t=1}^{x}(I(N_{c_t,n}=1)I(N_{c_t,N-n}=h_t))|\mathbf{N}_n=\mathbf{n}_n, K_n=k\right\}$$

$$= \sum_{x=1}^{z} S(z,x)x! \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}}$$

$$\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{E}\left\{\prod_{t=1}^{x}(I(N_{c_t,N-n}=h_t)) \mid \mathbf{N}_n=\mathbf{n}_n, K_n=k\right\}$$

$$= \sum_{x=1}^{z} S(z,x)x! \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}} \tag{27}$$

$$\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{P}(N_{c_1,N-n}=h_1,\ldots,N_{c_x,N-n}=h_x|\mathbf{N}_n=\mathbf{n}_n, K_n=k).$$

The conditional probability in (27) can be computed from Lemma 1 in Favaro et al. (2013). In particular, from Equation 38 and Equation 40 in Favaro et al. (2013) we have

i)

$$\mathrm{Pr}(N_{c_1,N-n}=h_1,\ldots,N_{c_x,N-n}=h_x \mid \mathbf{N}_n=\mathbf{n}_n, K_n=k, V_{N-n}=v)$$
$$= \frac{(N-n-v)!}{(N-n-v-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!}$$
$$\times \frac{(n-\sum_{t=1}^{x} n_{c_t}-(k-x)\alpha)_{(N-n-v-\sum_{t=1}^{x} h_t)}}{(n-k\alpha)_{(N-n-v)}}$$

ii)

$$\mathrm{Pr}(V_{N-n}=v \mid \mathbf{N}_n=\mathbf{n}_n, K_n=k)$$
$$= \binom{N-n}{v}(n-k\alpha)_{(N-n-v)} \sum_{j=0}^{v} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \frac{\mathscr{C}(v,j;\alpha)}{\alpha^j},$$

and

$$\mathrm{Pr}(N_{c_1,N-n}=h_1,\ldots,N_{c_x,N-n}=h_x \mid \mathbf{N}_n=\mathbf{n}_n, K_n=k) \tag{28}$$
$$= \sum_{v=0}^{N-n} \binom{N-n}{v}(n-k\alpha)_{(N-n-v)} \sum_{j=0}^{v} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \frac{\mathscr{C}(v,j;\alpha)}{\alpha^j}$$
$$\times \frac{(N-n-v)!}{(N-n-v-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!}$$
$$\times \frac{(n-\sum_{t=1}^{x} n_{c_t}-(k-x)\alpha)_{(N-n-v-\sum_{t=1}^{x} h_t)}}{(n-k\alpha)_{(N-n-v)}}$$

$$
\begin{aligned}
&= \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \sum_{v=j}^{N-n} \binom{N-n}{v} (n-k\alpha)_{(N-n-v)} \mathscr{C}(v,j;\alpha) \\
&\quad \times \frac{(N-n-v)!}{(N-n-v-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \\
&\qquad \times \frac{(n-\sum_{t=1}^{x} n_{c_t}-(k-x)\alpha)_{(N-n-v-\sum_{t=1}^{x} h_t)}}{(n-k\alpha)_{(N-n-v)}} \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \\
&\quad \times \sum_{v=j}^{N-n} \mathscr{C}(v,j;\alpha) \frac{(N-n-\sum_{t=1}^{x} h_t)!}{(N-n-v-\sum_{t=1}^{x} h_t)! v!} \\
&\qquad \times (n-\sum_{t=1}^{x} n_{c_t}-(k-x)\alpha)_{(N-n-v-\sum_{t=1}^{x} h_t)} \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \\
&\quad \times \sum_{v=j}^{N-n} \binom{N-n-\sum_{t=1}^{x} h_t}{v} \mathscr{C}(v,j;\alpha) \\
&\qquad \times \mathscr{C}(N-n-v-\sum_{t=1}^{x} h_t, 0; \alpha, -n+\sum_{t=1}^{x} n_{c_t}+(k-x)\alpha) \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \\
&\quad \times \mathscr{C}(N-n-\sum_{t=1}^{x} h_t, j; \alpha, -n+\sum_{t=1}^{x} n_{c_t}+(k-x)\alpha) \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \frac{1}{(\theta+n)_{(N-n)}} \\
&\quad \times \sum_{j=0}^{N-n} \left(\frac{\theta}{\alpha}+k\right)_{(j)} \mathscr{C}(N-n-\sum_{t=1}^{x} h_t, j; \alpha, -n+\sum_{t=1}^{x} n_{c_t}+(k-x)\alpha) \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \frac{1}{(\theta+n)_{(N-n)}} \\
&\quad \times (\theta+n-\sum_{i=1}^{x} n_{c_t}+x\alpha)_{(N-n-\sum_{t=1}^{x} h_t)}.
\end{aligned}
$$

Then, by combining Equation (27) with Equation (28) we can write the following identities

$$
\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})^z \,|\, X_1, \ldots, X_n\} \\
&= \sum_{x=1}^{z} S(z,x)x! \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}} \\
&\quad \times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{P}(N_{c_1,N-n}=h_1,\ldots,N_{c_x,N-n}=h_x|\mathbf{N}_n=\mathbf{n}_n, K_n=k) \\
&= \sum_{x=1}^{z} S(z,x)x! \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}} \\
&\quad \times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1) \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \\
&\quad\quad \times \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \frac{(\theta+n-\sum_{i=1}^{x} n_{c_t} + x\alpha)_{(N-n-\sum_{t=1}^{x} h_t)}}{(\theta+n)_{(N-n)}} \\
&= \sum_{x=1}^{r} S(r,x)x! \binom{m_1}{x} \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}} \\
&\quad \times \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(1-\alpha)_{(h_t)}}{h_t!} \frac{(\theta+n-x+x\alpha)_{(N-n-\sum_{t=1}^{x} h_t)}}{(\theta+n)_{(N-n)}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})_{[z]} \,|\, X_1, \ldots, X_n\} \qquad (29) \\
&\quad = \mathbb{E}\left\{ \left(\sum_{i=1}^{k} \left(\sum_{h=0}^{q} I(N_{i,n}=1) I(N_{i,N-n}=h) \right) \right)_{[z]} \,|\, \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\} \\
&\quad = z! \binom{m_1}{z} \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \\
&\quad\quad \times \frac{(N-n)!}{(N-n-\sum_{t=1}^{z} i_t)!} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \frac{(\theta+n-z+z\alpha)_{(N-n-\sum_{t=1}^{z} i_t)}}{(\theta+n)_{(N-n)}}.
\end{aligned}
$$

Equation (29) leads, by means of standard arguments on inversion formula, to the calculation of the conditional distribution of $\tau_{1,q}$ given $(X_1, \ldots, X_n)$. In particular, a Bayesian nonparametric estimator of $\tau_{1,q}$, is given by the posterior mean

$$
\begin{aligned}
\hat{\tau}_{1,q} &= \mathbb{E}\{\tau_{1,q} \,|\, X_1, \ldots, X_n\} \\
&= m_1 \sum_{i=0}^{q} \frac{(N-n)!}{(N-n-i)!} \frac{(1-\alpha)_{(i)}}{i!} \frac{(\theta+n-1+\alpha)_{(N-n-i)}}{(\theta+n)_{(N-n)}}.
\end{aligned}
$$

Note that $\hat{\tau}_{1,0}$ coincides with $\hat{\tau}_1$ in Theorem 1. We conclude the study of $\tau_{1,q}$ by considering the large n and large N asymptotic behaviour of the posterior (falling) factorial moment $\mathbb{E}\{(\tau_{1,q})_{[r]} \mid X_1, \ldots, X_n\}$. In particular, we write the following

$$
\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})_{[z]} \mid X_1, \ldots, X_n\} \\
&\quad = z!\binom{m_1}{z} \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \\
&\qquad \times \frac{(N-n)!}{(N-n-\sum_{t=1}^{z} i_t)!} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \frac{(\theta+n-z+z\alpha)_{(N-n-\sum_{t=1}^{z} i_t)}}{(\theta+n)_{(N-n)}} \\
&\quad = z!\binom{m_1}{z} \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \\
&\qquad \times \frac{\Gamma(N-n+1)/\Gamma(N)}{\Gamma(N-n-\sum_{t=1}^{z} i_t+1)/\Gamma(N)} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \frac{\frac{\Gamma(\theta-z+z\alpha+N-\sum_{t=1}^{z} i_t)/\Gamma(N)}{\Gamma(\Gamma(\theta+n-z+z\alpha)/\Gamma(n)}}{\frac{\Gamma(\theta+N)/\Gamma(N)}{\Gamma(\theta+n)/\Gamma(n)}} \\
&\quad \approx z!\binom{m_1}{z} \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \frac{N^{-n+1}}{N^{-n-\sum_{t=1}^{z} i_t+1}} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \frac{\frac{N^{\theta-r+r\alpha-\sum_{i=1}^{z} i_t}}{n^{\theta-z+z\alpha}}}{\frac{N^\theta}{n^\theta}} \\
&\quad = z!\binom{m_1}{z} \left[\left(\frac{n}{N}\right)^{1-\alpha}\right]^z \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \\
&\quad = z!\binom{m_1}{z} \left[\left(\frac{n}{N}\right)^{1-\alpha}\right]^z \left[\frac{\Gamma(2+q-\alpha)}{\Gamma(1+q)\Gamma(2-\alpha)}\right]^z \\
&\quad = z!\binom{m_1}{z} \left[\left(\frac{n}{N}\right)^{1-\alpha} \frac{\Gamma(2+q-\alpha)}{\Gamma(1+q)\Gamma(2-\alpha)}\right]^z. \qquad (30)
\end{aligned}
$$

If

$$
\left(\frac{n}{N}\right)^{1-\alpha} \frac{\Gamma(2+q-\alpha)}{\Gamma(1+q)\Gamma(2-\alpha)} \in (0,1) \qquad (31)
$$

then Equation (30) is the falling factorial moment of order z of a Binomial random variable with parameter $(m_1, (n/N)^{1-\alpha}\Gamma(2+q-\alpha)/\Gamma(1+q)\Gamma(2-\alpha))$, with m_1 being the number of trials and $(n/N)^{1-\alpha}\Gamma(2+q-\alpha)/\Gamma(1+q)\Gamma(2-\alpha)$ being the probability of success in a trial. Note that (31) is always satisfied for $q = 0$.

Acknowledgments

The authors are grateful to an Associate Editor and an anonymous Referee for all their critical comments, corrections, and suggestions which improved remarkably the present paper.

References

[1] Bacallado, S., Battiston, M., Favaro, S. and Trippa, L. (2015). Sufficientness postulates for Gibbs-type priors and hierarchial generalizations. *Statistical Science* **32**, 487–500. MR3730518

[2] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association* **85** 38–45.

[3] Canale, A., Lijoi, A., Nipoti, B. and Prünster, I. (2017). On the Pitman-Yor process with spike and slab base measure. *Biometrika*, **104** 681–697. MR3694590

[4] Carota, C., Filippone, M., Leombruni, R. and Polettini, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *The Annals of Applied Statistics* **9**, 525–546. MR3341126

[5] Carota, C., Filippone, M. and Polettini, S. (2018). Assessing Bayesian semi-parametric log-linear models: an application to disclosure risk estimation. *International Statistical Review*, to appear.

[6] Camerlenghi, F., Favaro, S., Naulet, Z. and Panero, F. (2020). Optimal disclosure risk assessment. *The Annals of Statistics*, **49**, 723–744. MR4255105

[7] Charalambides, C.A. (2005) *Combinatorial methods in discrete distributions*, Wiley Series in Probability and Statistics. MR2131068

[8] Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability*, **23**, 1721–1754. MR3114915

[9] Favaro, S., Naulet, Z. (2021). Near-optimal estimation of the unseen under regularly varying tail populations. *arXiv:2104.03251.*

[10] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230. MR0350949

[11] Fienberg, S.E. and Makov, U.E. (1998). Condentiality, uniqueness, and disclosure limitation for categorical data *J. Off. Stat.* **14**, 385–397.

[12] Johnson, N.L., Kemp, A.W. and Kotz, S. (2005) *Univariate discrete distributions*, Wiley Series in Probability and Statistics. MR2163227

[13] Manrique-Vallier, D. and Reiter, J.P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association* **107**, 1385–1394. MR3036402

[14] Manrique-Vallier, D. and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statististics*, **23** 1061–1079. MR3270711

[15] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92**, 21–39. MR1156448

[16] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158. MR1337249

[17] Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté

de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics, Springer New York. MR2245368

[18] Pitman, J. and Yor, M. (1999). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900. MR1434129

[19] Reiter, J.P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100**, 1103–1112. MR2236926

[20] Rinott, Y. and Shlomo, N. (2006). A generalized negative binomial smoothing model for sample disclosure risk estimation. In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, Springer, Berlin. MR2459186

[21] Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics* **14**, 373–383.

[22] Scarpa, B. and Dunson, D. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics*, **65** 772–780. MR2649850

[23] Skinner, and Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B* **64**, 855–867. MR1979391

[24] Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics* **10**, 31–51.

[25] Skinner, and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.* **103**, 989–1001. MR2462887

[26] Willenborg, L. and de Waal, T. (2001). *Elements of statistical disclosure control.* Springer, New York. MR1866909