Consistent estimation of small masses in feature sampling

Journal of Machine Learning Research 22 (2021) 1-28 Submitted 8/18; Revised 1/21; Published 1/21

Consistent estimation of small masses in feature sampling

Fadhel Ayed FADHEL.AYED@SOME.OX.AC.UK
Department of Statistics
University of Oxford,
24-29 St Giles', OX1 3LB Oxford, United Kingdom.

Marco Battiston M.BATTISTON@LANCASTER.AC.UK
Department of Mathematics and Statistics
Lancaster University,
Fylde Ave, Bailrigg, LA1 4YR Lancaster, United Kingdom.

Federico Camerlenghi FEDERICO.CAMERLENGHI@UNIMIB.IT
Department of Economics, Management and Statistics,
University of Milano - Bicocca,
Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy.

Stefano Favaro STEFANO.FAVARO@UNITO.IT
Department of Economics and Statistics
University of Torino and Collegio Carlo Alberto
Corso Unione Sovietica 218/bis, 10134, Torino, Italy.

Editor: David Blei

Abstract

Consider an (observable) random sample of size n from an infinite population of individuals, each individual being endowed with a finite set of "features" from a collection of features $(F_j)_{j\geq 1}$ with unknown probabilities $(p_j)_{j\geq 1}$, i.e., p_j is the probability that an individual displays feature F_j. Under this feature sampling framework, in recent years there has been a growing interest in estimating the sum of the probability masses p_j's of features observed with frequency $r \geq 0$ in the sample, here denoted by $M_{n,r}$. This is the natural feature sampling counterpart of the classical problem of estimating small probabilities in the species sampling framework, where each individual is endowed with only one feature (or "species"). In this paper we study the problem of consistent estimation of the small mass $M_{n,r}$. We first show that there do not exist universally consistent estimators, in the multiplicative sense, of the missing mass $M_{n,0}$. Then, we introduce an estimator of $M_{n,r}$ and identify sufficient conditions under which the estimator is consistent. In particular, we propose a nonparametric estimator $\hat{M}_{n,r}$ of $M_{n,r}$ which has the same analytic form of the celebrated Good–Turing estimator for small probabilities, with the sole difference that the two estimators have different ranges (supports). Then, we show that $\hat{M}_{n,r}$ is strongly consistent, in the multiplicative sense, under the assumption that $(p_j)_{j\geq 1}$ has regularly varying heavy tails.

Keywords: Feature sampling; Good–Turing estimator; missing mass; multiplicative consistency; nonparametric inference; regularly varying heavy-tailed distributions; species sampling.

1. Introduction

The estimation of small probabilities is a classical problem in statistics, dating back to the work of Alan M. Turing and Irving J. Good at Bletchley Park in the 1940s (Good, 1953). To define small probabilities, let consider the following species sampling framework: i) an infinite population of individuals, with each individual belonging to one "species" from a collection of (possibly infinite) species labelled by $(S_j)_{j\geq 1}$; ii) an unknown probability distribution $(q_j)_{j\geq 1}$, with q_j being the probability that an individual belongs to species S_j; iii) an (observable) random sample $(Z_1, \ldots, Z_n)$ of individuals from the population. If $\mathcal{S}_{n,r}$ is the set of labels of species with frequency $r \geq 0$ in the observable sample, with the convention that $\mathcal{S}_{n,0}$ are labels not in the sample, then the small probability of order $r \geq 0$ is defined as

$$P_{n,r} = \mathbb{P}(Z_{n+1} \in \mathcal{S}_{n,r} \,|\, Z_1, \ldots, Z_n) = \sum_{j\geq 1} q_j \mathbb{1}_{\{S_j \in \mathcal{S}_{n,r}\}}, \tag{1}$$

where $\mathbb{1}_{\{\}}$ denotes the indicator function. That is, $P_{n,r}$ is the total probability mass of species observed with frequency $r \geq 0$ in the sample $(Z_1, \ldots, Z_n)$. In particular, $P_{n,0}$ is referred to as the missing mass, i.e., the total probability mass of unseen species in the sample. The Good–Turing estimator (Good, 1953) is the most popular estimator of $P_{n,r}$. This is a nonparametric estimator, in the sense that it does not rely on any distributional assumption of the q_j's, and it has been the subject of numerous studies. These studies include, e.g., central limit theorems and large deviation principles (Zhang and Zhang, 2009; Gao, 2013; Grabchak and Zhang, 2017), admissibility and concentration properties (McAllester and Schapire, 2000; Ohannessian and Dahleh, 2012; Ben-Hamou et al., 2017), consistency and corresponding convergence rates (McAllester and Ortiz, 2003; Mossel and Ohannessian, 2019; Ayed et al., 2018), optimality and minimax properties (Orlitsky et al., 2003; Ayed et al., 2018).

The importance of estimating small probabilities has been growing dramatically in recent years, driven by applications in the broad areas of biological and physical sciences (Kroes et al., 1999; Gao et al., 2007; Daley and Smith, 2013), in linguistics (Gale and Sampson, 1995; Ohannessian and Dahleh, 2012), in machine learning (Bubeck et al., 2013; Cai et al., 2018), in information theory (Orlitsky et al., 2004; Ohannessian and Dahleh, 2012; Ben-Hamou et al., 2018) and in forensic DNA analysis (Anevski et al., 2017; Cereda and Gill, 2020). At the same time, there has been a growing interest, especially in ecology and biological sciences, in the estimation of the missing mass within the more general framework of feature sampling. See, e.g., Ionita-Laza et al. (2009), Ionita-Laza et al. (2010), Chao et al. (2014), Gravel (2014), Zou et al. (2016), and Ayed et al. (2019). The feature sampling framework generalizes the species sampling framework by allowing each individual in the population to belong to more than one "species", now called "feature". Formally, the feature sampling framework consists of: i) an infinite population of individuals, with each individual endowed with a finite set of features selected from a collection of (possibly infinite) features labelled by $(F_j)_{j\geq 1}$; ii) a collection of unknown probabilities $(p_j)_{j\geq 1}$, with p_j being the probability that an individual is endowed with feature F_j; iii) an (observable) random sample $(Y_1, \ldots, Y_n)$ of individuals from the population. In analogy to the species sampling framework, under the feature sampling framework the missing mass is defined as the sum of the probabilities p_j's of unseen features in the observed sample $(Y_1, \ldots, Y_n)$.

The Bernoulli product model is arguably the most popular model for estimating the missing mass under the feature sampling framework. It assumes that the ith element of the random sample $(Y_1, \ldots, Y_n)$ is a sequence $Y_i = (Y_{i,j})_{j\geq 1}$ of independent Bernoulli random variables with unknown feature probabilities $(p_j)_{j\geq 1}$, and that Y_r is independent of Y_s for any $r \neq s$. Therefore the random variable $X_{n,j} := \sum_{1\leq i\leq n} Y_{i,j}$, i.e., the number of times that feature F_j appears in the random sample $(Y_1, \ldots, Y_n)$, is distributed according to a Binomial distribution with parameter (n, p_j), for any $j \geq 1$. In analogy to the species sampling framework, given the random sample $(Y_1, \ldots, Y_n)$ under the Bernoulli product model we define the small mass of order $r \geq 0$ as

$$M_{n,r}(Y_1, \ldots, Y_n; (p_j)_{j\geq 1}) = \mathbb{E}\left[\sum_{j\geq 1} \mathbb{1}_{\{X_{n,j}=r, Y_{n+1,j}=1\}} \,|\, Y_1, \ldots, Y_n\right] = \sum_{j\geq 1} p_j \mathbb{1}_{\{X_{n,j}=r\}}. \quad (2)$$

That is, $M_{n,r}(Y_1, \ldots, Y_n; (p_j)_{j\geq 1})$ is the total mass of features observed with frequency $r \geq 0$ in the sample $(Y_1, \ldots, Y_n)$. In particular, we refer to $M_{n,0}(Y_1, \ldots, Y_n; (p_j)_{j\geq 1})$ as the missing mass, i.e., the total mass of unseen features in the observable sample. For ease of notation, in the rest of the paper we will not highlight the dependence on $(Y_1, \ldots, Y_n)$ and $(p_j)_{j\geq 1}$ in $M_{n,r}(Y_1, \ldots, Y_n; (p_j)_{j\geq 1})$, and we simply write $M_{n,r}$. Motivated by recent works on consistent estimation of $P_{n,r}$ (Ohannessian and Dahleh, 2012; Ben-Hamou et al., 2017; Grabchak and Zhang, 2017; Mossel and Ohannessian, 2019; Ayed et al., 2018) in this paper we investigate the problem of consistent estimation of $M_{n,r}$ under the Bernoulli product model.

As an extension of the main result of Mossel and Ohannessian (2019) to the feature sampling framework, we first show that there do not exist universally consistent estimators, in the multiplicative sense, of the missing mass $M_{n,0}$. That is, under the Bernoulli product model we prove that for any estimator $\hat{T}_{n,0}$ of $M_{n,0}$ there exists at least a choice of feature probabilities $(p_j)_{j\geq 1}$ for which $\hat{T}_{n,0}/M_{n,0}$ does not converge to 1 in probability, as $n \to +\infty$. Our strategy of proof differs from the constructive strategy of Mossel and Ohannessian (2019), and it relies on non-trivial generalizations of Bayesian nonparametric ideas and techniques developed in Ayed et al. (2018). In particular, the use of generalized Beta process prior (James, 2017) is critical to our strategy, which allows us to prove non-consistency of $\hat{T}_{n,0}$ by exploiting the posterior distribution of $M_{n,0}$ given $(Y_1, \ldots, Y_n)$. Based on our inconsistency result, we then consider the problem of introducing a consistent nonparametric estimator for $M_{n,r}$. This problem leads us to extend most of the results of Ohannessian and Dahleh (2012) and Ben-Hamou et al. (2017) to the feature sampling framework. We propose an estimator $\hat{M}_{n,r}$ of $M_{n,r}$ which has the same analytic form of the Good–Turing estimator of $P_{n,r}$, with the difference that the two estimators have different ranges (supports): while the Good–Turing estimator of $P_{n,r}$ takes values in the set $(0, 1)$, our estimator for $M_{n,r}$ takes values in $\mathbb{R}^+$, indeed the p_j's in (2) are not required to sum up to 1. Then, we show that $\hat{M}_{n,r}$ is strongly consistent, in the multiplicative sense, under the assumption that $(p_j)_{j\geq 1}$ has regularly varying heavy tails (Karlin, 1967). Proofs of our results relies on novel concentration inequalities for $M_{n,r}$, which may be of independent interest.

There is a growing interest in the estimation of $M_{n,0}$, mainly driven by applications in biological sciences. This is typically motivated by sampling procedures that are expensive, in terms of time and/or financial resources, and further draws are legitimated only by the

possibility of recording unseen features. In genetics, for instance, the ambitious prospect of growing databases to encompass hundreds of thousands of human genomes, makes important to quantify the power of large sequencing projects to discover new genetic variants (Auton et al., 2015). An accurate estimate of $M_{n,0}$ introduces a criterion for evaluating the effectiveness of further sampling, providing a roadmap for large-scale sequencing projects: one can fix a suitable threshold such that the sampling procedure takes place until the estimate of the total mass $M_{n,0}$ of unseen genetic variants becomes for the first time smaller than the threshold (Ionita-Laza et al., 2009; Gravel, 2014; Zou et al., 2016). The estimation of $M_{n,r}$ is also relevant in genetics, where there is a concrete interest in estimating the total mass of relatively rare genetic variants, i.e., small values of r, since these variants are known to play a critical role in disease predisposition (Cirulli and Goldstein, 2010; Bomba et al., 2017). To the best of our knowledge, the problem of estimating $M_{n,r}$ first appeared in Cai et al. (2018) within the context of learning augmented algorithms, with potential applications in computational biology (Zhang et al., 2014), password security (Schechter, 2010), games (Harrison, 2010) and social networks (Song et al., 2009). In particular, Cai et al. (2018) relies on Bayesian nonparametric estimates of $M_{n,r}$ to provide new insights on the count-min sketch, a time and memory efficient randomized data structure for estimating the number of times a symbol has been observed in a data stream (Cormode and Muthukrishnan, 2005).

The paper is structured as follows. Section 2 contains a brief review on consistent estimation of small probabilities. In Section 3 we prove that, under the Bernoulli product model, there do not exist universally consistent estimators, in the multiplicative sense, of the missing mass $M_{n,0}$. Section 4 contains novel exponential tail bounds for the small mass $M_{n,r}$, for $r \geq 0$, as well as exponential tail bounds for related statistics. In Section 5 we introduce a nonparametric estimator $\hat{M}_{n,r}$ of $M_{n,r}$ and we apply tail bounds of Section 4 to show that $\hat{M}_{n,r}$ is a consistent estimator under the assumption of regularly varying feature probabilities p_j's. Section 6 contains a discussion of our results and remaining open challenges. Some technical lemmas used in the proofs of Theorem 1 and Theorem 12 are deferred to Appendix A.

2. A review on consistent estimation of $P_{n,r}$

The Multinomial model is arguably the most popular model for estimating the small probabilities $P_{n,r}$. It assumes that the (observable) random sample $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$ from the population is a collection of independent and identically distributed random variables from an unknown discrete distribution $q = \sum_{j\geq 1} q_j \delta_{S_j}$. Both species' labels $(S_j)_{j\geq 1}$ and probability masses $(q_j)_{j\geq 1}$ are unknown and, without loss of generality, we assume that $(S_j)_{j\geq 1}$ is a sequence of distinct points in $[0,1]$. We can therefore consider as parameter space the set

$$\mathscr{Q} = \left\{ \sum_{j\geq 1} q_j \delta_{S_j} \,:\, S_j \in [0,1],\, q_j \in [0,1] \text{ and } \sum_{j\geq 1} q_j = 1 \right\}.$$

For a fixed n, an estimator $\hat{R}_{n,r} : [0,1]^n \to [0,1]$ of the missing mass $P_{n,r}$ is a measurable map whose argument is the observed sample $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$. We say that the estimator $\hat{R}_{n,r}$ is multiplicative consistent for $P_{n,r}$, under the parameter space $\mathscr{Q}$, if for every $\varepsilon > 0$

and for every $q \in \mathscr{Q}$

$$\lim_{n\to+\infty} \mathbb{Q}_{\mathbf{Z}_n|q}\left(\left|\frac{\hat{R}_{n,r}}{P_{n,r}} - 1\right| \geq \varepsilon\right) = 0, \tag{3}$$

where $\mathbb{Q}_{\mathbf{Z}_n|q}$ denotes the law of the observations $\mathbf{Z}_n$ under a Multinomial model of parameter q. The problem of consistent estimation, in the multiplicative sense, of $P_{n,r}$ has been the subject of recent works by Ohannessian and Dahleh (2012), Mossel and Ohannessian (2019), Ben-Hamou et al. (2017) and Ayed et al. (2018). Here we review the main contributions of these works.

The choice of the multiplicative loss function $L(\hat{R}_{n,r}, P_{n,r}) = |\hat{R}_{n,r}/P_{n,r} - 1|$ is known to be suitable for estimating small value parameters, such as $P_{n,r}$, since it allows to achieve more informative results. Besides the estimation of $P_{n,r}$, the multiplicative loss function has been used, for instance, in the estimation of small value probabilities using importance sampling (Chatterjee and Diaconis, 2018) and in the estimation of tail probabilities in extreme value theory (Beirlant and Devroye, 1999). Under the multiplicative loss function, Ohannessian and Dahleh (2012) studied the consistency of the Good–Turing estimator $\hat{P}_{n,r}$ of $P_{n,r}$, i.e.,

$$\hat{P}_{n,r} = (r+1)\frac{K_{n,r}}{n}$$

where $K_{n,r}$ is the number of distinct species appearing exactly r times in $\mathbf{Z}_n$. They showed that if $q \in \mathscr{Q}$ is a geometric distribution with small enough parameter, then there exists $\varepsilon > 0$ such that

$$\lim_{n\to+\infty} \mathbb{Q}_{\mathbf{Z}_n|q}\left(\left|\frac{\hat{P}_{n,0}}{P_{n,0}} - 1\right| \geq \varepsilon\right) > C, \tag{4}$$

for some $C > 0$. That is, $\hat{P}_{n,0}$ is not a universally consistent estimator, in the multiplicative sense, of the missing mass $P_{n,0}$. As discussed in Ohannessian and Dahleh (2012), the intuition behind (4) is that with a light-tailed distribution like the geometric distribution, there are not enough samples to learn the small probabilities well enough for consistency. Similar arguments shows that $\hat{P}_{n,r}$ is not a universally consistent estimator, in the multiplicative sense, of $P_{n,r}$.

Based on the non-consistency result (4), Ohannessian and Dahleh (2012) investigated conditions on $\mathscr{Q}$ to obtain multiplicative consistency for $P_{n,r}$. In particular, they showed that the assumption of regularly varying heavy-tailed distributions is sufficient to obtain strong multiplicative consistency for $P_{n,r}$. To define regularly varying heavy-tailed distributions, let $\nu(\mathrm{d}x) = \sum_{j\geq 1} \delta_{q_j}(\mathrm{d}x)$ and let $\bar{\nu}(x) = \nu[x,1]$. Then, following Karlin (1967), we say that $q \in \mathscr{Q}$ is a regularly varying heavy-tailed distribution, with regular variation index $\alpha \in (0,1)$ if

$$\bar{\nu}(x) \sim x^{-\alpha}\ell(1/x) \qquad x \downarrow 0$$

where $\ell(\cdot)$ is a slowly varying function, namely $\ell(ct)/\ell(t) \to 1$ as $t \to +\infty$ for all $c > 0$. Ohannessian and Dahleh (2012) showed that if $q \in \mathscr{Q}$ is a regularly varying heavy-tailed distribution with index α, then for every $r \geq 0$ there exist a universal constant a_r and distribution specific constants $b_r > 0$, $n_r < +\infty$ and $\varepsilon_r > 0$ such that for all $n > n_r$ and for all $\varepsilon \in (0, \varepsilon_r)$

$$\mathbb{Q}_{\mathbf{Z}_n|q}\left(\left|\frac{P_{n,r}}{\mathbb{E}[P_{n,r}]} - 1\right| > \varepsilon\right) \leq a_r \mathrm{e}^{-b_r\varepsilon^2 n^\alpha \ell(n)}. \tag{5}$$

Then, by combining (5) with the Borel–Cantelli lemma it follows that $\hat{P}_{n,r}/P_{n,r} \to 1$ almost surely, as $n \to +\infty$. That is, under the assumption of regularly varying heavy-tailed distributions, the Good–Turing estimator $\hat{P}_{n,r}$ is strongly consistent, in the multiplicative sense, for $P_{n,r}$.

Mossel and Ohannessian (2019) strengthened the result of non-consistency obtained in Ohannessian and Dahleh (2012). Specifically, let $\hat{R}_{n,0}$ be any estimator of the missing mass $P_{n,0}$. Then, Mossel and Ohannessian (2019) showed that there exist $\varepsilon > 0$ and $q \in \mathscr{Q}$ such that

$$\lim_{n\to+\infty} \mathbb{Q}_{\mathbf{Z}_n|q}\left(\left|\frac{\hat{R}_{n,0}}{P_{n,0}} - 1\right| \geq \varepsilon\right) > C, \tag{6}$$

for some $C > 0$. That is, there do not exist universally consistent estimators, in the multiplicative sense, of the missing mass $P_{n,0}$. Mossel and Ohannessian (2019) proved (6) by carefully defining a discrete distribution q that satisfies (6), and such a construction relies on a coupling of two generalized (dithered) geometric distributions. An alternative, and remarkably shorter, proof of (6) is given in Ayed et al. (2018). While the approach of Mossel and Ohannessian (2019) has the merit to be constructive, the approach of Ayed et al. (2018) has the merit to be direct by exploiting properties of the posterior distribution of $P_{n,0}$ under a Dirichlet process prior for q (Ferguson, 1973). Moreover, the alternative approach of Ayed et al. (2018) paved the way to study the rate of consistency of the Good–Turing estimator under the assumption of regularly varying q. In particular, Ayed et al. (2018) showed that the convergence rate $n^{-\alpha/2}$ is the best rate that any estimator $\hat{R}_{n,0}$ of $P_{n,0}$ can achieve, up to a slowly varying function, and that that the Good–Turing estimator $\hat{P}_{n,0}$ achieves that rate.

3. Non-existence of universally consistent estimators of $M_{n,0}$

Consider the Bernoulli product model described in the Introduction. Without loss of generality, we assume that each feature F_j is labeled by a value in $[0,1]$ and therefore $(F_j)_{j\geq 1}$ is a sequence of distinct points in $[0,1]$. Furthermore, the probabilities $(p_j)_{j\geq 1}$ are assumed to be summable, i.e., $\sum_{j\geq 1} p_j < +\infty$; this condition is needed in order to guarantee that every observation Y_i will display only a finite number of features almost surely. Indeed, $\sum_{j\geq 1} p_j < +\infty$ is equivalent to $\sum_{j\geq 1} \mathbb{P}(F_j \in Y_i) = \sum_{j\geq 1} \mathbb{E}[\mathbb{1}_{\{F_j\in Y_i\}}] < +\infty$, which in turns implies $\sum_{j\geq 1} \mathbb{1}_{\{F_j\in Y_i\}} < +\infty$ almost surely, by Tonelli–Fubini Theorem. The two unknown sequences $(F_j)_{j\geq 1}$ and $(p_j)_{j\geq 1}$ can be uniquely encoded in a finite measure on $[0,1]$, $\sum_{j\geq 1} p_j \delta_{F_j}(\cdot)$, with all masses at most one. We can therefore consider as parameter space the set

$$\mathscr{P} := \left\{ \sum_{j\geq 1} p_j \delta_{F_j} \,:\, F_j \in [0,1],\, p_j \in [0,1] \text{ and } \sum_{j\geq 1} p_j < +\infty \right\}. \tag{7}$$

Recall that $X_{n,j}$ denotes the number of times that feature F_j has been observed in the sample $(Y_1, \ldots, Y_n)$, that is $X_{n,j} = \sum_{1\leq i\leq n} Y_{i,j} = \sum_{1\leq i\leq n} \mathbb{1}_{\{F_j\in Y_i\}}$ is a Binomial random variable with parameter (n, p_j). For a fixed $n \geq 1$, an estimator $\hat{T}_{n,0} : (\{0,1\}^\infty)^n \to \mathbb{R}_+$ of the missing mass $M_{n,0}$ is a measurable map whose argument is the observed sample

$\mathbf{Y}_n = (Y_1, \ldots, Y_n)$. We say that the estimator $\hat{T}_{n,0}$ is multiplicative consistent for $M_{n,0}$, under the parameter space $\mathscr{P}$ if for every $\varepsilon > 0$ and every $p \in \mathscr{P}$,

$$\lim_{n \to +\infty} \mathbb{P}_{\mathbf{Y}_n|p} \left(\left| \frac{\hat{T}_{n,0}}{M_{n,0}} - 1 \right| \geq \varepsilon \right) = 0, \tag{8}$$

where $\mathbb{P}_{\mathbf{Y}_n|p}$ denotes the law of the observations $\mathbf{Y}_n$ under a Bernoulli product model of parameter p. Theorem 1 shows that there are no universally multiplicative consistent estimators of $M_{n,0}$ for the class $\mathscr{P}$. This means that for any estimator $\hat{T}_{n,0}$ of the missing mass, there exists at least one element $p \in \mathscr{P}$ for which $\hat{T}_{n,0}/M_{n,0}$ does not converge to 1 in probability, as $n \to +\infty$.

Theorem 1 *Under the feature allocation model, there are no universally consistent estimators of the missing mass, i.e., there are no estimators satisfying* (8)*. In particular, for every estimator $\hat{T}_{n,0}$, and for any $\varepsilon \in (0, 1/6)$ it is possible to find an element $p \in \mathscr{P}$*

$$\limsup_{n \to +\infty} \mathbb{P}_{\mathbf{Y}_n|p} \left(\left| \frac{\hat{T}_{n,0}}{M_{n,0}} - 1 \right| \geq \varepsilon \right) > C_\varepsilon. \tag{9}$$

for some strictly positive constant C_ε.

3.1 Proof of Theorem 1

The proof of this theorem will build on technical lemmas which statements and proofs are deferred to the appendix for clarity of the exposition. In order to prove Theorem 1, it is enough to show that for every estimator $\hat{T}_{n,0}$ and every $\varepsilon \in (0, 1/6)$,

$$\sup_{p \in \mathscr{P}} \limsup_{n \to +\infty} \mathbb{P}_{\mathbf{Y}_n|p} \left(\left| \frac{\hat{T}_{n,0}}{M_{n,0}} - 1 \right| \geq \varepsilon \right) > C_\varepsilon, \tag{10}$$

and therefore there exists a $p \in \mathscr{P}$ for which $\hat{T}_{n,0}$ is not consistent. Let us also notice that from Lemma 13 (stated in the Appendix), we know that for every $\varepsilon \in (0, 1/6)$,

$$\sup_{p \in \mathscr{P}} \limsup_{n \to +\infty} \mathbb{P}_{\mathbf{Y}_n|p} \left(\left| \frac{\hat{T}_{n,0}}{M_{n,0}} - 1 \right| \geq \varepsilon \right) \geq \sup_{p \in \mathscr{P}} \limsup_{n \to +\infty} \mathbb{P}_{\mathbf{Y}_n|p} \left(\left| \frac{M_{n,0}}{\hat{T}_{n,0}} - 1 \right| \geq 2\varepsilon \right). \tag{11}$$

Hence, denoting $\bar{\varepsilon} = 2\varepsilon \in (0, 1/3)$, it is sufficient to prove that

$$\sup_{p \in \mathscr{P}} \limsup_{n \to +\infty} \mathbb{P}_{\mathbf{Y}_n|p} \left(\left| \frac{M_{n,0}}{\hat{T}_{n,0}} - 1 \right| \geq \bar{\varepsilon} \right) > C_{\bar{\varepsilon}}, \tag{12}$$

for some strictly positive constant $C_{\bar{\varepsilon}}$.

The main idea of the proof is as follows: we lower bound the supremum over $\mathscr{P}$ in (11) by an average with respect to a (carefully chosen) prior for p; we swap the conditional distribution of $\mathbf{Y}_n|p$ and the marginal of p with the conditional of $p|\mathbf{Y}_n$ and the marginal of $\mathbf{Y}_n$; finally we lower bound the event probability with respect to the posterior of p given

$\mathbf{Y}_n$. Let us formalize the proof.
In the sequel, we denote by $\mathbb{E}_p$ the expectation with respect to the prior for p, $\mathbb{E}_{\mathbf{Y}_n}$ the expectation with respect to the marginal distribution of $\mathbf{Y}_n$ and $\mathbb{P}_{p|\mathbf{Y}_n}$ the probability under the posterior of p given $\mathbf{Y}_n$. We concentrate on the left hand side of (12), which satisfies

$$\begin{aligned}\sup_{p\in\mathscr{P}} \limsup_{n\to+\infty} \mathbb{P}_{\mathbf{Y}_n|p}\left(\left|\frac{M_{n,0}}{\hat{T}_{n,0}}-1\right|\geq\bar{\varepsilon}\right) &\geq \mathbb{E}_p\left[\limsup_{n\to+\infty}\mathbb{P}_{\mathbf{Y}_n|p}\left(\left|\frac{M_{n,0}}{\hat{T}_{n,0}}-1\right|\geq\bar{\varepsilon}\right)\right]\\ &\geq \limsup_{n\to+\infty}\mathbb{E}_{\mathbf{Y}_n}\left[\mathbb{P}_{p|\mathbf{Y}_n}\left(\left|\frac{M_{n,0}}{\hat{T}_{n,0}}-1\right|\geq\bar{\varepsilon}\right)\right]. \qquad (13)\end{aligned}$$

where we have applied reverse Fatou's lemma to exchange the lim sup with the expected value.

Our choice of the nonparametric prior for p is based on completely random measures (see Daley and Vere-Jones (2008)) and the generalized Indian Buffet process prior of James (2017). In particular, a prior for $p \in \mathscr{P}$ can be defined through a completely random measure $\tilde{N}(\cdot) = \sum_j s_j\delta_{F_j}(\cdot)$ on $[0,1]$, where $(\{s_j, F_j\})_{j\geq 1}$ is a Poisson point process on $\mathbb{R}^+ \times [0,1]$, by setting $p(\cdot) = \sum_j (1-e^{-s_j})\delta_{F_j}(\cdot) \in \mathscr{P}$. We select $\tilde{N}$ to be a completely random measure with Lévy intensity $\nu(\mathrm{d}s,\mathrm{d}F) = e^{-s}/s\,\mathrm{d}s\mathbb{1}_{(0,1)}(F)\mathrm{d}F$. The distribution of $\tilde{N}$ is completely characterized by its Laplace functional defined as,

$$\mathbb{E}\left[e^{-\int_{[0,1]} f(F)\tilde{N}(\mathrm{d}F)}\right] = \exp\left\{-\int_{\mathbb{R}^+\times[0,1]}(1-e^{-sf(F)})\nu(\mathrm{d}s,\mathrm{d}F)\right\}, \qquad (14)$$

for any measurable function $f : [0,1] \to \mathbb{R}^+$. See also Kingman (1993).

Theorem 3.1 of James (2017) provides us with a distributional equality for the posterior of $\tilde{N}$ conditionally on the observed sample $\mathbf{Y}_n$. Denoting by $F_1^*, \ldots, F_{k_n}^*$ the k_n distinct features out of $\mathbf{Y}_n$, the following distributional equality holds

$$\tilde{N}|\mathbf{Y}_n \stackrel{d}{=} \tilde{N}_n + \sum_{\ell=1}^{k_n} J_\ell\delta_{F_\ell^*} \qquad (15)$$

where the J_ℓ's are non-negative random jumps (see (James, 2017, Equation (3.4)) for their definition) and $\tilde{N}_n$ is a completely random measure with updated Lévy intensity $\nu_n(\mathrm{d}s,\mathrm{d}F) = e^{-sn}\nu(\mathrm{d}s,\mathrm{d}F)$, independent of $(J_\ell, F_\ell^*)_{\ell=1,\ldots,k_n}$ and of the observations $\mathbf{Y}_n$. In the sequel we will further denote by $(s'_j)_{j\geq 1}$ and $(F'_j)_{j\geq 1}$ the jumps and atoms of $\tilde{N}_n$, respectively.

Defining $A_n := \{F_1^*, \dots, F_{k_n}^*\}$, from (15) we have that, for any Borel set B in $\mathbb{R}^+$, the missing mass $M_{n,0}$ satisfies

$$\begin{aligned}\mathbb{P}_{p|\mathbf{Y}_n}(M_{n,0} \in B) &= \mathbb{P}_{p|\mathbf{Y}_n}\left(\sum_{j\geq 1} p_j \delta_{F_j}(A_n^c) \in B\right) \\ &= \mathbb{P}_{\tilde{N}|\mathbf{Y}_n}\left(\sum_{j\geq 1}(1-e^{-s_j})\delta_{F_j}(A_n^c) \in B\right) \\ &\overset{(15)}{=} \mathbb{P}_{\tilde{N}_n}\left(\sum_{j\geq 1}(1-e^{-s'_j}) \in B\right)\end{aligned} \tag{16}$$

showing that the posterior distribution of the missing mass $M_{n,0}$ is equal in distribution to the random variable $\sum_{j\geq 1}(1-e^{-s'_j})$, where $(s'_j)_{j\geq 1}$ are the jumps of $\tilde{N}_n$. Unfortunately, the density of $M_{n,0}$ is not available in closed form, nevertheless, thanks to Lemma 14, we know that

$$S_n := \sum_{j\geq 1} s'_j = \tilde{N}_n([0,1])$$

is gamma distributed with parameters $(1, n+1)$. Besides, Lemma 15, which builds on the fact that $1-e^{-x}$ is asymptotically equivalent to x as $x \to 0$, allows us to replace $M_{n,0}$ with S_n in our calculations. The rest of the proof will consist of two steps: i) by Lemma 15, we will show that any consistent estimator of $M_{n,0}$ is also a consistent estimator of S_n; ii) exploiting the closed form probability distribution function of S_n and the fact that S_n is independent of the observations $\mathbf{Y}_n$, we prove that there are no consistent estimators of this random variable for the multiplicative loss.
More formally, let $\bar{\varepsilon} \in (0, 1/3)$, the inverse triangular inequality entails

$$\begin{aligned}\left|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1\right| &= \left|\frac{M_{n,0}}{S_n}\left(\frac{S_n}{\hat{T}_{n,0}} - 1 + 1\right) - 1\right| \geq \left|\frac{M_{n,0}}{S_n}\left|\frac{S_n}{\hat{T}_{n,0}} - 1\right| - \left|\frac{M_{n,0}}{S_n} - 1\right|\right| \\ &\geq \frac{M_{n,0}}{S_n}\left|\frac{S_n}{\hat{T}_{n,0}} - 1\right| - \left|\frac{M_{n,0}}{S_n} - 1\right|\end{aligned} \tag{17}$$

which implies

$$\mathbb{P}_{p|\mathbf{Y}_n}\left(1 - \frac{\bar{\varepsilon}}{2} \leq \frac{M_{n,0}}{S_n} \leq 1\,,\, \left|\frac{S_n}{\hat{T}_{n,0}} - 1\right| > 3\bar{\varepsilon}\right) \leq \mathbb{P}_{p|\mathbf{Y}_n}\left(\left|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1\right| > \bar{\varepsilon}\right). \tag{18}$$

Indeed, thanks to (17), the two events together

$$1 - \frac{\bar{\varepsilon}}{2} \leq \frac{M_{n,0}}{S_n} \leq 1\,,\, \left|\frac{S_n}{\hat{T}_{n,0}} - 1\right| > 3\bar{\varepsilon}$$

imply that

$$\left|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1\right| \geq \frac{M_{n,0}}{S_n}\left|\frac{S_n}{\hat{T}_{n,0}} - 1\right| - \left|\frac{M_{n,0}}{S_n} - 1\right| \geq \left(1 - \frac{\bar{\varepsilon}}{2}\right)3\bar{\varepsilon} - \frac{\bar{\varepsilon}}{2} = \bar{\varepsilon}(5 - 3\bar{\varepsilon})/2 > \bar{\varepsilon}$$

where the last inequality follows from the fact that $\bar{\varepsilon} < 1$. Hence, from (18), we have that

$$\mathbb{P}_{p|\mathbf{Y}_n}\left(\left|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1\right| > \bar{\varepsilon}\right) \geq \mathbb{P}_{p|\mathbf{Y}_n}\left(1 - \frac{\bar{\varepsilon}}{2} \leq \frac{M_{n,0}}{S_n} \leq 1\right) - 1 + \mathbb{P}_{p|\mathbf{Y}_n}\left(\left|\frac{S_n}{\hat{T}_{n,0}} - 1\right| > 3\bar{\varepsilon}\right)$$

which may be plugged into (13) to obtain

$$\begin{aligned}\sup_{p\in\mathscr{P}} \limsup_{n\to+\infty} \mathbb{P}_{\mathbf{Y}_n|p}&\left(\left|\frac{\hat{T}_{n,0}}{M_{n,0}} - 1\right| \geq \bar{\varepsilon}\right) \\ &\geq \limsup_{n\to+\infty} \mathbb{E}_{\mathbf{Y}_n}\left[\mathbb{P}_{p|\mathbf{Y}_n}\left(1 - \frac{\bar{\varepsilon}}{2} \leq \frac{M_{n,0}}{S_n} \leq 1\right) - 1\right] \\ &\qquad + \inf_{\mathbf{Y}_n} \inf_{x>0} \mathbb{P}_{p|\mathbf{Y}_n}\left(\left|\frac{S_n}{x} - 1\right| > 3\bar{\varepsilon}\right).\end{aligned} \tag{19}$$

We are going to lower bound the r.h.s. of (19). With regard to the first term, Lemma 15 gives

$$\mathbb{P}_{p|\mathbf{Y}_n}\left(1 - \frac{\bar{\varepsilon}}{2} \leq \frac{M_{n,0}}{S_n} \leq 1\right) - 1 \geq -e^{-\bar{\varepsilon}(n+1)}. \tag{20}$$

Let us now consider the second term on the r.h.s. of (19). Using again the fact that S_n is gamma distributed and $\bar{\varepsilon} < 1/3$, we have

$$\begin{aligned}\mathbb{P}_{p|\mathbf{Y}_n}\left(\left|\frac{S_n}{x} - 1\right| > 3\bar{\varepsilon}\right) &= 1 - (n+1)\int_{(1-3\bar{\varepsilon})x}^{(1+3\bar{\varepsilon})x} e^{-s(n+1)}\mathrm{d}s \\ &= 1 + e^{-(1+3\bar{\varepsilon})x(n+1)} - e^{-(1-3\bar{\varepsilon})x(n+1)} \\ &\geq \inf_{y>0}\left[1 + e^{-(1+3\bar{\varepsilon})y} - e^{-(1-3\bar{\varepsilon})y}\right]\end{aligned}$$

it is now easy to see that the function $f(y) := 1 + e^{-(1+3\bar{\varepsilon})y} - e^{-(1-3\bar{\varepsilon})y}$ is strictly positive, continuous and admits a global minimum on $\mathbb{R}^+$ at the point $y^* = \log((1+3\bar{\varepsilon})/(1-3\bar{\varepsilon}))/(6\bar{\varepsilon})$, therefore

$$\mathbb{P}_{p|\mathbf{Y}_n}\left(\left|\frac{S_n}{x} - 1\right| > 3\bar{\varepsilon}\right) \geq f(y^*) =: C_{\bar{\varepsilon}} > 0. \tag{21}$$

Replacing the two bounds (20) and (21) in (19), for any $\bar{\varepsilon} \in (0, 1/3)$ we get

$$\sup_{(p_j)_{j\geq 1}\in\mathscr{P}} \lim_{n\to+\infty} \mathbb{P}\left(\left|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1\right| \geq \bar{\varepsilon}\right) \geq -\limsup_{n\to+\infty} e^{-\bar{\varepsilon}(n+1)} + C_{\bar{\varepsilon}} = C_{\bar{\varepsilon}} > 0$$

which completes the proof.

Remark 2 *We point out that the prior used to prove inconsistency almost surely samples measures with an approximately exponential tail decay. Indeed,* $(\log(1-p_j))_{j\geq 1}$ *is a gamma Process. Therefore, if we suppose that the* $(p_j)_j$ *are ordered, Kingman (1975) equation (65) implies that, almost surely,*

$$\log p_j \sim -jC.$$

As a consequence, for any estimator $\hat{T}_{n,0}$, there is a vector $(p_j)_{j\geq 1}$ with light tail for which the estimation of the missing mass is inconsistent for the multiplicative loss. Hence, we need to restrict our study to vectors of probabilities with "slow" enough tail decay in order to find consistent estimators of the missing mass. This will be investigated and formalized in Section 5.

4. Concentration inequalities for feature sampling

In this section we will establish exponential tail bounds for the small masses $M_{n,r}$ and the statistic $K_{n,r}$ defined by

$$K_{n,r} = \sum_{j\geq 1} \mathbb{1}_{\{X_{n,j}=r\}}, \quad \text{for } r \geq 1$$

which counts the number of features observed with frequency r in the sample $\mathbf{Y}_n$. The statistic $K_{n,r}$ is of interest in different applications of feature allocation models. We also define $K_n := \sum_{r\geq 1} K_{n,r}$, which represents the total number of distinct features out of the initial sample. The concentration inequalities, that we are going to state, will be exploited in Section 5 to prove the multiplicative consistency of the proposed estimator of $M_{n,r}$ under the assumption of regularly varying heavy tails $(p_j)_{j\geq 1}$, for any fixed $r \geq 0$. We also emphasize that our tail bounds are valid in full generality, i.e., without imposing further assumptions on the probability masses $(p_j)_{j\geq 1}$.

In order to derive the concentration inequalities for $K_{n,r}$ (resp. $M_{n,r}$) we will exploit Chernoff bounds, which require suitable inequalities on the logarithmic moment generating function (or log-Laplace transform) of $K_{n,r}$ (resp. $M_{n,r}$). To this end, we now remind some useful definitions from Boucheron et al. (2013) and Ben-Hamou et al. (2017) which serve to compare the tail behaviour of a generic random variable with respect to some reference distributions (Gaussian, gamma and Poisson).

Definition 3 *Let X be a real valued random variable defined on some probability space, then:*

i. X is sub-Gaussian on the right tail (resp. on the left tail) with variance factor v if for any $\lambda \geq 0$ (resp. $\lambda \leq 0$)

$$\log \mathbb{E}\left(e^{\lambda(X-\mathbb{E}[X])}\right) \leq \frac{v\lambda^2}{2}; \tag{22}$$

ii. X is sub-gamma on the right tail with variance factor v and scale parameter c if

$$\log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \frac{\lambda^2 v}{2(1-c\lambda)}, \quad \text{for any } \lambda \text{ satisfying } 0 \leq \lambda \leq 1/c; \tag{23}$$

iii. X is sub-gamma on the left tail with variance factor v and scale parameter c if $-X$ is sub-gamma on the right tail with variance factor v and scale parameter c;

iv. X is sub-Poisson with variance factor v if for all $\lambda \in \mathbb{R}$

$$\log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \phi(\lambda)v, \tag{24}$$

where $\phi(\lambda) = e^{\lambda} - 1 - \lambda$.

Note that a sub-Gaussian random variable is also sub-gamma for any choice of the scale parameter c, but in general the inverse is not true. As proved in the sequel, the bounds on the log-Laplace transforms (22)–(23) imply exponential tails bounds by means of the Chernoff inequality. See Boucheron et al. (2013) for other details. In Proposition 4, we prove exponential tail bounds for the moment generating function of the small mass $M_{n,r}$, showing that $M_{n,r}$ is sub-Gaussian on the left tail and sub-gamma on the right one. These bounds are then applied in Corollary 6 to derive concentration inequalities for $M_{n,r}$.

Proposition 4 *Consider an integer number r such that $0 \leq r < n-2$. On the left tail, the random variable $M_{n,r}$ is sub-Gaussian with variance factor $v^-_{n,r} := (r+1)(r+2)\mathbb{E}[K_{n+2,r+2}]/((n+1)(n+2))$, i.e., for any $\lambda \leq 0$ it holds*

$$\log \mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right] \leq \frac{\lambda^2 v^-_{n,r}}{2}. \tag{25}$$

On the right tail, the random variable $M_{n,r}$ satisfies

$$\log \mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right] \leq v^+_{n,r}\left\{\frac{1}{(1-\lambda/(n-r))^{r+1}} - 1 - \frac{\lambda(r+1)}{n-r}\right\}, \tag{26}$$

for any $0 \leq \lambda < (n-r)$, where $v^+_{n,r}$ is defined by

$$v^+_{n,r} := \frac{\mathbb{E}[K_{n-r}]\, n!}{(1-2/(n-r))(n-r)!(n-r)^r}.$$

Proof We first focus on the proof of (25). Let $\lambda \leq 0$, exploiting the independence of the random variables $X_{n,j}$'s and the elementary inequality $\log(z) \leq z-1$, valid for any $z > 0$, we obtain

$$\begin{aligned}
\log \mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right] &= \sum_{j\geq 1} \log \mathbb{E}\left[e^{\lambda(p_j \mathbb{1}_{\{X_{n,j}=r\}} - p_j \mathbb{P}(X_{n,j}=r))}\right] \\
&= \sum_{j\geq 1}\Big(-\lambda p_j \mathbb{P}(X_{n,j}=r) + \log(1+(e^{\lambda p_j}-1)\mathbb{P}(X_{n,j}=r))\Big) \\
&\leq \sum_{j\geq 1} \mathbb{P}(X_{n,j}=r)(e^{\lambda p_j}-1-\lambda p_j).
\end{aligned}$$

Since $\lambda \leq 0$ we get that

$$\begin{aligned}
\log \mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right] &\leq \sum_{j\geq 1} \mathbb{P}(X_{n,j}=r)\frac{(\lambda p_j)^2}{2} = \frac{\lambda^2}{2}\sum_{j\geq 1} p_j^2 \mathbb{P}(X_{n,j}=r) \\
&= \frac{\lambda^2}{2}\sum_{j\geq 1}\binom{n}{r} p_j^{r+2}(1-p_j)^{n-r} = \frac{\lambda^2}{2}\cdot\frac{n!(r+2)!}{r!(n+2)!}\mathbb{E}[K_{n+2,r+2}] = \frac{\lambda^2 v^-_{n,r}}{2}
\end{aligned}$$

then (25) now follows.

As for the second inequality (26), we can use the previous calculations to state that for any $\lambda \in [0, n-r)$

$$\log \mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right] \leq \sum_{j\geq 1} \mathbb{P}(X_{n,j}=r)(e^{\lambda p_j}-1-\lambda p_j) = \sum_{j\geq 1}\mathbb{P}(X_{n,j}=r)\sum_{k\geq 2}\frac{(\lambda p_j)^k}{k!}$$

$$= \sum_{k\geq 2}\sum_{j\geq 1}\mathbb{P}(X_{n,j}=r)\frac{(\lambda p_j)^k}{k!} = \sum_{k\geq 2}\sum_{j\geq 1}\binom{n}{r}p_j^r(1-p_j)^{n-r}\frac{(\lambda p_j)^k}{k!}$$
$$= \sum_{k\geq 2}\binom{n}{r}\frac{\lambda^k}{k!}\sum_{j\geq 1}p_j^{k+r}(1-p_j)^{n-r}$$

where we have used the Taylor series expansion of the exponential function. We now observe that $(1-p_j)^{n-r}\leq e^{-(n-r)p_j}$ and the bound is pretty accurate when n is high, then we get

$$\log\mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right] \leq \sum_{k\geq 2}\binom{n}{r}\frac{\lambda^k}{k!}\sum_{j\geq 1}p_j^{k+r}e^{-(n-r)p_j}$$
$$= \frac{\binom{n}{r}}{(n-r)^r}\sum_{k\geq 2}\left(\frac{\lambda}{n-r}\right)^k\frac{(k+r)!}{k!}\sum_{j\geq 1}\frac{(p_j(n-r))^{k+r}}{(k+r)!}e^{-(n-r)p_j}.$$

By observing that for any $k\geq 2$ the upper bound holds true

$$\frac{(p_j(n-r))^{k+r}}{(k+r)!}e^{-(n-r)p_j}\leq (e^{(n-r)p_j}-1)e^{-(n-r)p_j} = 1-e^{-(n-r)p_j},$$

we obtain the following inequality for the log-Laplace functional of our interest

$$\log\mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right]\leq \frac{\binom{n}{r}}{(n-r)^r}\sum_{k\geq 2}\left(\frac{\lambda}{n-r}\right)^k\frac{(k+r)!}{k!}\sum_{j\geq 1}[1-e^{-(n-r)p_j}].$$

Fixing the useful notation

$$\Phi_n := \sum_{j\geq 1}(1-e^{-np_j})$$

and evaluating the sum over k using the Euler gamma function, we get

$$\log\mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right]\leq \binom{n}{r}\frac{\Phi_{n-r}}{(n-r)^r}\sum_{k\geq 2}\left(\frac{\lambda}{n-r}\right)^k\frac{\Gamma(k+r+1)}{k!}$$
$$= \binom{n}{r}\frac{\Phi_{n-r}}{(n-r)^r}\sum_{k\geq 2}\left(\frac{\lambda}{n-r}\right)^k\frac{1}{k!}\int_0^\infty e^{-x}x^{k+r}\mathrm{d}x$$
$$= \binom{n}{r}\frac{\Phi_{n-r}}{(n-r)^r}\int_0^\infty e^{-x}x^r\sum_{k\geq 2}\left(\frac{\lambda x}{n-r}\right)^k\frac{1}{k!}\mathrm{d}x$$
$$= \binom{n}{r}\frac{\Phi_{n-r}}{(n-r)^r}\int_0^\infty e^{-x}x^r\left(e^{\lambda x/(n-r)}-1-\frac{\lambda x}{n-r}\right)\mathrm{d}x$$

where all the previous equalities are valid whenever $0\leq\lambda<(n-r)$. By solving the integrals on the right hand side of the last term in the previous chain of equations, we get

$$\log\mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right]\leq \frac{r!\Phi_{n-r}}{(n-r)^r}\binom{n}{r}\left\{\frac{1}{(1-\lambda/(n-r))^{r+1}}-1-\frac{(r+1)\lambda}{n-r}\right\}$$

$$= \frac{n!\Phi_{n-r}}{(n-r)!(n-r)^r}\left\{\frac{1}{(1-\lambda/(n-r))^{r+1}} - 1 - \frac{(r+1)\lambda}{n-r}\right\} \tag{27}$$

Proceeding along similar lines as in (Gnedin et al., 2007, Lemma 1), it is not difficult to see that for any $n > 2$

$$|\Phi_n - \mathbb{E}[K_n]| \leq \frac{2}{n}\Phi_{n,2} \leq \frac{2}{n}\Phi_n, \tag{28}$$

which entails $\Phi_n \leq \mathbb{E}[K_n]/(1-2/n)$, for any $n > 2$. The last inequality can be used to provide an upper bound for the r.h.s. of (27). To this end we can apply (28) with n replaced with $n-r$ to obtain

$$\Phi_{n-r} \leq \frac{\mathbb{E}[K_{n-r}]}{(1-2/(n-r))}$$

for any $n > r+2$. We use the previous inequality to bound Φ_{n-r} which appears in (27), and (26) easily follows. ■

We now specialize Proposition 4 when $r = 0$, proving that the missing mass $M_{n,0}$ is sub-Gaussian on the left tail and sub-gamma on the right one.

Proposition 5 *Let $n > 2$. On the left tail, the random variable $M_{n,0}$ is sub-Gaussian with variance factor $v_n^- := 2\mathbb{E}[K_{n+2,2}]/((n+2)\cdot(n+1))$, i.e., for any $\lambda \leq 0$ it holds*

$$\log \mathbb{E}\left[e^{\lambda[M_{n,0}-\mathbb{E}[M_{n,0}]]}\right] \leq \frac{\lambda^2 v_n^-}{2}. \tag{29}$$

On the right tail, the random variable $M_{n,0}$ is sub-gamma with variance factor $v_n^+ := 2\mathbb{E}[K_n]/(n^2-2n)$ and scale parameter $1/n$, i.e., for any $0 \leq \lambda < n$ one has

$$\log \mathbb{E}\left[e^{\lambda[M_{n,0}-\mathbb{E}[M_{n,0}]]}\right] \leq \frac{\lambda^2 v_n^+}{2(1-\lambda/n)}. \tag{30}$$

Proof It is sufficient to apply Proposition 4 when $r = 0$. ■

As already mentioned at the beginning of this section, the bounds on the log-Laplace obtained in Proposition 4 imply useful exponential tail bounds for $M_{n,r}$ which can be obtained via the Cramér-Chernoff method (see Boucheron et al. (2013)). More specifically we can state and prove the following

Corollary 6 *For any r satisfying $1 \leq r \leq n-2$ and $x \geq 0$, then*

$$\mathbb{P}(M_{n,r} - \mathbb{E}[M_{n,r}] \leq -x) \leq \exp\left\{-\frac{x^2}{2v_{n,r}^-}\right\}$$

and

$$\mathbb{P}(M_{n,r} - \mathbb{E}[M_{n,r}] \geq x)$$
$$\leq \exp\left\{-\left[\lambda_{\max}(x)x - v_{n,r}^+\left(\left(1+\frac{x(n-r)}{(r+1)v_{n,r}^+}\right)^{\frac{r+1}{r+2}} - 1 - \frac{\lambda_{\max}(x)(r+1)}{n-r}\right)\right]\right\}$$

where

$$\lambda_{\max}(x) = (n-r)\left[1-\left(1+\frac{x(n-r)}{(r+1)v_{n,r}^+}\right)^{-1/(r+2)}\right]. \tag{31}$$

Proof The first concentration follows from the fact that $M_{n,r}$ is a sub-Gaussian random variable on the left tail (see (Boucheron et al., 2013, Section 2.3)). In order to prove the second inequality we may use the general Cramér-Chernoff method as described in (Boucheron et al., 2013, Section 2.2). In Proposition 4 we have proved (26), bounding the log-Laplace of $M_{n,r} - \mathbb{E}[M_{n,r}]$ for any for any $0 \leq \lambda < (n-r)$. More precisely

$$\log \mathbb{E}\left[e^{\lambda[M_{n,r}-\mathbb{E}[M_{n,r}]]}\right] \leq v_{n,r}^{+}\left\{\frac{1}{(1-\lambda/(n-r))^{r+1}} - 1 - \frac{\lambda(r+1)}{n-r}\right\} =: \psi_{n,r}(\lambda), \tag{32}$$

for any $0 \leq \lambda < (n-r)$. The Cramér-Chernoff method prescribes to determine the Legendre transform of $\psi_{n,r}$, i.e.,

$$\psi_{n,r}^{*}(x) := \sup_{\lambda \geq 0}\{\lambda x - \psi_{n,r}(\lambda)\},$$

and it gives the so-called Chernoff inequality

$$\mathbb{P}(M_{n,r} - \mathbb{E}[M_{n,r}] \geq x) \leq \exp\{-\psi_{n,r}^{*}(x)\}. \tag{33}$$

We need only to prove that (33) coincides with the concentration inequality in the statement. By some elementary calculations, it is not difficult to see that the function $\lambda x - \psi_{n,r}(\lambda)$ attains its maximum over $\lambda \in (0, n-r)$ at the point

$$\lambda_{\max}(x) = (n-r)\left[1 - \left(1 + \frac{x(n-r)}{(r+1)v_{n,r}^{+}}\right)^{-1/(r+2)}\right],$$

hence

$$\begin{aligned}
\psi_{n,r}^{*}(x) &= \lambda_{\max}(x)x - \psi_{n,r}(\lambda_{\max}(x)) \\
&= \lambda_{\max}(x)x - v_{n,r}^{+}\left(\frac{1}{(1-\lambda_{\max}(x)/(n-r))^{r+1}} - 1 - \frac{\lambda_{\max}(x)(r+1)}{n-r}\right) \\
&= \lambda_{\max}(x)x - v_{n,r}^{+}\left(\left(1 + \frac{x(n-r)}{(r+1)v_{n,r}^{+}}\right)^{\frac{r+1}{r+2}} - 1 - \frac{\lambda_{\max}(x)(r+1)}{n-r}\right)
\end{aligned}$$

and putting this expression in (33) the second concentration inequality is proved, as well. ■

Due to the importance of the missing mass $M_{n,0}$ in several applications, we also specialize Corollary 6 to the case $r = 0$.

Corollary 7 *For any $n > 2$ and $x \geq 0$, the following probability bounds hold true*

$$\begin{aligned}
\mathbb{P}(M_{n,0} - \mathbb{E}[M_{n,0}] \leq -x) &\leq \exp\left\{-\frac{x^2}{2v_n^{-}}\right\}, \\
\mathbb{P}(M_{n,0} - \mathbb{E}[M_{n,0}] \geq x) &\leq \exp\left\{-v_n^{+}n^2\left[1 + \frac{x}{nv_n^{+}} - \sqrt{1 + \frac{x}{nv_n^{+}}}\right]\right\}.
\end{aligned}$$

Proof The two inequalities follow by the Chernoff bound and the log-Laplace bound proved in Proposition 5. This is a standard argument, see Boucheron et al. (2013) for details. ■

Proceeding along similar lines as before we show that $K_{n,r}$ is a sub-Poisson random variable, this result is implicitly proved in the Supplementary material by Ayed et al. (2019), but for the sake of completeness we report it here as well.

Proposition 8 *For any $r \geq 1$ and $n \geq 1$, the random variable $K_{n,r}$ is sub-Poisson with variance factor $\mathbb{E}[K_{n,r}]$. Indeed, for any $\lambda \in \mathbb{R}$ the following bound holds true*

$$\log \mathbb{E}[e^{\lambda(K_{n,r}-\mathbb{E}[K_{n,r}])}] \leq \phi(\lambda)\mathbb{E}[K_{n,r}], \tag{34}$$

where $\phi(\lambda) := e^{\lambda} - 1 - \lambda$.

Proof Exploiting the independence of the random variables $X_{n,j}$'s, for any $\lambda \in \mathbb{R}$ we can write:

$$\begin{aligned}
\log \mathbb{E}[e^{\lambda(K_{n,r}-\mathbb{E}[K_{n,r}])}] &= \sum_{j=1}^{\infty} \log \mathbb{E} \exp\left\{\lambda(\mathbb{1}_{\{X_{n,j}=r\}} - \mathbb{E}\mathbb{1}_{\{X_{n,j}=r\}})\right\} \\
&= \sum_{j=1}^{\infty} \left\{-\lambda \mathbb{P}(X_{n,j}=r) + \log(e^{\lambda}\mathbb{P}(X_{n,j}=r) + 1 - \mathbb{P}(X_{n,j}=r))\right\} \\
&\leq \sum_{j=1}^{\infty} \phi(\lambda)\mathbb{P}(X_{n,j}=r) = \phi(\lambda)\mathbb{E}[K_{n,r}]
\end{aligned}$$

where we have used the inequality $\log(z) \leq z - 1$, for any $z > 0$. ■

The previous proposition and the Chernoff bounds imply an exponential tail bound for $K_{n,r}$, indeed one can prove that

Corollary 9 *For any $n \geq 1$, $r \geq 1$ and $x \geq 0$ the following tail bound holds true*

$$\mathbb{P}(|K_{n,r} - \mathbb{E}[K_{n,r}]| \geq x) \leq 2\exp\left\{-\frac{x^2}{2(\mathbb{E}[K_{n,r}] + x/3)}\right\}. \tag{35}$$

Corollaries 7 and 9 provide us with concentration inequalities for the small masses $M_{n,r}$ and the statistics $K_{n,r}$, respectively, around their means, for any $r \geq 0$. These results hold true in general, without any further assumption on the probabilities $(p_j)_{j\geq 1}$. In the next Section, we will focus on the restricted class of regularly varying probabilities, and we define a nonparametric estimator of the small mass $M_{n,r}$, also considered by Ayed et al. (2018) for the special case $r = 0$. Thanks to the previous concentration inequalities, we are able to give provable guarantees in terms of consistency for that estimator of $M_{n,r}$ within the restricted class of regularly varying probabilities.

5. A consistent estimator for regularly varying feature probabilities

We now introduce a nonparametric estimator of the small mass $M_{n,r}$ in a quite natural way, i.e., by comparing expectations. We simply evaluate the expected value of $M_{n,r}$ and

we obtain

$$\begin{aligned}\mathbb{E}[M_{n,r}] &= \mathbb{E}\left(\sum_{j\geq 1} p_j \mathbb{1}_{\{X_{n,j}=r\}}\right) = \sum_{j\geq 1} p_j \mathbb{P}(X_{n,j}=r) = \binom{n}{r}\sum_{j\geq 1} p_j^{r+1}(1-p_j)^{n-r} \\ &= \frac{r+1}{n+1}\sum_{j\geq 1}\mathbb{P}(X_{n+1,j}=r+1) = \frac{r+1}{n+1}\mathbb{E}[K_{n+1,r+1}].\end{aligned} \tag{36}$$

By comparing the expected values of the first and the last term in (36) we are led to define the following nonparametric estimator of the small mass $M_{n,r}$

$$\hat{M}_{n,r} := \frac{r+1}{n} K_{n,r+1} \tag{37}$$

where $n+1$ has been replaced by n so that all the quantities are computable at time n. It is worth to stress that $\hat{M}_{n,r}$ is a *nonparametric* estimator of the mass of features with frequency r in the sample, since it does not rely on any parametric assumption on the p_j's. Moreover, it is a feasible quantity form a computational standpoint, being easy to implement and evaluate for any value of n. The (37) has the same parametric form of the Good–Turing estimator (Good, 1953), however the two estimators have different ranges: while the estimator of the missing mass in species sampling takes values in $[0,1]$, the same estimator in the feature sampling framework takes positive values. Finally $\hat{M}_{n,0}$ has been already introduced by Ayed et al. (2018), who have extensively discussed its interpretations both as a Jackknife estimator in the sense of Quenouille (1956) and as a non-parametric empirical Bayes estimator in the same spirit as Efron and Morris (1973). A somewhat related derivation of (37) can be also found in Chao and Colwell (2017).

Here we want to study the consistency of (37). In Section 3 we proved that, without imposing further assumptions on the features' proportions, any estimator of the missing mass is inconsistent for at least one choice of the proportions (Theorem 1). We then study the consistency of (37) under the ubiquitous assumption of heavy tailed probabilities $(p_j)_{j\geq 1}$. We rely on the theory of regular variation by Karamata, J. (1930, 1933) (see also Karlin (1967)) to define a suitable class of heavy-tailed $(p_j)_{j\geq 1}$, showing that, under this class, $\hat{M}_{n,r}$ turns out to be multiplicative consistent, for any $r \geq 0$.

We use the limiting notation $f \simeq g$ to mean $f/g \to 1$; we further write $f \lesssim g$ if there exists a fixed constant $C > 0$ such that $f \leq Cg$. As in Karlin (1967) we give the following definition.

Definition 10 *Let $\nu(\mathrm{d}x) := \sum_{i\geq 1}\delta_{p_i}(\mathrm{d}x)$ and define the measure $\overline{\nu}(x) := \nu[x,1]$, which is the cumulative count of all features having no less than a certain probability mass. We say that $(p_j)_{j\geq 1}$ is regularly varying with regular variation index $\alpha \in (0,1)$ if $\overline{\nu}(x) \simeq x^{-\alpha}\ell(1/x)$ as $x \downarrow 0$, where $\ell(t)$ is a slowly varying function, that is $\ell(ct)/\ell(t) \to 1$ as $t \to +\infty$ for all $c > 0$.*

Let us remark that if we denote $(p_{[j]})_{j\geq 1}$ the sorted probabilities in decreasing order, definition 10 is equivalent to

$$p_{[j]} \simeq j^{-1/\alpha}\ell_*(j),$$

as $j \to \infty$, where ℓ_* is another slowly varying function. For simplicity, the relation between ℓ, ℓ_* and α is skipped here, interested readers can refer to Lemma 22 and Proposition 23

of Gnedin et al. (2007). Definition 10 is in the same spirit as Karlin (1967), but for our purposes here we consider the case $\sum_{j\geq 1} p_j < +\infty$, while in Karlin (1967) the p_j's satisfy the more restrictive condition $\sum_{j\geq 1} p_j = 1$. The next theorem is similar to a result proved by Karlin (1967) and provides the first order asymptotic of $\mathbb{E}K_{n,r}$, used later to prove consistency of $\hat{M}_{n,r}$.

Theorem 11 *Let $(p_j)_{j\geq 1}$ be regularly varying with $\alpha \in (0,1)$. If $\Gamma(\cdot)$ denotes the gamma function, then as $n \to +\infty$, $\mathbb{E}[K_{n,r}] \simeq \frac{\alpha\Gamma(r-\alpha)}{r!} n^{\alpha}\ell(n)$ and $\mathbb{E}[K_n] \simeq \Gamma(1-\alpha)n^{\alpha}\ell(n)$.*

Proof We first define the quantity

$$\Phi_{n,r} := \sum_{j\geq 1} \frac{(np_j)^r}{r!} e^{-np_j}, \quad r \geq 1,$$

which can be considered an asymptotic approximation of $\mathbb{E}K_{n,r}$. Indeed, in order to prove the theorem, we first show that $\Phi_{n,r} \simeq \frac{\alpha\Gamma(r-\alpha)}{r!} n^{\alpha}\ell(n)$ as $n \to +\infty$, and then we prove that $\Phi_{n,r} \simeq \mathbb{E}[K_{n,r}]$. In order to prove the former asymptotic equivalence it is worth noticing that (Gnedin et al., 2007, Proposition 13) applies also for the feature setting under regularly varying heavy tails, indeed the measure defined by $\nu_r(\mathrm{d}p) := p^r \nu(\mathrm{d}p)$ is such that

$$\nu_r([0,p]) \simeq \frac{\alpha}{r-\alpha} p^{r-\alpha}\ell(1/p), \qquad \text{as } p \to 0. \tag{38}$$

Since $\Phi_{n,r} = n^r/r! \int_0^1 e^{-np}\nu_r(\mathrm{d}p)$ is the Laplace transform of ν_r multiplied by a suitable quantity, we can apply Tauberian theorems to connect the asymptotic behaviour of the cumulative distribution function of ν_r given in (38) to that of $\Phi_{n,r}$. In particular, from Tauberian theorems (see Feller (1971)), we obtain

$$\Phi_{n,r} = \frac{n^r}{r!}\int_0^1 e^{-np}\nu_r(\mathrm{d}p) \simeq \frac{n^r}{r!}\alpha\Gamma(r-\alpha)n^{-(r-\alpha)}\ell(n) = \alpha\frac{\Gamma(r-\alpha)}{r!}n^{\alpha}\ell(n), \tag{39}$$

as $n \to +\infty$. As a byproduct of (39), we get $\Phi_{n,r} \to +\infty$. Finally to show $\Phi_{n,r} \simeq \mathbb{E}[K_{n,r}]$, we can easily observe that (Gnedin et al., 2007, Lemma 1) applies in this setting as well, hence there exists a constant c such that

$$|\mathbb{E}[K_{n,r}] - \Phi_{n,r}| \leq \frac{c}{n}\max\{\Phi_{n,r}, \Phi_{n,r+2}\} \to 0, \tag{40}$$

as $n \to +\infty$. From (40), along with $\Phi_{n,r} \to +\infty$, we obtain

$$\Big|\frac{\mathbb{E}[K_{n,r}]}{\Phi_{n,r}} - 1\Big| = \frac{|\mathbb{E}[K_{n,r}] - \Phi_{n,r}|}{\Phi_{n,r}} \to 0, \quad \text{as } n \to +\infty,$$

in other words we have shown that $\Phi_{n,r} \simeq \mathbb{E}[K_{n,r}]$ as $n \to +\infty$. The asymptotic for $\mathbb{E}[K_n]$ can be proved in a similar fashion. To see this, first of all define

$$\Phi_n := \sum_{j\geq 1}[1 - e^{-np_j}]$$

which can be considered as an asymptotic approximation of $\mathbb{E}[K_n]$ thanks to the inequality (28), indeed let us evaluate

$$\Big|\frac{\mathbb{E}[K_n]}{\Phi_n}-1\Big| = \frac{|\mathbb{E}[K_n]-\Phi_n|}{\Phi_n} \overset{(28)}{\leq} \frac{2}{n} \to 0$$

which implies that $\mathbb{E}[K_n] \simeq \Phi_n$ as n grows to ∞. It remains to determine the asymptotic behavior of Φ_n, applying the integration by parts formula and Tauberian theorems (Feller, 1971, Theorem 4, Section 5, Chapter 13) we obtain

$$\Phi_n = \int_0^1 (1-e^{-np})\nu(\mathrm{d}p) = \int_0^1 ne^{-np}\bar{\nu}(p)\mathrm{d}p \simeq n\,\Gamma(1-\alpha)n^{-(1-\alpha)}\ell(n).$$

therefore the conclusion follows since $\mathbb{E}[K_n] \simeq \Phi_n \simeq \Gamma(1-\alpha)n^{\alpha}\ell(n)$. ■

We are now ready to prove that $\hat{M}_{n,r}$ is multiplicative consistent for any fixed $r \geq 0$, when the feature probabilities $(p_j)_{j\geq 1}$ are regularly varying. In the proof we will employ the concentration inequalities of Section 4, which are tuned under the assumption of regular variation by means of Theorem 11. Technical details are deferred to the Appendix (see Lemma 16).

Theorem 12 *Let $(p_j)_{j\geq 1}$ be regularly varying with index $\alpha \in (0,1)$. Fix $r \geq 0$ and let $\hat{M}_{n,r} := (r+1)K_{n,r+1}/n$ be the nonparametric estimator of $M_{n,r}$ in a sample of size n, then $\hat{M}_{n,r}$ is strongly multiplicative consistent, i.e., $M_{n,r}/\hat{M}_{n,r} \xrightarrow{a.s.} 1$.*

Proof In order to prove the multiplicative consistency we first show that $K_{n,r}/\mathbb{E}[K_{n,r}] \xrightarrow{a.s.} 1$ and that $M_{n,r}/\mathbb{E}[M_{n,r}] \xrightarrow{a.s.} 1$. As for the former convergence, we can use the concentration inequality (35) given in Corollary 9, which, for any $\varepsilon > 0$, gives

$$\mathbb{P}(|K_{n,r}/\mathbb{E}[K_{n,r}]-1| \geq \varepsilon) \leq 2\exp\left\{-\frac{\varepsilon^2\mathbb{E}[K_{n,r}]}{2(1+\varepsilon/3)}\right\}. \tag{41}$$

When $\varepsilon > 0$ is fixed, we can exploit the asymptotic equivalence $\mathbb{E}[K_{n,r}] \simeq (r!)^{-1}\alpha\Gamma(r-\alpha)n^{\alpha}\ell(n)$ in Theorem 11 to state that there exists a suitable constant $C > 0$

$$\sum_{n\geq 1}\mathbb{P}(|K_{n,1}/\mathbb{E}[K_{n,1}]-1| \geq \varepsilon) \overset{(41)}{\lesssim} \sum_{n\geq 1} 2\exp\{-Cn^{\alpha}\ell(n)\} < +\infty,$$

which implies that for any $\varepsilon > 0$, $\mathbb{P}(\limsup_n(|K_{n,r}/\mathbb{E}[K_{n,r}]-1| \geq \varepsilon)) = 0$ by the first Borel–Cantelli lemma, hence $K_{n,r}/\mathbb{E}[K_{n,r}] \xrightarrow{a.s.} 1$ for any fixed $r \geq 1$.

Analogously we may use Corollary 6 to prove the almost sure convergence to 1 of the ratio $M_{n,r}/\mathbb{E}[M_{n,r}]$. Indeed, for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}(|M_{n,r}/\mathbb{E}[M_{n,r}]-1| \geq \varepsilon)\\
&\quad= \mathbb{P}(M_{n,r}-\mathbb{E}[M_{n,r}] \geq \varepsilon\mathbb{E}[M_{n,r}]) + \mathbb{P}(M_{n,r}-\mathbb{E}[M_{n,r}] \leq -\varepsilon\mathbb{E}M_{n,r})\\
&\quad\leq \exp\left\{-\frac{\varepsilon^2(\mathbb{E}[M_{n,r}])^2}{2v_{n,r}^-}\right\} + \exp\Big\{-\Big[\lambda_{\max}(\varepsilon\mathbb{E}M_{n,r})\varepsilon\mathbb{E}M_{n,r}\\
&\qquad - v_{n,r}^+\left(\left(1+\frac{\varepsilon\mathbb{E}M_{n,r}(n-r)}{(r+1)v_{n,r}^+}\right)^{\frac{r+1}{r+2}} - 1 - \frac{\lambda_{\max}(\varepsilon\mathbb{E}M_{n,r})(r+1)}{n-r}\right)\Big]\Big\},
\end{aligned} \tag{42}$$

where $\lambda_{\max}$ has been defined in (31). By observing that $\mathbb{E}[M_{n,r}] = (r+1)\mathbb{E}[K_{n+1,r+1}]/(n+1)$ and using again Theorem 11, it is not difficult to determine the asymptotic behavior of the previous upper bound as $n \to \infty$ to conclude that for any fixed $\varepsilon > 0$

$$\sum_{n\geq 1} \mathbb{P}(|M_{n,r}/\mathbb{E}[M_{n,r}] - 1| \geq \varepsilon) \lesssim \sum_{n\geq 1} \exp\{-Cn^{\alpha}\ell(n)\} < +\infty \tag{43}$$

for a suitable constant $C > 0$. See Lemma 16 for technical details. By the first Borel–Cantelli lemma, we get $M_{n,r}/\mathbb{E}[M_{n,r}] \xrightarrow{a.s.} 1$, as well.

Thanks to the previous considerations, the consistency of $\hat{M}_{n,r}$ easily follows, indeed

$$\frac{M_{n,r}}{\hat{M}_{n,r}} = \frac{M_{n,r}}{\mathbb{E}[M_{n,r}]} \cdot \frac{n\mathbb{E}[M_{n,r}]}{(r+1)\mathbb{E}[K_{n,r+1}]} \cdot \frac{\mathbb{E}[K_{n,r+1}]}{K_{n,r+1}} \xrightarrow{a.s.} 1,$$

since all the ratios on the r.h.s. converge to 1 almost surely. ■

6. Discussion

The importance of estimating the small mass $M_{n,r}$ has emerged in ecology and biological sciences (Ionita-Laza et al., 2009, 2010; Chao et al., 2014; Gravel, 2014; Zou et al., 2016; Ayed et al., 2019), and most recently in machine learning (Cai et al., 2018) within the context of learning augmented algorithms. In this paper we first studied the problem of consistent estimation of $M_{n,r}$, thus extending to the feature sampling framework some recent results on the consistent estimation of the small probability $P_{n,r}$ (Ohannessian and Dahleh, 2012; Ben-Hamou et al., 2017; Grabchak and Zhang, 2017; Mossel and Ohannessian, 2019; Ayed et al., 2018). In particular,

i) we showed that that there do not exist universally consistent estimators, in the multiplicative sense, of the missing mass $M_{n,0}$; that is, under the Bernoulli product model we prove that for any estimator $\hat{T}_{n,0}$ of $M_{n,0}$ there exists at least a choice of feature probabilities $(p_j)_{j\geq 1}$ for which $\hat{T}_{n,0}/M_{n,0}$ does not converge to 1 in probability, as $n \to +\infty$;

ii) we introduced a nonparametric estimator $\hat{M}_{n,r}$ of $M_{n,r}$ which has the same analytic form of the Good–Turing estimator of $P_{n,r}$, and we showed that $\hat{M}_{n,r}$ is strongly consistent, in the multiplicative sense, under feature probabilities with regularly varying heavy tails.

These results rely on novel exponential tail bounds for the small mass $M_{n,r}$, and for related counting statistics, which are of independent interest. To the best of our knowledge, our study is the first theoretical account to the problem of estimating the small mass $M_{n,r}$, for $r \geq 0$.

Our work paves the way to explore new research directions in the estimation of the small mass $M_{n,r}$, both theoretical and methodological. From a theoretical perspective, it remains an interesting open problem the study of the rate of consistency of the nonparametric estimator $\hat{M}_{n,r}$. With this regards, we retain that Bayesian nonparametric ideas and

techniques developed in Section 3 can be usefully exploited to show optimality of the rate of consistency of $\hat{M}_{n,r}$ under feature probabilities with regularly varying heavy tails (Ayed et al., 2018, 2019). From a methodological perspective, our results allow to extend the realm of applicability of the estimation of $M_{n,r}$. In principle, any application involving the estimation of small probabilities $P_{n,r}$ may be reconsidered under the more general feature sampling framework. For instance, exponential tail bounds introduced in Section 4 can be usefully exploited to extend the upper confidence bound multi-armed bandit strategy of Bubeck et al. (2013), based on a confidence interval for the Good–Turing estimator $\hat{P}_{n,0}$, to the feature sampling framework.

Acknowledgments

The authors are grateful to two anonymous Referees for all their comments and corrections which improved remarkably the paper. Federico Camerlenghi and 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. Federico Camerlenghi and Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022. Federico Camerlenghi is a member of the *Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni* (GNAMPA) of the *Istituto Nazionale di Alta Matematica* (INdAM). Federico Camerlenghi is also affiliated to Collegio Carlo Alberto, Piazza V. Arbarello 8, Torino, Italy, and to the Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University, Milano, Italy. Stefano Favaro is also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

Appendix A. Technical lemmas

Here we state and prove the technical lemmas used in the proof of Theorem 1 and in the proof of Theorem 12.

Lemma 13 *For every $\varepsilon \in (0, 1/6)$,*

$$\mathbb{P}(|\frac{\hat{T}_{n,0}}{M_{n,0}} - 1| \geq \varepsilon) \geq \mathbb{P}(|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1| \geq 2\varepsilon)$$

Proof Suppose that $\left|\frac{\hat{T}_{n,0}}{M_{n,0}} - 1\right| < \varepsilon$, then multiplying by $M_{n,0}$ it comes that

$$-M_{n,0}\varepsilon < \hat{T}_{n,0} - M_{n,0} < M_{n,0}\varepsilon. \tag{44}$$

From the lower bound of (44), it comes that $\hat{T}_{n,0} > (1-\varepsilon)M_{n,0} > \frac{M_{n,0}}{2}$, leading to $\frac{M_{n,0}}{\hat{T}_{n,0}} < 2$. Therefore, dividing (44) by $\hat{T}_{n,0}$ and using previous inequality successively gives

$$\left|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1\right| \leq \frac{M_{n,0}}{\hat{T}_{n,0}}\varepsilon < 2\varepsilon$$

Considering the complements of the two events, it follows that

$$\left|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1\right| \geq 2\varepsilon \Rightarrow \left|\frac{\hat{T}_{n,0}}{M_{n,0}} - 1\right| \geq \varepsilon,$$

and, as a consequence, $\mathbb{P}(|\frac{\hat{T}_{n,0}}{M_{n,0}} - 1| \geq \varepsilon) \geq \mathbb{P}(|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1| \geq 2\varepsilon)$, proving (11). ■

In the following two lemmas, we denote by $\tilde{N}_n$ a completely random measure with Lévy intensity $\nu_n(\mathrm{d}s, \mathrm{d}F) = e^{-sn}\nu(\mathrm{d}s, \mathrm{d}F) = e^{-s(n+1)}/s\,\mathrm{d}s\mathbb{1}_{(0,1)}(F)\mathrm{d}F$. We further denote $(s'_j)_{j\geq 1}$ the jumps of the completely random measure $\tilde{N}_n$.

Lemma 14 *The random variable $S_n := \tilde{N}_n([0,1])$ is gamma distributed, with parameters (1, n+1).*

Proof This result is straightforward when computing the moment generating function of S_n Indeed, for every $x \in \mathbb{R}$ we have

$$\begin{aligned}\mathbb{E}[e^{xS_n}] &= \mathbb{E}\left[\exp\left\{x\int_{[0,1]}\tilde{N}_n(\mathrm{d}F)\right\}\right] \\ &\overset{(14)}{=} \exp\left\{-\int_0^1\int_0^{+\infty}(1-e^{xs})e^{-sn}\nu(\mathrm{d}s,\mathrm{d}\theta)\right\} \\ &= \exp\left\{-\int_0^{+\infty}(1-e^{xs})\frac{e^{-s(n+1)}}{s}\mathrm{d}s\right\} = \left(1-\frac{x}{n+1}\right)^{-1},\end{aligned}$$

which is the characteristic function of a Gamma$(1, n+1)$ random variable. ■

Lemma 15 *Let $S_n := \sum_{j\geq 1} s'_j = \tilde{N}_n([0,1])$ and $M_{n,0} = \sum_{j\geq 1}(1-e^{-s'_j})$. Then for any $\bar{\varepsilon} > 0$ the following inequality holds*

$$\mathbb{P}\left(1-\frac{\bar{\varepsilon}}{2} \leq \frac{M_{n,0}}{S_n} \leq 1\right) \geq 1 - e^{-\bar{\varepsilon}(n+1)},$$

implying in particular, using Borel–Cantelli lemma, that $M_{n,0}$ and S_n are almost surely equivalent.

Proof Let us observe that the elementary inequality

$$x - x^2/2 \leq 1 - e^{-x} \leq x,$$

for $x > 0$, implies that for all $j \geq 1$

$$s'_j - \frac{1}{2}(s'_j)^2 \leq 1 - e^{-s'_j} \leq s_j.$$

Summing over j the previous equation and observing that $S_n^2 \geq \sum_{j\geq 1}(s_j')^2$, we obtain

$$S_n - \frac{1}{2}S_n^2 \leq S_n - \frac{1}{2}\sum_{j\geq 1}(s_j')^2 \leq M_{n,0} \leq S_n,$$

and dividing by S_n, we get

$$1 - \frac{1}{2}S_n \leq \frac{M_{n,0}}{S_n} \leq 1.$$

As a simple consequence of the last inequality, for any $\bar{\varepsilon}n > 0$, the event $\{S_n \leq \bar{\varepsilon}\}$ implies the validity of $\{1 - \bar{\varepsilon}/2 \leq M_{n,0}/S_n \leq 1\}$ and therefore we can upper bound the first term in (19) as follows

$$\begin{aligned}\mathbb{P}\left(1 - \frac{\bar{\varepsilon}}{2} \leq \frac{M_{n,0}}{S_n} \leq 1\right) - 1 &\geq \mathbb{P}\left(S_n \leq \bar{\varepsilon}\right) - 1 \\ &= (n+1)\int_0^{\bar{\varepsilon}} e^{-x(n+1)}\mathrm{d}x - 1 = -e^{-\bar{\varepsilon}(n+1)},\end{aligned} \tag{45}$$

where we have used the fact that the posterior distribution of S_n is Gamma$(1, n+1)$. ■

The next lemma has been used in the proof of Theorem 12.

Lemma 16 *The upper bound appearing on the r.h.s. of Equation* (42) *satisfies*

$$\begin{aligned}&\exp\left\{-\frac{\varepsilon^2(\mathbb{E}[M_{n,r}])^2}{2v_{n,r}^-}\right\} + \exp\Big\{-\Big[\lambda_{\max}(\varepsilon\mathbb{E}M_{n,r})\varepsilon\mathbb{E}M_{n,r} \\ &\quad - v_{n,r}^+\Big(\Big(1 + \frac{\varepsilon\mathbb{E}M_{n,r}(n-r)}{(r+1)v_{n,r}^+}\Big)^{\frac{r+1}{r+2}} - 1 - \frac{\lambda_{\max}(\varepsilon\mathbb{E}M_{n,r})(r+1)}{n-r}\Big)\Big]\Big\} \\ &\qquad\qquad \lesssim \exp\{-Cn^\alpha\ell(n)\},\end{aligned} \tag{46}$$

where $\lambda_{\max}$ is defined in (31)*, and C is a suitable positive constant.*

Proof As for the first exponential function in (46), one can remember that $\mathbb{E}[M_{n,r}] = (r+1)\mathbb{E}[K_{n+1,r+1}]/(n+1)$ and $v_{n,r}^- = (r+1)(r+2)\mathbb{E}[K_{n+2,r+2}]/((n+1)(n+2))$, thus an application of Theorem 11 leads us to conclude that

$$\exp\left\{-\frac{\varepsilon^2(\mathbb{E}[M_{n,r}])^2}{2v_{n,r}^-}\right\} \simeq \exp\{-\bar{C}_1 n^\alpha \ell(n)\} \tag{47}$$

as $n \to +\infty$, where $\bar{C}_1 > 0$ is a constant depending on r and ε.

We now concentrate on the second exponential function in (46). We define

$$\begin{aligned}A_n := \frac{1}{v_{n,r}^+}\Big[&\lambda_{\max}(\varepsilon\mathbb{E}M_{n,r})\varepsilon\mathbb{E}M_{n,r} \\ &- v_{n,r}^+\Big(\Big(1 + \frac{\varepsilon\mathbb{E}M_{n,r}(n-r)}{(r+1)v_{n,r}^+}\Big)^{\frac{r+1}{r+2}} - 1 - \frac{\lambda_{\max}(\varepsilon\mathbb{E}M_{n,r})(r+1)}{n-r}\Big)\Big],\end{aligned} \tag{48}$$

and we first show that A_n can be written as a function of the term $\lambda_n := \lambda_{\max}(\varepsilon \mathbb{E} M_{n,r}))$. From the definition of λ_n and thanks to (31), it follows that

$$\left(1 + \frac{\varepsilon \mathbb{E} M_{n,r}(n-r)}{(r+1)v_{n,r}^{+}}\right)^{-\frac{1}{r+2}} = 1 - \frac{\lambda_n}{n-r},$$

thus one has

$$\left(1 + \frac{\varepsilon \mathbb{E} M_{n,r}(n-r)}{(r+1)v_{n,r}^{+}}\right)^{\frac{r+1}{r+2}} - 1 - \frac{(r+1)\lambda_n}{n-r} = \left(1 - \frac{\lambda_n}{n-r}\right)^{-(r+1)} - 1 - \frac{(r+1)\lambda_n}{n-r}; \quad (49)$$

analogously, exploiting (31) again, it is immediate to see that

$$\frac{\varepsilon \mathbb{E} M_{n,r}\lambda_n}{v_{n,r}^{+}} = \frac{(r+1)\lambda_n}{n-r}\left(\left(1 - \frac{\lambda_n}{n-r}\right)^{-(r+2)} - 1\right). \quad (50)$$

Thanks to Theorem 11, we know that, as $n \to +\infty$, the following asymptotic relations hold true: $v_{n,r}^{+} \simeq C_1 n^{\alpha}\ell(n)$ and $(n-r)\mathbb{E} M_{n,r} \simeq C_2 n^{\alpha}\ell(n)$ for suitable positive constants $C_1, C_2 > 0$. As a consequence of these relations and by (31), it is easy to show that $\lambda_n = \lambda_{\max}(\varepsilon \mathbb{E} M_{n,r})) \simeq C_3(n-r)$, with $C_3 \in (0,1)$. Since A_n can be expressed as the sum of the two terms in (49) and (50), which depend only on $\lambda_n \simeq C_3(n-r)$, it follows that, as $n \to +\infty$, A_n converges to

$$\begin{aligned} A_\infty &= (r+1)C_3[(1-C_3)^{-(r+2)} - 1] - (1-C_3)^{-(r+1)} + 1 + (r+1)C_3 \\ &= (r+1)C_3(1-C_3)^{-(r+2)} - (1-C_3)^{-(r+1)} + 1 \\ &= \left((r+2)C_3 - 1 + (1-C_3)^{r+2}\right)(1-C_3)^{-(r+2)}, \end{aligned}$$

and Bernoulli's inequality gives that $A_\infty > 0$. Therefore we conclude that

$$\exp(-v_{n,r}^{+}A_n) \lesssim \exp(-\bar{C}_2 n^{\alpha}\ell(n)), \quad (51)$$

with $\bar{C}_2 > 0$. The thesis now follows as a consequence of (47) and (51). ■

References

Anevski, D., Gill, R.D. and Zohren, S. (2017). Estimating a probability mass function with unknown labels. *Annals of Statistics* **45**, 2708–2735.

Auton, A. et al. (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.

Ayed, F., Battiston, M., Camerlenghi, F. and Favaro, S. (2018). On consistent estimation of the missing mass. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, to appear. *Preprint arXiv:1806.09712*.

Ayed, F., Battiston, M., Camerlenghi, F. and Favaro, S. (2019). A Good–Turing estimator for feature allocation models. *Electronic Journal of Statistics* **13**, 3775–3804.

Ben-Hamou, A., Boucheron, S. and Ohannessian, M.I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **23**, 249–287.

Ben-Hamou, A., Boucheron, S. and Gassiat, E. (2018). Pattern coding meets censoring: (almost) adaptive coding on countable alphabets. *Preprint: arXiv:1608.08367*

Beirlant, J. and Devroye, L. (1999) On the impossibility of estimating densities in the extreme tail. *Statistics and Probability Letters* **43**, 57–64

Bomba, L., Walter, K. and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology* **18**, 77.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities.* Oxford University Press.

Bubeck, S., Ernst, D., and Garivier, A. (2013). Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality. *Journal of Machine Learning Research* **14**, 601–623.

Cai, D., Mitzenmacher, M. and Adams, R.P. (2018). A Bayesian nonparametric view on count–min sketch. In *Advances in Neural Information Processing Systems.*

Cereda, G. and Gill, D.R. (2020). A nonparametric Bayesian approach to the rare type match problem. *Preprint: arXiv:1908.02954.*

Chao, A. and Colwell, R.K. (2017). Thirty years of progeny from Chao's inequality: estimating and comparing richness with incidence data and incomplete sampling. *Statistics and Operation Research Transactions* **41**, 3–54.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. and Ellison, A.M. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* **84**, 45–67.

Chatterjee, S. and Diaconis, P. (2018) The sample size required in importance sampling. *Annals of Applied Probability* **28**, 1099–1135

Cirulli, E.T. and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**, 415.

Colwell, R., Chao, A., Gotelli, N.J., Lin, S., Mao, C.X., Chazdon, R.L. and Longino, J.T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* **5**, 3–21.

Cormode, G. and Muthukrishnan, S. (2005). An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms* **55**, 58–75.

Daley, T. and Smith, A.D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods* **10**, 325–327.

Daley, D.J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes*, Volume I and Volume II. Springer, New York.

Efron, B. and Morris, C (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Association* **68**, 117–130.

Feller, W. (1971). *An introduction to probability theory and its applications*, Volume II. Wiley, NY.

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

Gale, W.A. and Sampson, G. (1995). Good–Turing frequency estimation without tears. *Journal of Quantitative Linguistics* **2**, 217–237.

Gao, F. (2013). Moderate deviations for a nonparametric estimator of sample coverage. *Annals of Statistics* **41**, 641-669.

Gao, Z., Tseng, C.H., Pei, Z. an Blaser, M.J. (2007). Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences of USA* **104**, 2927–2932.

Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys* **4**, 146–171.

Goldwater, S., Griffiths, T., and Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*.

Good, I.J.(1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237-264.

Grabchak, M. and Zhang, Z. (2017). Asymptotic properties of Turing's formula in relative error. *Machine Learning* **106**, 1771–1785.

Gravel, S. Henn, B.M., Gutenkunst, R.N. Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D. and Altshuler, D.L. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of USA* **108**, 11983–11988.

Gravel, S. (2014). Predicting discovery rates of genomic features. *Genetics* **197**, 601–610.

Harrison, B.A. (2010). *Move prediction in the game of go.* PhD thesis, Citeseer, 2010.

Ionita-Laza, I., Lange, C. and Laird, N.M. (2009). Estimating the number of unseen variants in the human genome. *Proceeding of the National Academy of Sciences of USA* **106**, 5008–5013.

Ionita-Laza, I., Lange, C. and Laird, N.M. (2010). On the optimal design of genetic variant discovery studies. *Statistical Applications in Genetics and Molecular Biology* **9**.

James, L.F. (2017). Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors. *Annals of Statistics* **45**, 2016–2045.

Karamata, J. (1930). Sur un mode de croissance régulière des fonctions. *Mathematica* **4**, 38–53.

Karamata, J. (1933). Sur un mode de croissance régulière. Théorèmes fondamentaux. *Bull. Soc. Math. France*, **61**, 55–62.

Karczewski, K. J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P. et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *Preprint BioRxiv:531210.*

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.*, **17**, 373–401.

Kingman, J.F.C.(1975). Random discrete distributions. *Journal of the Royal Statistical Society Series B* **37**, 1–15.

Kingman, J.F.C. (1993). *Poisson processes.* Oxford University Press, Oxford.

Kroes, I., Lepp, P.W. and Relman, D.A. (1999). Bacterial diversity within the human subgingival crevice. *Proceeding of the National Academy of Sciences of USA* **96**, 14547–14552.

McAllester, D. and Ortiz, L. (2003). Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research* **4**, 895–911.

McAllester, D. and Schapire, R.E. (2000). On the convergence rate of Good–Turing estimators. *Proceedings of the Conference on Computational Learning Theory.*

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* **1**, 226–251.

Mossel, E. and Ohannessian, M.I. (2019). On the impossibility of learning the missing mass. *Entropy* **21**, 28.

Navarro, D.J. and Griffiths, T.L. (2010). A nonparametric Bayesian model for inferring features from similarity judgments. In *Advances in Neural Information Processing Systems.*

Newman, M.E.J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, **45**, 167–256.

Ohannessian, M.I. and Dahleh, M.A. (2012). Rare probability estimation under regularly varying heavy tails. In *Proceedings of the 25th Annual Conference on Learning Theory*, **23**, 1–24.

Orlitsky, A., Santhanam, N.P. and Zhang, J. (2003). Always Good–Turing: asymptotically optimal probability estimation. *Science* **302**, 427-431.

Orlitsky, A., Santhanam, N.P. and Zhang, J. (2004). Universal compression of memoryless sources over unknown alphabets. *IEEE Transaction on Information Theory*, **50**, 1469–1481.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Schechter, S., Herley, C. and Mitzenmacher, M. (2010). Popularity is everything: A new approach to protecting passwords from statistical-guessing attacks. In *Proceedings of the USENIX Conference on Hot Topics in Security*.

Song, H.H., Cho, T.W., Dave, V., Zhang, Y. and Qiu, L. (2009). Scalable proximity estimation and link prediction in online social networks. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*.

Zhang, Q., Pell, J., Canino-Koning, R., Howe, A.C. and Brown. C.T. (2014). These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *PLOS ONE* **9**, e101271.

Zhang, C.H. and Zhang, Z. (2009). Asymptotic normality of a nonparametric estimator of sample coverage. *Annals of Statistics* **37**, 2582-2595.

Zou, J., Valiant, G., Valiant, P., Karczewski, K., Chan, S.O., Samocha, K., Lek, M., Sunyaev, S., Daly, M. and MacArthur, D.G. (2016). Quantifying the unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications* **7**.