A Good-Turing estimator for feature allocation models

Fadhel Ayed[1]**, Marco Battiston**[2]**,**
Federico Camerlenghi[3,*] **and Stefano Favaro**[4,†]

[1]*Department of Statistics,*
University of Oxford,
24-29 St Giles', OX1 3LB,
Oxford, United Kingdom
e-mail: fadhel.ayed@gmail.com

[2]*Department of Mathematics and Statistics,*
Lancaster University,
Fylde Ave, Bailrigg, LA1 4YR,
Lancaster, United Kingdom
e-mail: marco.battiston@stats.ox.ac.uk

[3]*Department of Economics, Management and Statistics,*
University of Milano - Bicocca,
Piazza dell'Ateneo Nuovo 1, 20126,
Milano, Italy
e-mail: federico.camerlenghi@unimib.it

[4]*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

Abstract: Feature allocation models generalize classical species sampling models by allowing every observation to belong to more than one species, now called features. Under the popular Bernoulli product model for feature allocation, we assume n observable samples and we consider the problem of estimating the expected number M_n of hitherto unseen features that would be observed if one additional individual was sampled. The interest in estimating M_n is motivated by numerous applied problems where the sampling procedure is expensive, in terms of time and/or financial resources allocated, and further samples can be only motivated by the possibility of recording new unobserved features. We consider a nonparametric estimator $\hat{M}_n$ of M_n which has the same analytic form of the popular Good-Turing estimator of the missing mass in the context of species sampling models. We show that $\hat{M}_n$ admits a natural interpretation both as a jackknife estimator and as a nonparametric empirical Bayes estimator. Furthermore, we give provable guarantees for the performance of $\hat{M}_n$ in terms of minimax rate optimality, and we provide with an interesting connection between $\hat{M}_n$ and the Good-Turing estimator for species sampling. Finally, we derive non-asymptotic confidence intervals for $\hat{M}_n$, which are easily computable and do not rely on any asymptotic approximation. Our approach is illustrated with synthetic data and SNP data from the ENCODE sequencing genome project.

---

*Also affiliated to Collegio Carlo Alberto, Torino and BIDSA, Bocconi University, Milano.
†Also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

1. Introduction

Feature allocation models generalize classical species sampling models by allowing every observation to belong to more than one species, now called features. In particular, every observation is endowed with a (unknown) finite set of features selected from a (possibly infinite) collection of features $(F_j)_{j\geq1}$. We conveniently represent each observation as a binary sequence, with entries indicating the presence (1) or absence (0) of each feature. Every feature F_j is associated to an unknown probability p_j, and each observation displays feature F_j with probability p_j, for $j \geq 1$. The Bernoulli product model, or binary independence model, is arguably the most popular feature allocation model. It models the i-th observation as a sequence $Y_i = (Y_{i,j})_{j\geq1}$ of independent Bernoulli random variables with unknown success probabilities $(p_j)_{j\geq1}$, with the assumption that Y_r is independent of Y_s for any $r \neq s$. Bernoulli product models have found applications in several scientific disciplines. They have been applied extensively in ecology when animals are captured using traps and each observation is an incidence vector collecting the presence or absence of each species in the traps (e.g., Colwell et al. [6], Chao et al. [5] and Chao and Colwell [4]). Besides ecology, Bernoulli product models found applications in the broad area of biosciences (e.g., Ionita-Laza et al. [15], Gravel [13] and Zou et al. [28]); in the analysis of choice behaviour arising from psychology and marketing (Görür et al. [11]); in binary matrix factorization for dyadic data (Meeds et al. [18]); in graphical models (e.g., Wood et al. [27] and Wood and Griffiths [26]); in the analysis of similarity judgement matrices (Navarro and Griffiths [20]); in network data analysis (Miller et al. [19]).

Let $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ denote a collection of n random samples collected under the Bernoulli product model with unknown feature probabilities $(p_j)_{j\geq1}$. Furthermore, let $X_{n,j} = \sum_{1\leq i\leq n} Y_{i,j}$ be the number of times that feature F_j has been observed in $\mathbf{Y}_n$. That is, $X_{n,j}$ is a Binomial random variable with parameters (n, p_j) for any $j \geq 1$. In this paper we consider the problem of estimating the conditional expected number, given the random sample $\mathbf{Y}_n$, of hitherto unseen features that would be observed if one additional sample Y_{n+1} was collected, i.e.

$$M_n(\mathbf{Y}_n, (p_j)_{j\geq1}) = \mathbb{E}\left(\sum_{j\geq1}\mathbb{I}\{X_{n,j} = 0, Y_{n+1,j} = 1\} \,|\, \mathbf{Y}_n\right) \\ = \sum_{j\geq1} p_j\mathbb{I}\{X_{n,j} = 0\}, \tag{1.1}$$

where $\mathbb{I}$ is the indicator function. For easiness of notation, in the rest of the paper we will not highlight the dependence on $\mathbf{Y}_n$ and $(p_j)_{j \geq 1}$, and simply write M_n instead of $M_n(\mathbf{Y}_n, (p_j)_{j \geq 1})$. The statistic M_n is referred to as the missing mass, namely the sum of the probability masses of unobserved features in the first n samples.

Interest in estimating the missing mass is motivated by numerous applied problems where the sampling procedure is expensive, in terms of time and/or financial resources, and further draws can be only motivated by the possibility of recording new unobserved 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. [1]). An accurate estimate of M_n will enable better study design and quantitative evaluation of the potential and limitations of these datasets. Indeed one can fix a suitable threshold such that the sequencing procedure takes place until the estimate of M_n becomes for the first time smaller than the threshold. This introduces a criterion for evaluating the effectiveness of further sampling, providing a roadmap for large-scale sequencing projects. A Bayesian parametric estimator of (1.1) has been introduced in Ionita-Laza et al. [15], and it relies on a Beta prior distribution for the unknown probabilities p_j's. A limitation of this approach is that it requires parametric forms for the distribution of variant frequencies, which requires some model of demography and selection. For instance, the Beta prior distribution is a reasonable assumption for neutrally evolving variants but may not be appropriate for deleterious mutations. To overcome this drawback, a nonparametric approach to estimate M_n has been proposed in the recent work of Zou et al. [28]. This is a purely algorithmic type approach, whose core is a linear programming based algorithm.

In this paper, we consider an estimator $\hat{M}_n$ of M_n which has the same analytic form of the popular Good-Turing estimator of the missing mass under the multinomial model for species sampling (e.g., Good [10] and Robbins [23]), with the difference that the two estimators have different ranges (supports). We refer to $\hat{M}_n$ as the Good-Turing estimator for feature allocation models. As in Zou et al. [28], our approach is nonparametric in the sense that it does not rely on any distributional assumption on the unknown p_j's. The use of the Good-Turing estimator for estimating M_n was first considered in Chao and Colwell [4], where also a bootstrap procedure is presented to approximate the variance of the estimator. Our work delves into the Good-Turing estimator for feature allocation models, thus providing theoretical guarantees for its use. We show that $\hat{M}_n$ admits a natural interpretation both as a jackknife (resampling) estimator (Quenouille [21] and Tukey [25]) and as a nonparametric empirical Bayes estimator in the sense of Efron and Morris [9]. Theoretical properties of the estimator $\hat{M}_n$ are investigated. Specifically, we first provide a lower bound for the minimax risk under a squared loss function and then we show that the mean squared error of the proposed estimator achieves the optimal minimax rate for the estimation of M_n. This asymptotic analysis provides with an interesting connection between the Good-Turing estimators for species sampling models

and for feature allocation models in terms of the limiting Poisson regime of the Binomial distribution. Finally, we derive a non-asymptotic confidence interval for $\hat{M}_n$ by means of novel Bernstein type concentration inequalities which are of separate interest; the confidence interval is easy to compute and it does not rely on any asymptotic assumption or any parametric constraint on the unknown p_j's. We illustrate the proposed methodology by the analysis of various synthetic data and SNP datasets from the ENCODE (http://www.hapmap.org/downloads/encode1.html.en) sequencing genome project.

The paper is structured as follows. In Section 2 we introduce the Good-Turing estimator $\hat{M}_n$ for the missing mass M_n, we present its interpretation as a jackknife estimator and a nonparametric empirical Bayes estimator, and we give provable guarantees for its performance. In Section 3 we derive a non-asymptotic confidence interval for $\hat{M}_n$, in Section 4 we discuss the use of $\hat{M}_n$ for designing cost-effective feature inventories, and in Section 5 we present a numerical illustration of $\hat{M}_n$ with synthetic and real data. Some concluding remarks on future works are discussed in Section 6. Proofs are deferred to Appendix A.

2. A Good-Turing estimator for M_n

Let $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ be a collection of n random samples under the Bernoulli product model with unknown feature probabilities $(p_j)_{j \geq 1}$ such that $\sum_{j \geq 1} p_j < +\infty$. That is, $Y_i = (Y_{i,j})_{j \geq 1}$ is a sequence of independent Bernoulli random variables, with p_j being the success probability of $Y_{i,j}$ for any $i = 1, \ldots, n$, and Y_r is independent of Y_s for any $r \neq s$. Note that the assumption $\sum_{j \geq 1} p_j < +\infty$ implies that each Y_i displays finitely many features almost surely; indeed, by monotone convergence, we have $\sum_{j \geq 1} p_j < +\infty$ if and only if $\mathbb{E}(\sum_{j \geq 1} Y_{i,j}) < +\infty$, which in turns implies that $\sum_{j \geq 1} Y_{i,j} < +\infty$ almost surely. If $X_{n,j}$ denotes the number of times that feature F_j has been observed in the random sample $\mathbf{Y}_n$, then

$$K_{n,r} = \sum_{j \geq 1} \mathbb{I}\{X_{n,j} = r\}$$

is the number of features with frequency r in $\mathbf{Y}_n$. Let K_n be the total number of features in $\mathbf{Y}_n$, i.e. $K_n = \sum_{r \geq 1} K_{n,r}$. For any two sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$, write $a_n \simeq b_n$ to mean that $a_n/b_n \to 1$ as $n \to +\infty$. An intuitive estimator of M_n can be deduced from a comparison between expectations of M_n and $K_{n,1}$. Specifically, since $X_{n,j}$ is a Binomial random variable with parameter (n, p_j),

$$\mathbb{E}(M_n) = \sum_{j \geq 1} p_j \mathbb{P}(X_{n,j} = 0) = \sum_{j \geq 1} p_j (1 - p_j)^n$$

$$= \frac{1}{\binom{n+1}{1}} \sum_{j \geq 1} \binom{n+1}{1} p_j (1 - p_j)^{(n+1)-1} = \frac{1}{n+1} \mathbb{E}(K_{n+1,1}) \simeq \frac{1}{n} \mathbb{E}(K_{n,1})$$

and set

$$\hat{M}_n = \frac{K_{n,1}}{n}$$

as an estimator of M_n, for any $n \geq 1$. The estimator $\hat{M}_n$ is nonparametric, in the sense that it does not rely on any distributional assumptions on the feature probabilities $(p_j)_{j \geq 1}$. See Chao and Colwell [4] for a somehow related derivation of $\hat{M}_n$.

The estimator $\hat{M}_n$ turns out to have the same analytic form of the popular Good-Turing estimator of the missing mass for species sampling models. The Good-Turing estimator first appeared in Good [10] as a nonparametric empirical Bayes estimator under the classical multinomial model for species sampling, i.e., $(Y_1, \ldots, Y_n)$ are n random samples from a population of individuals belonging to a (possibly infinite) collection of species $(S_j)_{j \geq 1}$ with unknown proportions $(p_j)_{j \geq 1}$ such that $\sum_{j \geq 1} p_j = 1$. Under the multinomial model every observation is endowed with one species selected from $(S_j)_{j \geq 1}$, and hence $K_{n,1} \in \{0, 1, \ldots, n\}$. Therefore, although the estimator $\hat{M}_n$ has the same analytic form of the Good-Turing estimator, the two estimators have different ranges (supports): while the Good-Turing estimator for species sampling models takes values in $[0, 1]$, the Good-Turing estimator for feature allocation models takes positive (finite) values.

Hereafter we show that: i) the estimator $\hat{M}_n$ admits natural interpretations both as a jackknife (resampling) estimator and as a nonparametric empirical Bayes estimator in the sense of Efron and Morris [9]; ii) the mean squared error of $\hat{M}_n$ converges to zero at the best possible rate for the estimation of the missing mass M_n; iii) the estimator $\hat{M}_n$ is linked to the Good-Turing estimator for species sampling models in terms of the limiting Poisson regime of the Binomial distribution.

2.1. Interpretations of $\hat{M}_n$

We first show that the estimator $\hat{M}_n$ admits a natural interpretation as a jackknife estimator in the sense of Quenouille [21] and Tukey [25]. See also the monograph by Efron [8] and references therein for a comprehensive account on jackknife (resampling) estimators. Let $K_n(j)$ denote the number of distinct features observed in the sample $\mathbf{Y}_n$ after the removal of the j-th sample Y_j. It is easy to show that the missing mass M_{n-1} equals the posterior expected value of $K_n - K_{n-1}(j)$, given all the samples but Y_j, i.e., $M_{n-1} = \mathbb{E}[K_n - K_{n-1}(j)| \{Y_1, \ldots, Y_n\} \setminus Y_j]$. Accordingly, the difference $K_n - K_{n-1}(j)$ provides with an unbiased estimator of M_{n-1}. The jackknife estimator is then $\sum_{1 \leq j \leq n} (K_n - K_{n-1}(j))/n$ which equals $\hat{M}_{n-1} = K_{n,1}/n$; indeed, for any fixed j, $K_n - K_{n-1}(j)$ coincides with the number of features displayed only by the j-th sample Y_j.

The estimator $\hat{M}_n$ also admits a natural interpretation as a nonparametric empirical Bayes estimator in the sense of Efron and Morris [9]. A Bayesian nonparametric approach to estimate M_n relies on the specification of a prior distribution for the unknown feature probabilities $(p_j)_{j \geq 1}$. Here we consider the three parameters Beta process prior (e.g., Teh and Görür [24] and James [16]). This is a well-known generalization of the celebrated Beta process prior of Hjort

[14], and it is defined as the distribution of a completely random measure (Daley and Vere-Jones [7]) with intensity measure $\nu(\mathrm{d}s, \mathrm{d}f) = \rho(s)\mathrm{d}s\mathbb{I}(0,1)(f)\mathrm{d}f$ where $\rho(s) = \vartheta s^{-1-\alpha}(1-s)^{\beta+\alpha-1}/\Gamma(1-\alpha)$, for $\alpha \in (0,1)$, $\beta > -\alpha$ and $\vartheta > 0$, with $\Gamma(\cdot)$ being the Gamma function. In particular, under the three parameters Beta process prior, the Bayesian point estimator of M_n under squared loss function is

$$\tilde{M}_n(\vartheta, \alpha, \beta) = \frac{\vartheta\Gamma(\alpha + \beta + n)}{\Gamma(\beta + n + 1)}. \tag{2.1}$$

We refer to the Appendix A.1 for further details on the three parameters Beta process prior, and on the derivation of (2.1). In particular, it is shown in Appendix A.1 that

$$\hat{\vartheta}_n = \frac{K_{n,1}\Gamma(\beta + n + 1)}{\Gamma(\beta + \alpha + n)n}$$

is a consistent estimator of ϑ for any $\alpha \in (0,1)$ and $\beta > -\alpha$, and plugging $\hat{\vartheta}_n$ into (2.1) we obtain

$$\tilde{M}_n(\hat{\vartheta}_n, \alpha, \beta) = \frac{K_{n,1}}{n} = \hat{M}_n.$$

In other terms, the proposed estimator $\hat{M}_n$ coincides with the nonparametric empirical Bayes estimator $\tilde{M}_n(\hat{\vartheta}_n, \alpha, \beta)$ of the missing mass. $\tilde{M}_n(\hat{\vartheta}_n, \alpha, \beta)$ is obtained by assigning the three parameters Beta process prior to $(p_j)_{j\geq 1}$ and then estimating its mass parameter ϑ with an appropriately chosen consistent estimator.

2.2. Optimality of $\hat{M}_n$

We start by showing that the mean squared error (L_2-risk) of the estimator $\hat{M}_n$ goes to zero at the best possible rate. This result legitimates the use of $\hat{M}_n$ as an estimator of the missing mass M_n. For any $0 < W < +\infty$ let $\mathscr{P}_W = \{(p_j)_{j\geq 1} : p_j \in (0,1) \text{ and } \sum_{j\geq 1} p_j \leq W\}$ be a class of features probabilities. Also, let $\hat{T}_n$ denote a whichever estimator of the missing mass M_n based on the random sample $\mathbf{Y}_n$, that is a map on the n-fold product space of observations $\{0,1\}^\infty \times \cdots \times \{0,1\}^\infty$ and taking values in $\mathbb{R}^+$. For a specific collection of feature probabilities $(p_j)_{j\geq 1}$ in the class $\mathscr{P}_W$, the L_2-risk of the estimator $\hat{T}_n$ is defined as follows

$$R_n(\hat{T}_n; (p_j)_{j\geq 1}) = \mathbb{E}[(M_n - \hat{T}_n)^2] = \mathrm{Bias}(\hat{T}_n)^2 + \mathrm{Var}(\hat{T}_n - M_n),$$

and the minimax risk is

$$R_n^* = \inf_{\hat{T}_n} \sup_{(p_j)_{j\geq 1} \in \mathscr{P}_W} R_n(\hat{T}_n; (p_j)_{j\geq 1}), \tag{2.2}$$

i.e. the infimum, with respect to the set of possible estimators $\hat{T}_n$ of the missing mass M_n, of the worst-case risk over the class of feature probabilities $\mathscr{P}_W$. The next theorem provides with an upper bound for $R_n(\hat{M}_n; (p_j)_{j\geq 1})$ and a lower bound for R_n^*.

Theorem 2.1. *Let $(p_j)_{j\geq 1}$ be feature probabilities in the class $\mathscr{P}_W$. Then,*

i) for any $n \geq 1$

$$\sup_{(p_j)_{j\geq 1}\in\mathscr{P}_W} R_n(\hat{M}_n; (p_j)_{j\geq 1}) \leq \frac{1}{n^2}W^2 + \frac{2n+1}{n(n+1)}W; \qquad (2.3)$$

ii) for any $n \geq 2$

$$R_n^* \geq \frac{2W}{9(3n+1)} - \frac{14}{n^2}. \qquad (2.4)$$

See Appendix A.2 for the proof of Theorem 2.1. For any two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, write $a_n \lesssim b_n$ to mean that there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for all n and $a_n \asymp b_n$ to mean that we have both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Part i) of Theorem 2.1 shows that $\sup_{(p_j)_{j\geq 1}\in\mathscr{P}_W} R_n(\hat{M}_n; (p_j)_{j\geq 1}) \lesssim W/n$, namely as n goes to infinity the mean squared error of $\hat{M}_n$ goes to zero at rate W/n. Part ii) of Theorem 2.1 provides with a lower bound for the minimax risk, showing that $R^* \gtrsim W/n$. In other terms, this shows that the mean squared error of $\hat{M}_n$ goes to zero at the best possible rate. Theorem 2.1 then shows that the estimator $\hat{M}_n$ is essentially rate optimal due to the matching minimax lower bound in the class $\mathscr{P}_W$ of admissible probabilities' masses up to a constant factor.

In Theorem 2.1 we considered the class $\mathscr{P}_W$ of feature probabilities having total mass bounded by W. We highlight the crucial role or this assumption in order to provide asymptotic results. First, notice from Theorem 2.1 that, for a fixed value of the sample size n, the minimax rate increases linearly in W. This implies that, for a fixed n, the estimation problem can be made as hard as desired if no bounds are imposed on the sum of the possible vectors of probabilities $(p_j)_{j\geq 1}$. Since at first glance this result can seem slightly counter intuitive, let us consider an illustrative example which should clarify why the estimation difficulty of the problem is proportional to W. To simplify the argument, suppose that W is a strictly positive integer and consider W frequency vectors $(p_j^{(0)})_{j\geq 1}, \ldots, (p_j^{(W-1)})_{j\geq 1} \in \mathscr{P}_1$, i.e. such that $\sum_j p_j^{(r)} \leq 1$ for all $r \in \{0, \ldots, W-1\}$. Now, let us construct another vector of frequencies $P = (p_j)_{j\geq 1} \in \mathscr{P}_W$ obtained by setting $p_{qW+r} = p_q^{(r)}$ for $q \geq 0$ and $0 \leq r \leq W-1$. The resulting vector P is the concatenation of $(p_j^{(0)})_{j\geq 1}, \ldots, (p_j^{(W-1)})_{j\geq 1}$, such that in the first W entries of P we put the first elements $(p_1^{(0)}, \ldots, p_1^{(W-1)})$, followed by the second ones $(p_2^{(0)}, \ldots, p_2^{(W-1)})$, the third ones $(p_3^{(0)}, \ldots, p_3^{(W-1)})$ and so on. Furthermore, any observation $(Y_{i,j})_{j\geq 1}$ sampled from the Bernoulli product model with frequencies given by this P can be rewritten, following a similar construction, as the concatenation of W observations $Y_i^{(r)} = (Y_{i,j}^{(r)})_{j\geq 1}$, $r \in \{0, \ldots, W-1\}$, each of them sampled from a binomial product model with the corresponding frequencies $(p_j^{(r)})_{j\geq 1}$. Hence, the missing mass for a random sample from P can be related to the missing masses of each of its W subcom-

ponents, by

$$M_n(\mathbf{Y}_n; P) = \sum_{r=0}^{W-1} M_n^{(r)}(\mathbf{Y}_n^{(r)}; (p_j^{(r)})_{j \geq 1}),$$

where $\mathbf{Y}^{(r)} = (Y_1^{(r)}, \ldots, Y_n^{(r)})$. From this construction, we can see that by trying to estimate the missing mass on the left hand side, we are basically trying to estimate a sum of W unrelated quantities, which explains why the error is linear in W.

In order to apply Theorem 2.1 and evaluate the performances of the Good-Turing estimator compared to the minimax rate, we would need to know an upper bound W, i.e. an upper bound on the total mass of the unknown vector of probabilities $(p_j)_{j \geq 1}$. In real life applications, W is unlikely to be known. However, we can easily estimate it. Specifically, since the total mass W is the expected number of features displayed per observation, we can use as a consistent estimator for W the estimator $\hat{W}_n := \sum_{j \geq 1} X_{n,j}/n$. Besides the following concentration inequality for $\hat{W}_n$ around its mean W may be useful to measure the probability that W_n deviates from W. In particular it can be proved that for any $\delta \in (0, 1)$

$$\mathbb{P}\left(\hat{W}_n - W > \sqrt{\frac{4W}{n}\log(1/\delta)} + \frac{1}{n}\log(1/\delta)\right) \leq \delta \tag{2.5}$$

and

$$\mathbb{P}\left(-(\hat{W}_n - W) > \sqrt{\frac{2W}{n}\log(1/\delta)}\right) \leq \delta. \tag{2.6}$$

The proofs of these exponential tail bounds rely on suitable bounds on the corresponding log-Laplace transform of $\hat{W}_n - \mathbb{E}\hat{W}_n$. See Section A.3 for details. Equation (2.6) leads to an upper bound on W, indeed with probability $1 - \delta$ one has

$$W \leq \left(\sqrt{\hat{W}_n + \frac{\log(1/\delta)}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}\right)^2. \tag{2.7}$$

Analogously, one may directly apply Equation (2.5) in order to find a lower bound for W. More precisely, it is straightforward to see the validity of the following inequality

$$\sqrt{W} \geq \sqrt{\hat{W}_n} - \sqrt{\frac{\log(1/\delta)}{n}} \tag{2.8}$$

with probability bigger than $1 - \delta$. Therefore (2.7)–(2.8) result in a confidence intervals for W.

Finally, we conclude this subsection by remarking that the proposed estimator displays some bias. The formula of the theoretical bias can be found in the proof of Theorem 2.1. However, the effect of this bias becomes negligible for any reasonable sample size. As a simple illustration, in Table 1, we display the theoretical bias of the estimator for different sample sizes, when the frequencies

TABLE 1

Bias of $\hat{M}_n$: $p_j = 1/j^s$ for $j \leq 10^5$ and $p_j = 0$ for $j > 10^5$, and different values of s.

s	$n = 10$	$n = 50$	$n = 100$	$n = 1000$
0.6	1.310	0.459	0.241	0.001
0.8	0.268	0.075	0.042	0.003
1.0	0.100	0.020	0.010	0.001
1.2	0.052	0.008	0.004	0.000
1.4	0.033	0.004	0.002	0.000
1.6	0.023	0.002	0.001	0.000

TABLE 2

$Bias(\hat{M}_n)^2/R_n$ *100: $p_j = 1/j^s$ for $j \leq 10^5$ and $p_j = 0$ for $j > 10^5$, and different values of s.

s	$n = 10$	$n = 50$	$n = 100$	$n = 1000$
0.6	16.06	13.05	9.57	0.89
0.8	2.99	1.44	1.08	0.19
1.0	1.35	0.34	0.19	0.03
1.2	0.94	0.16	0.07	0.01
1.4	0.76	0.10	0.04	0.00
1.6	0.64	0.07	0.03	0.00

$(p_j)_{j \geq 1}$ are chosen according to the Zipf's Law for different values of the parameter s (the same simulation setting will be reconsidered and better described in Section 5). In Table 2, we report for each combination of sample size n and parameter s the percentage of the risk that is due to the bias. As can be seen from Table 1 and 2, expect for very small sample size and Zipf parameter, under which most of the frequencies are high, the bias is always almost negligible and does not contribute much to the risk of estimator, which is mainly due to its variance.

2.3. Connection to the Good-Turing estimator for species sampling models

We relate the Good-Turing estimator for feature allocation models with the classical Good-Turing estimator for species sampling. This link relies on the well-known limiting Poisson approximation of Binomial random variables. Theorem 2.1 states that, in the feature allocation models, the Good-Turing estimator achieves a risk of order $R_n \asymp \frac{W}{n}$, while it is known from Rajaraman et al. [22] that the risk $\tilde{R}_n$ of the Good-Turing estimator in the species sampling case asymptotically behaves as $\tilde{R}_n \asymp 1/n$. In order to compare the two models, we will consider the limiting scenario when $W \to 0$. Let us consider a vector of feature frequencies $(p_j)_{j \geq 1}$, such that $\sum_j p_j = W = \frac{\lambda}{n}$ for some positive value λ and denote $\tilde{p}_j = \frac{p_j}{\sum_{j \geq 1} p_j}$ the normalized probability vector. Applying the large n Poisson approximation of the binomial distribution, it follows that each $X_{n,j}$ is now approximately distributed according to a Poisson distribution with mean $\lambda \tilde{p}_j$. Therefore, the approximated model for large n boils down to sample first an effective size $\tilde{n} = \sum_{j \geq 1} X_{n,j}$, distributed according to a Poisson with parameter λ, and, conditionally on it, sample $\tilde{n}$ independent and identically

distributed observations $(\tilde{Y}_1, \ldots, \tilde{Y}_{\tilde{n}})$ from probability vector $(\tilde{p}_j)_{j \geq 1}$. Hence it is equivalent to a species sampling model where the sample size $\tilde{n}$ is assumed to be random. Denote $\tilde{M}_{\tilde{n}}$ the missing mass in the corresponding species sampling model. Now, by noticing that $M_n = \frac{\lambda}{n} \tilde{M}_{\tilde{n}}$, it follows that $R_n \approx \frac{\lambda^2}{n^2} \mathbb{E}_{\tilde{n}}(\tilde{R}_{\tilde{n}})$. Therefore, we expect the quantity $\frac{n^2}{\lambda^2} R_n$ to have the same asymptotic behaviour of $\mathbb{E}_{\tilde{n}}\left(\frac{1}{\tilde{n}} \mathbb{I}\{\tilde{n} > 0\}\right)$. Indeed, from Theorem 2.1 and $W = \frac{\lambda}{n}$, we get that $\frac{n^2}{\lambda^2} R_n \asymp \frac{1}{\lambda} = \frac{1}{\mathbb{E}(\tilde{n})}$.

We conclude by remarking that the above construction also provides a justification for the Good-Turing estimator in the feature allocation models. Indeed, in the context of feature allocation models, we want to estimate $M_n = W \tilde{M}_{\tilde{n}}$, where both W and $\tilde{M}_{\tilde{n}}$ are unknown and need to be estimated. In order to estimate $\tilde{M}_{\tilde{n}}$, we can use the Good-Turing estimator for species models, which here turns to be

$$\tilde{M}_{\tilde{n}}^{\mathrm{GT}} = \frac{K_{n,1}}{\tilde{n}} = \frac{K_{n,1}}{\sum_{j \geq 1} X_{n,j}}.$$

However, we also need to estimate $W = \sum_j p_j$ in order to make use of $\tilde{M}_{\tilde{n}}^{GT}$ as an estimator of M_n. As pointed out at the end of the previous subsection, a consistent estimator of W is $\hat{W} = \frac{\sum_j X_{n,j}}{n}$. Then, by combining the estimator $\hat{W}$ for W with the estimator $\tilde{M}_{\tilde{n}}^{\mathrm{GT}}$ for $\tilde{M}_{\tilde{n}}$, we obtain the following estimato of M_n,

$$\hat{M}_n = \hat{W} \tilde{M}_{\tilde{n}}^{\mathrm{GT}} = \frac{\sum_{j \geq 1} X_{n,j}}{n} \frac{K_{n,1}}{\sum_{j \geq 1} X_{n,j}} = \frac{K_{n,1}}{n},$$

which turns out to be exactly the Good-Turing estimator for feature allocation models.

3. A confidence interval for $\hat{M}_n$

In this section, we consider the problem of uncertainty quantification of the proposed estimator of the missing mass. In particular, we exploit tools from concentration inequalities for sum of independent random variables (Boucheron et al. [3]) to introduce a non-asymptotic level-δ confidence interval for $\hat{M}_n$, for any $\delta \in (0, 1)$.

Theorem 3.1. *Let $(p_j)_{j \geq 1}$ be any sequence of feature probabilities s.t. $\sum_{j \geq 1} p_j < +\infty$, and set $c_\delta(x) = (\sqrt{\log(1/\delta)/2} + \sqrt{7 \log(1/\delta)/6 + x})^2$ for any nonnegative integer x and $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$*

$$\hat{M}_n - L_\delta(K_{n,1}, K_{n,2}) \leq M_n \leq \hat{M}_n + U_\delta(K_n),$$

where

$$L_\delta(K_{n,1}, K_{n,2}) = \frac{2c_\delta(K_{n,2})}{n(n-1)} + \frac{\log(6/\delta)}{n} + \sqrt{2 \log\left(\frac{6}{\delta}\right)\left(\frac{4c_\delta(K_{n,1})}{n(n-1)} + \frac{2c_\delta(K_{n,2})}{n^2}\right)}$$

and

$$U_\delta(K_n) = \frac{\log(6/\delta)}{n-1} + \sqrt{2\log\left(\frac{6}{\delta}\right)\frac{4c_\delta(K_n)}{(n-1)^2(1-2/n)}}.$$

This theorem can be proven in two steps. We first derive bounds on the log-Laplace transform of $D_n = \hat{M}_n - M_n$, from which we can obtain a confidence interval for D_n as a function of $\mathbb{E}K_{n,1}$, $\mathbb{E}K_{n,2}$ and $\mathbb{E}K_n$. Then, by deriving bounds on the log-Laplace of $K_{n,1}$, $K_{n,2}$ and K_n, we are able to bound with high probability the deviation of these random variables from their mean. Theorem 3.1 then follows by plugging-in the results of step 2 in the bounds of step 1. The details are left in Appendix A.4. For any $\delta \in (0,1)$, Theorem 3.1 provides with a level $1-\delta$ confidence interval for $\hat{M}_n$ that holds for every sample size n and for every possible values of the feature probabilities $(p_j)_{j\geq 1}$. Indeed, this is derived by applying finite sample concentration inequalities, without using any asymptotic approximation, and it does not rely on any distributional assumption on the feature probabilities $(p_j)_{j\geq 1}$. Note that the proposed confidence interval can be easily evaluated without resorting to any simulation based strategy. It is enough to count the total number of features and number of features with frequency one and two in the observable sample, i.e. K_n, $K_{n,1}$ and $K_{n,2}$, and plug these quantities into $U_\delta(K_n)$ and $L_\delta(K_{n,1}, K_{n,2})$ to be added and subtracted to $\hat{M}_n$.

4. A stopping rule for the discovery process

As we recalled in the Introduction, interest in estimating missing mass M_n is motivated by the design of feature inventories that are cost-effective in terms of the number of features discovered and the amount of resources allocated in the discovery process. This is a fundamental issue in numerous scientific disciplines where the sampling procedure is expensive, in terms of time and/or financial resources. Feature inventories must then be designed in such a way to redirect the search of new features to more productive sites, methods or time periods whenever the sampling effort becomes unprofitable. In such a context, the estimator $\hat{M}_n$ is the key ingredient for defining an adaptive sequential approach in terms a stopping rule for the discovery process. Specifically, let h be an utility function, defined on the integers, such that $h(k)$ is the gain of observing k features; assume that h in non-decreasing and concave. If c is the cost associated with each sampling step, and K_n is the number of features in the sample $\mathbf{Y}_n$, then the stopping rule may be defined as

$$n^* = \inf\{n \geq 0 : [h(K_n + \hat{M}_n) - h(K_n)] \leq c\}.$$

This brief discussion highlights how to exploit the estimator $\hat{M}_n$ within the context of designing cost-effective feature inventories. In particular, it gives rises to the challenging problem of finding the time n^* at which it is optimal to conclude the discovery process, i.e. the time that maximizes the expected payoff $\mathbb{E}(h(K_n) - cn)$.

TABLE 3

Synthetic data: missing mass estimation with $p_j = 1/j^s$ for $j \leq 10^5$ and $p_j = 0$ for $j > 10^5$, and different values of s.

s	$n = 50$			$n = 250$		
	M_n	$\hat{M}_n$	95% CI	M_n	$\hat{M}_n$	95% CI
0.6	183.81	184.66	$(174.35, 198.12)$	105.38	105.61	$(101.74, 110.53)$
0.8	33.79	33.67	$(29.45, 39.51)$	26.40	26.05	$(24.40, 28.30)$
1.0	7.02	7.02	$(4.88, 9.91)$	5.41	5.43	$(4.66, 6.49)$
1.2	1.92	1.87	$(0.54, 3.60)$	1.35	1.35	$(0.92, 1.93)$
1.4	0.71	0.72	$(0, 1.98)$	0.44	0.44	$(0.15, 0.81)$
1.6	0.34	0.36	$(0, 1.39)$	0.18	0.18	$(0, 0.46)$

s	$n = 1000$		
	M_n	$\hat{M}_n$	95% CI
0.6	27.84	27.89	$(26.66, 29.65)$
0.8	17.18	17.17	$(16.47, 18.13)$
1.0	4.04	4.03	$(3.70, 4.48)$
1.2	0.97	0.97	$(0.80, 1.20)$
1.4	0.29	0.29	$(0.19, 0.43)$
1.6	0.12	0.11	$(0.04, 0.21)$

5. Numerical illustration

In this section we illustrate the experimental performance of the estimator $\hat{M}_n$ by the analysis of various synthetic data, and by the analysis SNP datasets from a genome project. In particular, let N denote the total number of possible features in the whole population of individuals. With regards to synthetic data, we present a numerical illustration by setting $p_j = 1/j^s$ for $j \leq N$ and $p_j = 0$ for $j > N$, with s being a nonnegative parameter. Note that the feature probability masses p_j's correspond to the unnormalized masses of the ubiquitous Zipf distribution. Recall that the parameter s controls how the total mass is spread among the features: the lager s, the smaller is the number of features with high probability. In other terms the larger s, the larger is the number of features with small frequencies (rare features). Hereafter we set $N = 10^5$, and we consider the following values for the discount parameter s: 0.6, 0.8, 1.0, 1.2, 1.4, 1.6. For each of these values of the parameter s, we take 100 samples of size $n = 50, 250, 1000$ from the population $(p_j)_{1 \leq j \leq N}$. Table 1 displays the true value of the missing mass M_n, the estimated value $\hat{M}_n$ and the corresponding 95% confidence interval (CI). All the experiments are averaged over the 100 samples. Results in Table 1 show that $\hat{M}_n$ provides good estimates of true value of M_n in all the scenarios that we considered. In addition, confidence intervals are quite narrow around the estimator and contain always the true value of the missing mass.

We conclude with an application to SNP data. In this context, each sample represents a genome sequence and we are interested in variations with respect to the reference genome. Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, which is called a nucleotide. For example, at a specific position in the human genome, the C (Cytosine) nucleotide may

TABLE 4
Real data: missing mass estimation for the ENCODE sequencing genome project.

Population	$n = 20$		$n = 40$	
	$\hat{M}_n$	95% CI	$\hat{M}_n$	95% CI
CEPH	55.6	$(38.7, 88.9)$	22.3	$(15.6, 38.9)$
CHB	50.0	$(37.3, 81.3)$	17.6	$(12.6, 32.8)$
JPT	61.7	$(46.9, 93.4)$	26.7	$(20.9, 42.2)$
YRI	88.3	$(65.2, 125.2)$	26.6	$(18.9, 44.7)$

appear in most individuals, but the position can be replaced by an A (Adenine) in a small group of people. This means that there is a SNP at that position of the DNA. SNPs are important biological markers, helping scientists locate genes that are associated with diseases. We use the SNP datasets from the ENCODE sequencing genome project (http://www.hapmap.org/downloads/encode1.html.en). The same project was analyzed in Ionita-Laza et al. [15]. Ten 500-kb regions of the genome were sequenced in 209 unrelated DNA samples: 60 Yoruba (YRI), 60 CEPH European (CEPH), 45 Han Chinese (CHB), and 44 Japanese (JPT). These regions were chosen to be representative of the genome in general, including various chromosomes, recombination rates, gene density, and values of nontranscribed conservation with mouse. To make results comparable across the 4 populations (YRI, CEPH, CHB, and JPT), we consider only $n = 20$ of the sequenced individuals for each dataset. Table 2 displays the estimated values and 95% confidence interval (CI). For samples of 20 individuals, the YRI population displays the highest estimate of the missing mass. This agrees with results in Ionita-Laza et al. [15], showing that the African population is the most diverse. We also consider increasing the sample size to $n = 40$ of the sequenced individuals for each dataset. Table 2 shows how the missing mass decreases with respect to the case $n = 20$. This highlights the saturation effect in discovering new variants. The discovery process is very efficient in the beginning, but after many individuals are sequenced, each additional individual contributes less and less to the pool of the newly discovered variants.

6. Concluding remarks

In this paper we delved into the Good-Turing estimator for feature allocation models, thus providing theoretical guarantees for its use. In particular, we proved that the mean squared error of $\hat{M}_n$ goes to zero at the best possible rate as the sample size n goes to infinity. Our results are simple, intuitive and easily implementable from practitioners. It distinguishes from the approaches of Ionita-Laza et al. [15] and Zou et al. [28] for being the first nonparametric statistical approach for estimating the missing mass in feature allocation models. In particular, differently from Ionita-Laza et al. [15] and Zou et al. [28], we associated to the proposed estimator an exact (non-asymptotic) quantification of its uncertainty. Results and techniques presente in this paper paves the way to new research directions in the context of feature allocation problems. In particular, three promising directions are: i) estimating the conditional expected number,

given an observable sample $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$, of features with frequency $r > 0$ in $\mathbf{Y}_n$ that would be observed in one additional sample (Chao and Colwell [4]); ii) solving the optimal stopping problem discussed in Section 4; iii) estimating the conditional expected number, given an observable sample $\mathbf{Y}_n$, of new features that would be observed in $m > 1$ additional (unobservable) samples (Colwell et al. [6]).

Appendix A: Proofs

A.1. Nonparametric empirical Bayes

In the present section we derive $\hat{M}_n$ as a nonparametric empirical Bayes estimator in the sense of Efron and Morris [9]. Recall that F_j denotes the label of feature j, in the sequel assumed to be $F_j \in (0, 1)$. $Y_i = (Y_{i,j})_{j \geq 1}$ denotes the observation of i-th individual. Each entry $Y_{i,j}$ in this sequence is distributed according to a Bernoulli distribution with parameter p_j and is independent of the others. $Y_{i,j} = 1$ (resp. $Y_{i,j} = 0$) indicates the presence (resp. absence) of feature j in the i-th individual. The Bayesian nonparametric approach to estimate the missing mass M_n requires a prior specification for the p_j's: we resort to the three-parameter Beta process introduced by [24]. Such a prior distribution is defined as the distribution of a Completely Random Measure (CRM) $\tilde{\mu}$ (see e.g. [7]), with Lévy intensity $\nu(\mathrm{d}s, \mathrm{d}f) = \rho(s)\mathrm{d}s\mathbb{I}(0,1)(f)\mathrm{d}f$, where $\rho(s) = \vartheta/\Gamma(1 - \alpha)s^{-1-\alpha}(1 - s)^{\beta+\alpha-1}$, being $\alpha, \beta \in (0, 1)$ and $\vartheta > 0$.

Let $\mathbf{Y}_n$ be a random sample of size n and $F_1^*, \ldots, F_{K_n}^*$ be the K_n distinct features observed in it, the posterior estimate of M_n, under a square loss function, equals

$$\tilde{M}_n(\vartheta, \alpha, \beta) = E[\tilde{\mu}((0, 1) \setminus \{F_1^*, \ldots, F_{K_n}^*\})|\mathbf{Y}_n]. \tag{A.1}$$

We can characterize the posterior distribution of $\tilde{\mu}$ in (A.1) resorting to [16, Theorem 3.1 (ii)], which gives

$$\tilde{\mu}|\mathbf{Y}_n \stackrel{d}{=} \tilde{\mu}_n + \sum_{\ell=1}^{K_n} J_\ell \delta_{F_\ell^*}$$

where the J_ℓ's are jumps having a density on the positive real line and $\tilde{\mu}_n$ is a CRM having updated Lévy intensity given by $\nu_n(\mathrm{d}s\,\mathrm{d}f) = (1-s)^n\rho(s)\mathrm{d}s\mathbb{I}(0,1)(f)\mathrm{d}f$. Hence the Bayesian estimator in (A.1) boils down to

$$\tilde{M}_n(\vartheta, \alpha, \beta) = \mathbb{E}\left[\tilde{\mu}_n(\{F_1^*, \ldots, F_{K_n}^*\}^c) + \sum_{\ell=1}^{K_n} J_\ell \delta_{F_\ell^*}(\{F_1^*, \ldots, F_{K_n}^*\}^c)\right]$$

$$= \mathbb{E}[\tilde{\mu}_n((0, 1))] = -\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}[e^{-u\tilde{\mu}_n((0,1))}]\Big|_{u=0} = \vartheta\frac{\Gamma(\alpha + \beta + n)}{\Gamma(\beta + n + 1)}$$

where the second equality follows from the fact that the base measure of $\tilde{\mu}_n$ is diffuse and the last equality follows from the availability of the Laplace

functional and from standard calculations. We are now going to show that $\hat{\vartheta}_n = (K_{n,1}\Gamma(\beta + n + 1))/(\Gamma(\beta + \alpha + n)n)$ is a consistent estimator of ϑ, then we will conclude that the empirical Bayes estimator $\tilde{M}_n(\hat{\vartheta}_n, \alpha, \beta)$ coincides with $\hat{M}_n$. The consistency of $\hat{\vartheta}_n$ can be established resorting to the regular variation theory by [17, 12]. More specifically we are able to show that the tail integral of $\rho(s)$ is a regularly varying function, having regular variation index α, indeed

$$\overline{\rho}(s) := \int_x^1 \rho(s)\mathrm{d}s = \frac{\vartheta}{\Gamma(1-\alpha)}\int_x^1 s^{-1-\alpha}(1-s)^{\beta+\alpha-1}\mathrm{d}s$$

fulfills

$$\lim_{x\to 0}\frac{\overline{\rho}(s)}{x^{-\alpha}} = \frac{\vartheta}{\Gamma(1-\alpha)}\lim_{x\to 0}\frac{x^{-1-\alpha}(1-x)^{\beta+\alpha-1}}{\alpha x^{-\alpha-1}} = \frac{\vartheta}{\Gamma(1-\alpha)\alpha}.$$

Hence we just proved that $\overline{\rho}(x) \simeq x^{-\alpha}\vartheta/(\Gamma(1-\alpha)\alpha)$ as $n \to +\infty$, and therefore an application of [12, Corollary 21] gives the asymptotic relation $\frac{K_{n,1}}{n^\alpha} \to \vartheta$. Moreover, from Stirling's approximation, $\frac{\Gamma(\alpha+\beta+n)}{\Gamma(n+1+\beta)} \simeq n^{\alpha-1}$. Therefore, $\hat{\vartheta}_n = \frac{K_{n,1}\Gamma(\beta+n+1)}{\Gamma(\beta+\alpha+n)n} \to \vartheta$ and $\hat{\vartheta}_n$ is a consistent estimator for ϑ. Plugging $\hat{\vartheta}_n$ into the posterior estimate $\tilde{M}_n(\vartheta, \alpha, \beta)$, we obtain $\tilde{M}_n(\hat{\vartheta}_n, \alpha, \beta) = \frac{K_{n,1}}{n} = \hat{M}_n$. $\qquad\square$

A.2. Proof of Theorem 2.1

We prove part i) (upper bound) and ii) (lower bound) of Theorem 2.1 separately.

We first focus on the upper bound for the L_2-risk of $\hat{M}_n$. In the sequel we denote by $M_{n,1}$ the total mass of features with frequency 1 in a sample of size n, formally $M_{n,1} = \sum_{j\geq 1} p_j\mathbb{I}\{X_{n,j} = 1\}$, we observe that the expectation of $M_{n,1}$ equals

$$\mathbb{E}(M_{n,1}) = \sum_{j\geq 1} p_j\mathbb{P}(X_{n,j} = 1) = \sum_{j\geq 1}\binom{n}{1}p_j^2(1-p_j)^{n-1}$$

and obviously $\mathbb{E}[M_{n,1}] \leq W$, for any sequence $(p_j)_{j\geq 1}$ belonging to the class $\mathscr{P}_W$.

In order to bound the L_2-risk of the estimator $\hat{M}_n$ from above for any $(p_j)_{j\geq 1} \in \mathscr{P}_W$, we remind that $R_n(\hat{M}_n; (p_j)_{j\geq 1}) = \mathbb{E}(\hat{M}_n - M_n)^2$ may be seen as the sum of the squared bias and the variance. The upper bound for the bias can be easily proved as follows

$$\begin{aligned}\mathbb{E}(\hat{M}_n - M_n) &= \frac{1}{n}\sum_{j\geq 1}\binom{n}{1}p_j^1(1-p_j)^{n-1} - \sum_{j\geq 1}p_j(1-p_j)^n \\ &= \sum_{j\geq 1}p_j(1-p_j)^{n-1} - \sum_{j\geq 1}p_j(1-p_j)^n = \sum_{j\geq 1}p_j^2(1-p_j)^{n-1} \\ &= \frac{1}{\binom{n}{1}}\sum_{j\geq 1}\binom{n}{1}p_j^2(1-p_j)^{n-1} = \frac{1}{n}\mathbb{E}(M_{n,1}) \leq \frac{1}{n}W.\end{aligned}$$

As for the variance bound, we note that

$$\mathrm{Var}(\hat{M}_n - M_n) = \sum_{j \geq 1} \mathrm{Var}\left(\frac{1}{n}\mathbb{I}\{X_{n,j} = 1\} - p_j\mathbb{I}\{X_{n,j} = 0\}\right)$$

$$\leq \sum_{j \geq 1} \mathbb{E}\left(\frac{1}{n}\mathbb{I}\{X_{n,j} = 1\} - p_j\mathbb{I}\{X_{n,j} = 0\}\right)^2,$$

where we have exploited the independence of the random variables $X_{n,j}$. We now observe that the events $\{X_{n,j} = 1\}$ and $\{X_{n,j} = 0\}$ are incompatible, hence we get

$$\mathrm{Var}(\hat{M}_n - M_n) \leq \sum_{j \geq 1} \mathbb{E}\left(\frac{1}{n}\mathbb{I}\{X_{n,j} = 1\} - p_j\mathbb{I}\{X_{n,j} = 0\}\right)^2$$

$$= \sum_{j \geq 1}\left(\left(\frac{1}{n}\right)^2 \mathbb{P}(X_{n,j} = 1) + p_j^2\mathbb{P}(X_{n,j} = 0)\right)$$

$$= \left(\frac{1}{n}\right)^2 \sum_{j \geq 1} \mathbb{P}(X_{n,j} = 1) + \sum_{j \geq 1} p_j^2\mathbb{P}(X_{n,j} = 0)$$

$$= \frac{1}{n^2} \sum_{j \geq 1} \binom{n}{1} p_j(1 - p_j)^{n-1} + \sum_{j \geq 1} p_j^2(1 - p_j)^n$$

$$= \frac{1}{n} \sum_{j \geq 1} p_j(1 - p_j)^{n-1} + \frac{1}{\binom{n+1}{1}} \sum_{j \geq 1} \binom{n+1}{1} p_j^2(1 - p_j)^n$$

$$= \frac{1}{n}\mathbb{E}(M_{n-1}) + \frac{1}{n+1}\mathbb{E}(M_{n+1,1}) \leq \frac{2n+1}{n(n+1)}W.$$

Putting together the bound on the bias and the variance we immediately obtain (2.3).

We now focus on the proof of the lower bound of the minimax risk R_n^*, which has been defined in (2.2). Let us first notice that, since $(p_j)_{j \geq 1}$ belongs to $\mathscr{P}_W$, then $M_n \leq W$ almost surely, and hence $(M_n - \hat{T}_n)^2 \geq (M_n - \min(\hat{T}_n, W))^2$. Therefore, we can restrict the minimum in (2.2) over all the possible estimators less than W, hence the problem boils down to determine an estimate from below for $\inf_{\hat{T}_n \leq W} \sup_{(p_j)_j \in \mathscr{P}_W} \mathbb{E}[(M_n - \hat{T}_n)^2]$. The proof of this estimate relies on a Bayesian parametric approach based on the randomization of the probability masses p_j's. More specifically we consider the randomized vector $\mathsf{P} = (\mathsf{p}_1, \cdots, \mathsf{p}_m, 0, 0, \ldots)$ where m is a Poisson random variable with mean $\lambda = nW$ and p_j are independent beta random variables with parameters $a = 1$ and $b = 2n - 1$. We denote by $\mathsf{P}_F := \sum_{j=1}^m \mathsf{p}_j$ the total mass of the random sequence $(\mathsf{p}_j)_{j \geq 1}$; we further define $X := (X_{n,j})_{j \geq 1}$ and it is worth noticing that, conditionally on P, the $X_{n,j}$'s are independent, having a Binomial distribution with parameters (n, p_j). For generic random variables U, V we will write E_V (resp. $E_{V|U}$) to mean that the expectation is made with respect to V (resp. to

V given U). It is easily shown that

$$
\inf_{\hat{T}_n \leq W} \sup_{(p_j)_j \in \mathscr{P}_W} \mathbb{E}[(M_n - \hat{T}_n)^2]
$$

$$
\geq \inf_{\hat{T}_n \leq W} \mathbb{E}_\mathsf{P} \left\{ \mathbb{I}\{\mathsf{P} \in \mathscr{P}_W\} \mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2 | \mathsf{P}] \right\}
$$

$$
\geq \inf_{\hat{T}_n \leq W} \left\{ \mathbb{E}_\mathsf{P} \mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2 | \mathsf{P}] \right.
$$

$$
\left. - \mathbb{E}_\mathsf{P}(1 - \mathbb{I}\{\mathsf{P} \in \mathscr{P}_W\}) \mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2 | \mathsf{P}] \right\}
$$

$$
\geq \inf_{\hat{T}_n \leq W} \left\{ \mathbb{E}_\mathsf{P} \mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2 | \mathsf{P}] - \mathbb{E}_\mathsf{P}(1 - \mathbb{I}\{\mathsf{P} \in \mathscr{P}_W\}) \mathsf{P}_F^2 \right\}
$$

$$
\geq \inf_{\hat{T}_n \leq W} \mathbb{E}_\mathsf{P} \mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2 | \mathsf{P}] - \mathbb{E}(\mathsf{P}_F^2 \mathbb{I}\{\mathsf{P}_F > W\}) \qquad (\text{A.2})
$$

We bound separately the two terms on the r.h.s. of (A.2). We start by deriving an upper bound for the term on the right. Using Fubini's Theorem, it comes that

$$
\mathbb{E}(\mathsf{P}_F^2 \mathbb{I}\{\mathsf{P}_F > W\})
$$

$$
= \int s^2 \mathbb{I}\{s > W\} \mathbb{P}_{\mathsf{P}_F}(\mathrm{d}s) = 2 \int \int t \mathbb{I}\{t < s\} \mathrm{d}t \mathbb{I}\{s > W\} \mathbb{P}_{\mathsf{P}_F}(\mathrm{d}s)
$$

$$
= 2 \int t \int \mathbb{I}\{t < s\} \mathbb{I}\{s > W\} \mathbb{P}_{\mathsf{P}_F}(ds)\mathrm{d}t = 2 \int t \mathbb{P}(\mathsf{P}_F > \max(t, W))\mathrm{d}t
$$

which leads to the more convenient form

$$
\mathbb{E}(\mathsf{P}_F^2 \mathbb{I}\{\mathsf{P}_F > W\}) = W^2 \mathbb{P}(\mathsf{P}_F > W) + 2 \int_W^\infty t \mathbb{P}(\mathsf{P}_F > t)\mathrm{d}t. \qquad (\text{A.3})
$$

In order to provide an upper bound for the probability of the event $\{\mathsf{P}_F > t\}$, we use a Markov inequality with third centralized moment. We first evaluate the mean of the total mass,

$$
\mathbb{E}\mathsf{P}_F = \mathbb{E}[\ \mathbb{E}(\sum_{j=1}^m p_j | m)\ ] = \mathbb{E}\frac{ma}{a+b} = \mathbb{E}\frac{m}{2n} = \frac{W}{2}.
$$

Now, denoting $\mu_3(Y)$ the centralized third moment of the r.v. Y, the law of total cumulant implies that

$$
\mathbb{E}(\mathsf{P}_F - \mathbb{E}\mathsf{P}_F)^3 = \mu_3(\mathsf{P}_F)
$$

$$
= \mathbb{E}[\mu_3(\mathsf{P}_F | m)] + \mu_3(\mathbb{E}[\mathsf{P}_F | m]) + 3\mathrm{Cov}(\mathbb{E}[\mathsf{P}_F | m], \mathrm{Var}[\mathsf{P}_F | m])
$$

$$
= \mathbb{E}[\mu_3(\mathsf{P}_F | m)] + \mu_3(\frac{m}{2n}) + 3\mathrm{Cov}\Big(\frac{m}{2n}, \frac{mab}{(a+b)^2(a+b+1)}\Big)
$$

$$
= \mathbb{E}[\mu_3(\mathsf{P}_F | m)] + \frac{1}{8n^3}\mu_3(m) + 3\frac{2n-1}{8n^3(2n+1)}\mathrm{Var}(m)
$$

m is Poisson distributed, so its variance and third centralized moment are equal to its mean $\lambda = nW$. Further noticing that $\frac{2n-1}{2n+1} < 1$, we upper bound the third centralized moment by $\mathbb{E}[\mu_3(\mathsf{P}_F|m)] + \frac{W}{2n^2}$. Conditioning on m, the total mass P_F is a sum of independent and identically distributed random variables so $\mu_3(\mathsf{P}_F|m) = m\mu_3(p_1)$. Since p_1 is Beta distributed with parameters a, and b, its third centralized moment is given by

$$\mu_3(p_1) = [\text{Var}(p_1)]^{3/2}\frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}} = \frac{2ab(b-a)}{(a+b)^3(a+b+1)(a+b+2)}$$

$$= \frac{4(2n-1)(n-1)}{(2n)^3(2n+1)2(n+1)} = \frac{(2n-1)(n-1)}{4n^3(2n+1)(n+1)}.$$

Using the fact that $\frac{(2n-1)(n-1)}{(2n+1)(n+1)} \leq 1 \leq 2$, we successively find that $\mathbb{E}[\mu_3(\mathsf{P}_F|m)] \leq \frac{W}{2n^2}$ and finally $\mu_3(\mathsf{P}_F) \leq \frac{W}{n^2}$, obtaining the inequality we need to bound $\mathbb{P}(\mathsf{P}_F > t)$. Indeed, Markov's inequality implies that

$$\mathbb{P}(\mathsf{P}_F > t) = \mathbb{P}\Big(\mathsf{P}_F - \mathbb{E}\mathsf{P}_F > t - W/2\Big) = \mathbb{P}\Big((\mathsf{P}_F - \mathbb{E}\mathsf{P}_F)^3 > (t - W/2)^3\Big)$$

$$\leq \frac{\mu_3(\mathsf{P}_F)}{(t - W/2)^3} \leq \frac{W}{n^2(t - W/2)^3}$$

Now we can use the last inequality in (A.3) and find

$$\mathbb{E}(\mathsf{P}_F^2\mathbb{I}\{\mathsf{P}_F > W\}) = W^2\mathbb{P}(\mathsf{P}_F > W) + 2\int_W^\infty t\mathbb{P}(\mathsf{P}_F > t)\mathrm{d}t$$

$$\leq W^2\frac{8W}{n^2W^3} + \frac{2W}{n^2}\int_W^\infty \frac{t}{(t - W/2)^3}\mathrm{d}t$$

$$= \frac{8}{n^2} + \frac{2W}{n^2}\Big[\int_W^\infty \frac{1}{(t - W/2)^2}\mathrm{d}t + \int_W^\infty \frac{W/2}{(t - W/2)^3}\mathrm{d}t\Big]$$

$$= \frac{8}{n^2} + \frac{2W}{n^2}\Big[\frac{2}{W} + \frac{1}{W}\Big] = \frac{14}{n^2}$$

We now provide a lower bound for the first term on the r.h.s. of (A.2)

$$\inf_{\hat{T}_n \leq W} \mathbb{E}_\mathsf{P}\mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2|\mathsf{P}]$$

$$= \inf_{\hat{T}_n \leq W} \mathbb{E}_m\mathbb{E}_{\mathsf{P}|m}\Big\{\mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2|\mathsf{P}] \mid m\Big\}$$

$$\geq \mathbb{E}_m \inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathsf{P}|m}\Big\{\mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2|\mathsf{P}] \mid m\Big\}. \qquad (A.4)$$

Let $m \in \mathbb{N}$, with the convention that the risk is 0 when $m = 0$. We will now lower bound inside the first expectation in the previous expression, using the fact that given X and m, we know the posterior of P_F.

$$\inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathsf{P}|m}\mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2|\mathsf{P}]$$

$$= \inf_{\hat{T}_n \leq W} \mathbb{E}_{X|m} \mathbb{E}_{\mathsf{P}|X,m}[(M_n - \hat{T}_n)^2 | X, m]$$

$$\geq \mathbb{E}_{X|m} \inf_{\hat{T}_n} \mathbb{E}_{\mathsf{P}|X,m}[(M_n - \hat{T}_n)^2 | X, m]$$

$$\geq \mathbb{E}_{X|m} \mathrm{Var}_{\mathsf{P}|X,m}(M_n | X, m) = \mathbb{E}_{X|m} \mathrm{Var}_{\mathsf{P}|X,m} \left(\sum_{j=1}^{m} p_j \mathbb{I}\{X_{n,j} = 0\} \Big| X, m \right)$$

$$= \mathbb{E}_{X|m} \left[\sum_{j=1}^{m} \mathbb{I}\{X_{n,j} = 0\} \frac{a(n+b)}{(a+n+b)^2(a+n+b+1)} \right]$$

where we used the fact that $\mathsf{p}_j | X_{n,j} = x$ is a beta random variable $\mathrm{Beta}(x + a, n - x + b)$. Then,

$$\inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathsf{P}|m} \mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2 | \mathsf{P}]$$

$$\geq \sum_{j=1}^{m} \mathbb{P}(X_{n,j} = 0) \frac{a(n+b)}{(a+n+b)^2(a+n+b+1)}$$

$$= \sum_{j=1}^{m} \mathbb{E}_{\mathsf{p}_j}[(1 - \mathsf{p}_j)^n] \frac{a(n+b)}{(a+n+b)^2(a+n+b+1)}$$

$$= \sum_{j=1}^{m} \frac{\Gamma(b+n)\Gamma(a+b)}{\Gamma(b)\Gamma(a+b+n)} \frac{a(n+b)}{(a+n+b)^2(a+n+b+1)} \tag{A.5}$$

where we observed that $(1 - \mathsf{p}_j) \sim \mathrm{Beta}(b, a)$ to obtain (A.5). Therefore we can derive the bound

$$\inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathsf{P}|m} \mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2 | \mathsf{P}] \geq m \frac{\Gamma(b+n)\Gamma(1+b)}{\Gamma(b)\Gamma(1+b+n)} \frac{n+b}{(n+b+1)^2(n+b+2)}$$

$$= \frac{mb}{(n+b+1)^2(n+b+2)}$$

$$= \frac{2mn}{9n^2(3n+1)} = \frac{2m}{9n(3n+1)}$$

Together with (A.4), the previous inequality implies that

$$\inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathsf{P}} \mathbb{E}_{X|\mathsf{P}}[(M_n - \hat{T}_n)^2 | \mathsf{P}] \geq \frac{2W}{9(3n+1)}.$$

Plugging in (A.2) the previous results, we obtain

$$R_n^* = \inf_{\hat{T}_n} \sup_{(p_j)_j \in \mathscr{P}_W} \mathbb{E}[(M_n - \hat{T}_n)^2] \geq \frac{2W}{9(3n+1)} - \frac{14}{n^2},$$

which concludes the proof. Notice that here we are only interested in showing that the minimax rate is or order $\frac{W}{n}$. We could have obtained sharper bounds (with a better constant) by for instance taking $b = (1 + \epsilon)n - 1$ instead of $b = 2n - 1$. $\square$

A.3. Details for the determiantion of (2.5) and (2.6)

The concentration inequalities (2.5)–(2.6) can be proved my means of suitable bounds on the log-Laplace transform of $\hat{W}_n - W$, using the techniques of [3]. To this end we recall that a random variable Z is sub-Gaussian on the right tail (resp. left tail) with variance factor v if for all $\lambda \geq 0$ (resp. $\lambda \leq 0$) one has

$$\log \mathbb{E} e^{\lambda(Z - \mathbb{E} Z)} \leq \frac{v\lambda^2}{2}. \tag{A.6}$$

We will also say that a random variable Z is sub-Gamma on the right tail with variance factor v and scale parameter c if

$$\log \mathbb{E} e^{\lambda(Z - \mathbb{E} Z)} \leq \frac{\lambda^2 v}{2(1 - c\lambda)}, \quad \text{for any } \lambda \text{ such that } 0 \leq \lambda \leq 1/c.$$

Furthermore Z is said to be sub-Gamma on the left tail with variance factor v and scale parameter c, if $-Z$ is sub-Gamma on the right tail with variance factor v and scale parameter c. These types of bounds on the log-Laplace transform imply exponential tail bounds, useful to find confidence intervals, for example if Z is sub-Gamma on the right tail as above then $\mathbb{P}(Z > \mathbb{E} Z + \sqrt{2vs} + cs) \leq e^{-s}$ for any $s \geq 0$. Analogously if Z is sub-Gaussian on the left tail (i.e. (A.6) holds true for all $\lambda \leq 0$), then $\mathbb{P}(Z < \mathbb{E} Z - \sqrt{2vs}) \leq e^{-s}$ for any $s \geq 0$. See [3] for additional details on this subject. In the next proposition we prove that $\hat{W}_n$ is sub-Gamma on the right tail and sub-Gaussian on the left tail, providing suitable bounds on the log-Laplace transform. As just explained these bounds on the log-Laplace immediately implies the exponential tail bounds (2.5) and (2.6).

Proposition A.1. *The random variable* $\hat{W}_n := \sum_{j \geq 1} X_{n,j}/n$ *is an unbiased estimator of* W. *In addition* $\hat{W}_n$ *is sub-Gamma on the right tail with variance factor* $w_n^+ := 2W/n$ *and scale factor* $1/n$, *i.e. for any* $0 < \lambda < n$

$$\log \mathbb{E} \left(e^{\lambda(\hat{W}_n - W)} \right) \leq \frac{w_n^+ \lambda^2}{2(1 - \lambda/n)}; \tag{A.7}$$

on the left tail, $\hat{W}_n$ *is sub-Gaussian with variance factor* $w_n^- = W/n$, *namely for all* $\lambda \geq 0$ *we have*

$$\log \mathbb{E} \left(e^{-\lambda(\hat{W}_n - W)} \right) \leq \frac{\lambda^2 w_n^-}{2}. \tag{A.8}$$

Proof. One can immediately show that $\mathbb{E}(\hat{W}_n) = W$. We first prove (A.7), that is to say $\hat{W}_n - W$ is sub-Gamma on the right tail, indeed, by the independence of the $X_{n,j}$'s, we get

$$\log \mathbb{E} \left(e^{\lambda(\hat{W}_n - W)} \right) = \sum_{j \geq 1} \log \mathbb{E} \left(e^{\lambda/n(X_{n,j} - \mathbb{E} X_{n,j})} \right)$$

$$= \sum_{j \geq 1} \left(-\lambda p_j + \log \mathbb{E} e^{\lambda/n X_{n,j}} \right)$$

$$= \sum_{j \geq 1} \left(-\lambda p_j + n \log(1 - p_j + p_j e^{\lambda/n}) \right)$$

having observed that $X_{n,j}$ is a Binomial random variable with parameters n and p_j. An application of the inequality $\log(z) \leq z - 1$, for $z > 0$, yields

$$\log \mathbb{E} \left(e^{\lambda(\hat{W}_n - W)} \right) \leq n \sum_{j \geq 1} p_j (e^{\lambda/n} - 1 - \lambda/n) = n\phi(\lambda/n)\mathbb{E}\hat{W}_n \qquad \text{(A.9)}$$

where $\phi(t) = e^t - 1 - t$. By a series expansion of the exponential function, we obtain

$$\log \mathbb{E} \left(e^{\lambda(\hat{W}_n - W)} \right) \leq n \sum_{k \geq 2} \frac{1}{k!} \left(\frac{\lambda}{n} \right)^k \mathbb{E}\hat{W}_n \leq n \sum_{k \geq 2} \left(\frac{\lambda}{n} \right)^k \mathbb{E}\hat{W}_n$$

$$= n \cdot \frac{\lambda^2}{n^2} \cdot \frac{1}{1 - \lambda/n} \mathbb{E}\hat{W}_n = \frac{w_n^+ \lambda^2}{2(1 - \lambda/n)}$$

and then (A.7) follows for any $\lambda \in (0, n)$. The proof of (A.8) goes along similar lines, indeed proceeding as before one can prove that for any $\lambda \geq 0$

$$\log \mathbb{E} \left(e^{-\lambda(\hat{W}_n - W)} \right) \leq n\phi(-\lambda/n)\mathbb{E}\hat{W}_n$$

and then (A.8) follows by observing that $\phi(-\lambda/n) \leq \lambda^2/(2n^2)$. □

A.4. Proof of Theorem 3.1 and related results

In this section we focus on Theorem 3.1. First of all we introduce some preliminary results which are useful to prove the theorem and are interesting in their own right. In the next proposition we derive concentration inequalities for both $K_{n,r}$ and K_n.

Proposition A.2. *For any $n \geq 1$, $r \geq 1$ and $\delta \in (0, 1)$ we have*

$$\mathbb{P} \left(|K_{n,r} - \mathbb{E}(K_{n,r})| \leq \frac{2}{3} \log \left(\frac{2}{\delta} \right) + \sqrt{2\mathbb{E}(K_{n,r}) \log \left(\frac{2}{\delta} \right)} \right) \geq 1 - \delta \quad \text{(A.10)}$$

and

$$\mathbb{P} \left(|K_n - \mathbb{E}(K_n)| \leq \frac{2}{3} \log \left(\frac{2}{\delta} \right) + \sqrt{2\mathbb{E}(K_n) \log \left(\frac{2}{\delta} \right)} \right) \geq 1 - \delta. \quad \text{(A.11)}$$

Proof. Let us focus on the derivation of (A.10). First of all we will determine a variance bound for the random variable $K_{n,r}$, which will be employed to derive

a corresponding bound on the log-Laplace transform of $(K_{n,r} - \mathbb{E}(K_{n,r}))$. Then the result will follow by a suitable application of the Bernstein inequality.

Thanks to the independence of the random variables $X_{n,j}$'s, the variance of $K_{n,r}$ may be bounded as follows

$$\mathrm{Var}(K_{n,r}) = \sum_{j \geq 1} \mathrm{Var}\left(\mathbb{I}\{X_{n,j} = r\}\right) \leq \sum_{j \geq 1} \mathbb{E}\left(\mathbb{I}\{X_{n,j} = r\}\right) = \mathbb{E}(K_{n,r}). \quad \text{(A.12)}$$

We now establish the bound on the log-Laplace. Since $K_{n,r}$ is the sum of independent random variables, for any $t \in \mathbb{R}$ we can write:

$$\log \mathbb{E}(e^{t(K_{n,r} - \mathbb{E}(K_{n,r}))}) = \sum_{j=1}^{\infty} \log \mathbb{E} \exp\left\{(\mathbb{I}\{X_{n,j} = r\} - \mathbb{E}\mathbb{I}\{X_{n,j} = r\})\right\}$$

$$\leq \phi(|t|) \sum_{j=1}^{\infty} \mathrm{Var}(\mathbb{I}\{X_{n,j} = r\}) = \phi(|t|)\mathrm{Var}(K_{n,r})$$

$$\overset{\text{(A.12)}}{\leq} \phi(|t|)\mathbb{E}(K_{n,r})$$

where we have implicitly defined the function $\phi(t) = e^t - 1 - t$ and we have applied the Bennett inequality. The validity of the previous bound on the log-Laplace implies that for any $x \geq 0$

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

thanks to the Bernstein inequality. If one defines $s := x^2/(2(\mathbb{E}(K_{n,r}) + x/3))$ the previous inequality boils down to the following one

$$\mathbb{P}\left(|K_{n,r} - \mathbb{E}(K_{n,r})| \geq \frac{2}{3}s + \sqrt{2s\mathbb{E}(K_{n,r})}\right) \leq 2e^{-s}$$

which in turn implies the validity of (A.10), as a consequence of the change of variable $2e^{-s} = \delta$ and the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, valid for any positive real numbers a, b. The proof of (A.11) goes along similar lines, having observed that $\mathrm{Var}(K_n) \leq \mathbb{E}(K_n)$. □

The next Remark is a simple consequence of Proposition A.2 and will be used in the derivation of the confidence interval for $\hat{M}_n$ to upper bound the expected values $\mathbb{E}(K_{n,r})$ and $\mathbb{E}(K_n)$ in terms of the two statistics $K_{n,r}$ and K_n respectively.

Remark A.1. *The concentration inequality for $K_{n,r}$ (A.10) implies that with probability bigger than $1 - \delta$*

$$K_{n,r} \geq \mathbb{E}(K_{n,r}) - \frac{2}{3}\log\left(1/\delta\right) - \sqrt{2\log\left(1/\delta\right)\mathbb{E}(K_{n,r})}.$$

Solving the previous inequality with respect to $\sqrt{\mathbb{E}(K_{n,r})}$, we can conclude that with probability at least $1 - \delta$

$$\sqrt{\mathbb{E}(K_{n,r})} \leq \sqrt{\frac{1}{2} \log{(1/\delta)}} + \sqrt{\frac{7}{6} \log{(1/\delta)} + K_{n,r}}.$$

Analogously we can employ the concentration inequality for K_n (see (A.11)) to state that the event

$$\sqrt{\mathbb{E}(K_n)} \leq \sqrt{\frac{1}{2} \log{(1/\delta)}} + \sqrt{\frac{7}{6} \log{(1/\delta)} + K_n}$$

has probability at least $1 - \delta$.

The confidence interval we suggest is mainly based on a concentration inequality for $\hat{M}_n - M_n$. Such concentration result is proved my means of suitable bounds on the log-Laplace transform of $\hat{M}_n - M_n$, as in Section A.3, more precisely in Proposition A.3 we will prove that $\hat{M}_n - M_n$ is a sub-Gamma random variable on the right and the left tails (see [3]). In order to provide the bounds on the log-Laplace transform of $\hat{M}_n - M_n$, the following lemma is required.

Lemma A.1. *For any $n \geq 1$, we have*

$$\mathbb{E}(\hat{M}_n - M_n) \leq \frac{2}{n(n-1)} \mathbb{E}(K_{n,2}), \tag{A.13}$$

and

$$\mathrm{Var}(\hat{M}_n - M_n) \leq \frac{1}{n^2} \mathbb{E}(K_{n,1}) + \frac{2}{n(n-1)} \mathbb{E}(K_{n,2}). \tag{A.14}$$

Proof. First of all, we focus on the proof of (A.13), straightforward calculations lead to

$$\begin{aligned}
\mathbb{E}(\hat{M}_n - M_n) &= \frac{1}{n} \sum_{j \geq 1} \binom{n}{1} p_j (1 - p_j)^{n-1} - \sum_{j \geq 1} p_j (1 - p_j)^n \\
&= \sum_{j \geq 1} p_j (1 - p_j)^{n-1} - \sum_{j \geq 1} p_j (1 - p_j)^n \\
&= \sum_{j \geq 1} p_j^2 (1 - p_j)^{n-1} \leq \sum_{j \geq 1} p_j^2 (1 - p_j)^{n-2} \\
&= \frac{2}{n(n-1)} \sum_{j \geq 1} \binom{n}{2} p_j^2 (1 - p_j)^{n-2} = \frac{2}{n(n-1)} \mathbb{E}(K_{n,2})
\end{aligned}$$

where we have observed that $(1 - p_j)^{n+1-2} \leq (1 - p_j)^{n-2}$, which is legitimated by the fact that n is a sample size and can be assumed sufficiently large.

Second, we focus on the proof of the variance bound (A.14). Exploiting the independence of the random variables $X_{n,j}$'s, we get

$$\mathrm{Var}(\hat{M}_n - M_n) = \sum_{j \geq 1} \mathrm{Var}\left(\frac{1}{n}\mathbb{I}\{X_{n,j} = 1\} - p_j\mathbb{I}\{X_{n,j} = 0\}\right)$$

$$\leq \sum_{j \geq 1} \mathbb{E}\left(\frac{1}{n}\mathbb{I}\{X_{n,j} = 1\} - p_j\mathbb{I}\{X_{n,j} = 0\}\right)^2.$$

We now observe that the events $\{X_{n,j} = 1\}$ and $\{X_{n,j} = 0\}$ are incompatible, hence

$$\mathrm{Var}(\hat{M}_n - M_n) \leq \sum_{j \geq 1} \mathbb{E}\left(\frac{1}{n}\mathbb{I}\{X_{n,j} = 1\} - p_j\mathbb{I}\{X_{n,j} = 0\}\right)^2$$

$$= \sum_{j \geq 1} \left(\left(\frac{1}{n}\right)^2 \mathbb{P}(X_{n,j} = 1) + p_j^2 \mathbb{P}(X_{n,j} = 0)\right)$$

$$= \frac{1}{n^2}\sum_{j \geq 1}\mathbb{P}(X_{n,j} = 1) + \sum_{j \geq 1}p_j^2\mathbb{P}(X_{n,j} = 0)$$

$$= \frac{1}{n^2}\mathbb{E}(K_{n,1}) + \sum_{j \geq 1}p_j^2(1 - p_j)^n \leq \frac{1}{n^2}\mathbb{E}(K_{n,1}) + \sum_{j \geq 1}p_j^2(1 - p_j)^{n-2}$$

$$= \frac{1}{n^2}\mathbb{E}(K_{n,1}) + \frac{2}{n(n-1)}\mathbb{E}(K_{n,2})$$

and the bound on the variance now follows. $\qquad\square$

To fix the notation, it is worth to define the quantities

$$\mathbb{E}[K(n)] = \sum_{j \geq 1}(1 - e^{-np_j}), \quad \mathbb{E}[K_k(n)] = \sum_{j \geq 1}e^{-np_j}\frac{(np_j)^k}{k!}$$

for any $k \geq 1$. It can be easily seen that $\mathbb{E}[K_k(n)] \leq \mathbb{E}[K(n)]$, and in addition $\mathbb{E}[K(n)]$ is an increasing function of n. We are now ready to prove that $\hat{M}_n - M_n$ is sub-Gamma on its tails.

Proposition A.3. *The variable $D_n = \hat{M}_n - M_n$ is sub–gamma on the right tail with variance factor $v_n^+ = 2\{\mathbb{E}(K_{n,1}) + 2n^2\mathbb{E}(K_{n,2})/[n \cdot (n-1)]\}/n^2$ and scale factor $1/n$, i.e. for any $0 < \lambda < n$*

$$\log \mathbb{E}\left(e^{\lambda[D_n - \mathbb{E}(D_n)]}\right) \leq \frac{v_n^+\lambda^2}{2(1 - \lambda/n)}. \tag{A.15}$$

On the left tail, D_n is again sub–gamma with variance factor $v_n^- = 4\mathbb{E}[K(n)]/(n-1)^2$ and scale parameter $1/(n-1)$, i.e. for any $0 < \lambda < (n-1)$

$$\log \mathbb{E}\left(e^{-\lambda[D_n - \mathbb{E}(D_n)]}\right) \leq \frac{v_n^-\lambda^2}{2[1 - \lambda/(n-1)]}. \tag{A.16}$$

Proof. First we prove (A.15), observe that when $\lambda > 0$

$$\log \mathbb{E}\left(e^{\lambda[D_n - \mathbb{E}(D_n)]}\right)$$
$$= \sum_{j \geq 1} \log \mathbb{E} \exp\left\{\frac{\lambda}{n}\left[\mathbb{I}\{X_{n,j} = 1\} - np_j\mathbb{I}\{X_{n,j} = 0\}\right.\right.$$
$$\left.\left. - \mathbb{E}\left(\mathbb{I}\{X_{n,j} = 1\} - np_j\mathbb{I}\{X_{n,j} = 0\}\right)\right]\right\}.$$

For the sake of simplifying notation, define $H_j := \mathbb{I}\{X_{n,j} = 1\} - np_j\mathbb{I}\{X_{n,j} = 0\} \leq 1$, hence we have

$$\log \mathbb{E}\left(e^{\lambda[D_n - \mathbb{E}(D_n)]}\right) = \sum_{j \geq 1} \log \mathbb{E} \exp\left\{\frac{\lambda}{n}(H_j - \mathbb{E}H_j)\right\}$$
$$= \sum_{j \geq 1}\left(\log \mathbb{E}[e^{(\lambda/n)H_j}] - \frac{\lambda}{n}\mathbb{E}H_j\right)$$
$$\leq \sum_{j \geq 1}\left(\mathbb{E}[e^{(\lambda/n)H_j} - 1] - \frac{\lambda}{n}\mathbb{E}H_j\right)$$

where we used the inequality $\log(z) \leq z - 1$ for any positive z. Now, observing that the function $f(z) := \frac{e^z - 1 - z}{z^2}$ is increasing and that $H_j \leq 1$, we get

$$\log \mathbb{E}\left(e^{\lambda[D_n - \mathbb{E}(D_n)]}\right) \leq \sum_{j \geq 1} \mathbb{E}\left[e^{(\lambda/n)H_j} - 1 - \frac{\lambda}{n}H_j\right]$$
$$= \sum_{j \geq 1} \mathbb{E}\left\{\frac{e^{(\lambda/n)H_j} - 1 - \frac{\lambda}{n}H_j}{[(\lambda/n)H_j]^2}[(\lambda/n)H_j]^2\right\}$$
$$\leq \sum_{j \geq 1} \mathbb{E}(H_j^2)(e^{\lambda/n} - 1 - \lambda/n)$$
$$= \phi(\lambda/n) \sum_{j \geq 1} \mathbb{E}\left(\mathbb{I}\{X_{n,j} = 1\} - np_j\mathbb{I}\{X_{n,j} = 0\}\right)^2$$
$$\leq \sum_{k \geq 2}\left(\frac{\lambda}{n}\right)^k \sum_{j \geq 1} \mathbb{E}\left(\mathbb{I}\{X_{n,j} = 1\} - np_j\mathbb{I}\{X_{n,j} = 0\}\right)^2$$
$$= \frac{\lambda^2}{1 - \lambda/n} \sum_{j \geq 1} \mathbb{E}\left(\frac{1}{n}\mathbb{I}\{X_{n,j} = 1\} - p_j\mathbb{I}\{X_{n,j} = 0\}\right)^2$$

now one can proceed along similar arguments as in the proof of (A.14) to obtain

$$\log \mathbb{E}\left(e^{\lambda[D_n - \mathbb{E}(D_n)]}\right) \leq \frac{\lambda^2}{1 - \lambda/n}\left(\frac{1}{n^2}\mathbb{E}(K_{n,1}) + \frac{2}{n(n-1)}\mathbb{E}(K_{n,2})\right)$$

hence (A.15) now follows.

Finally we prove (A.16), analogous calculations as before show that

$$\log \mathbb{E}\left(e^{-\lambda[D_n - \mathbb{E}(D_n)]}\right) = \sum_{j \geq 1} \log \mathbb{E} \exp\left\{\lambda p_j (W_j - \mathbb{E}W_j)\right\}$$

having defined $W_j := \mathbb{I}\{X_{n,j} = 0\} - \frac{1}{np_j}\mathbb{I}\{X_{n,j} = 1\} \leq 1$. Thanks to the inequality $\log(z) \leq z - 1$, which is valid for any positive z, and applying similar considerations as those used in the first part of the proof, we have

$$\log \mathbb{E}\left(e^{-\lambda[D_n - \mathbb{E}(D_n)]}\right) = \sum_{j \geq 1}\left\{\log[E(e^{\lambda p_j W_j})] - \lambda p_j \mathbb{E}W_j\right\}$$

$$\leq \sum_{j \geq 1}\mathbb{E}[e^{\lambda p_j W_j} - 1 - \lambda p_j W_j] = \sum_{j \geq 1}\mathbb{E}\left[\frac{e^{\lambda p_j W_j} - 1 - \lambda p_j W_j}{(\lambda p_j W_j)^2}(\lambda p_j W_j)^2\right]$$

$$\leq \sum_{j \geq 1}\mathbb{E}\left[\frac{e^{\lambda p_j} - 1 - \lambda p_j}{(\lambda p_j)^2}(\lambda p_j W_j)^2\right] = \sum_{j \geq 1}\mathbb{E}W_j^2 \phi(\lambda p_j)$$

$$= \sum_{j \geq 1}\phi(\lambda p_j)\mathbb{E}\left(\mathbb{I}\{X_{n,j} = 0\} - \frac{1}{np_j}\mathbb{I}\{X_{n,j} = 1\}\right)^2$$

$$\leq \sum_{j \geq 1}\phi(\lambda p_j)\mathbb{E}\left(\mathbb{I}\{X_{n,j} = 0\} + \frac{1}{n^2 p_j^2}\mathbb{I}\{X_{n,j} = 1\}\right)$$

$$= \sum_{j \geq 1}\phi(\lambda p_j)\left[(1 - p_j)^n + \frac{1}{np_j}(1 - p_j)^{n-1}\right]$$

$$= \sum_{j \geq 1}\sum_{k \geq 2}\frac{(\lambda p_j)^k}{k!}\left[(1 - p_j)^n + \frac{1}{np_j}(1 - p_j)^{n-1}\right]$$

$$\leq \sum_{k \geq 2}\lambda^k \sum_{j \geq 1}\left(\frac{p_j^k}{k!}e^{-np_j} + \frac{p_j^{k-1}}{nk!}e^{-(n-1)p_j}\right),$$

where we have observed that $(1 - p)^n \leq e^{-np}$. Recalling the definition of $K_k(n)$ and that $\mathbb{E}[K_k(n)] \leq \mathbb{E}[K(n)]$, we get

$$\log \mathbb{E}\left(e^{-\lambda[D_n - \mathbb{E}(D_n)]}\right)$$

$$\leq \sum_{k \geq 2}\lambda^k \left\{\frac{1}{n^k}\sum_{j \geq 1}\frac{(np_j)^k}{k!}e^{-np_j} + \frac{1}{kn(n-1)^{k-1}}\sum_{j \geq 1}\frac{[(n-1)p_j]^{k-1}}{(k-1)!}e^{-(n-1)p_j}\right\}$$

$$= \sum_{k \geq 2}\lambda^k \left\{\frac{1}{n^k}\mathbb{E}[K_k(n)] + \frac{1}{kn(n-1)^{k-1}}\mathbb{E}[K_{k-1}(n-1)]\right\}$$

$$\leq \sum_{k \geq 2}\left(\frac{\lambda}{n-1}\right)^k \left\{\mathbb{E}[K_k(n)] + \frac{1}{k}\mathbb{E}[K_{k-1}(n-1)]\right\}$$

$$\leq \{\mathbb{E}[K(n)] + \mathbb{E}[K(n-1)]\} \sum_{k \geq 2} \left(\frac{\lambda}{n-1}\right)^k \leq \frac{2\mathbb{E}[K(n)]}{(n-1)^2} \frac{\lambda^2}{1 - \lambda/(n-1)},$$

hence (A.16) has been now established. □

Finally we prove Theorem 3.1.

Proof of Theorem 3.1. The result is a consequence of the two bounds that we are now going to prove separately:

$$\mathbb{P}\left(M_n \geq \hat{M}_n - \hat{m}_n - \sqrt{2\hat{v}_n^+ \log(1/\delta)} - \log(1/\delta)/n\right) \geq 1 - 4\delta \qquad (\text{A.17})$$

$$\mathbb{P}\left(M_n \leq \hat{M}_n + \sqrt{2\hat{v}_n^- \log(1/\delta)} + \log(1/\delta)/(n-1)\right) \geq 1 - 2\delta, \qquad (\text{A.18})$$

where $\hat{m}_n, \hat{v}_n^-$ and $\hat{v}_n^+$ are suitable quantities that will be defined in the proof. First we discuss how to determine (A.17), we remind that $D_n = \hat{M}_n - M_n$ is a sub–Gamma random variable, as shown in Proposition A.3, hence the following holds (see [2])

$$\mathbb{P}(D_n > \mathbb{E}(D_n) + \sqrt{2v_n^+ s} + s/n) \leq e^{-s}$$

for any $s \geq 0$. Putting $e^{-s} = \delta$, we obtain

$$\mathbb{P}(D_n \leq \mathbb{E}(D_n) + \sqrt{2v_n^+ \log(1/\delta)} + \log(1/\delta)/n) \geq 1 - \delta. \qquad (\text{A.19})$$

On the basis of (A.13), we can upper bound $\mathbb{E}(D_n)$ in (A.19) as

$$\mathbb{E}(D_n) \leq \frac{2}{n(n-1)}\mathbb{E}(K_{n,2}),$$

and, as a consequence of Remark A.3.2, the following inequality is true with probability bigger than $1 - \delta$

$$\mathbb{E}(D_n) \leq \frac{2}{(n-1)n} \left(\sqrt{\frac{1}{2}\log(1/\delta)} + \sqrt{\frac{7}{6}\log(1/\delta) + K_{n,2}}\right)^2 =: \hat{m}_n. \qquad (\text{A.20})$$

Remark A.3.2 also applies to upper bound the expected values in v_n^+ with the corresponding empirical values, indeed the following bound holds true

$$v_n^+ \leq \frac{2}{n^2} \left(\sqrt{\frac{1}{2}\log(1/\delta)} + \sqrt{\frac{7}{6}\log(1/\delta) + K_{n,1}}\right)^2$$
$$+ \frac{4}{n(n-1)} \left(\sqrt{\frac{1}{2}\log(1/\delta)} + \sqrt{\frac{7}{6}\log(1/\delta) + K_{n,2}}\right)^2 =: \hat{v}_n^+ \qquad (\text{A.21})$$

with probability bigger than $1 - 2\delta$. Using the inequalities (A.20)–(A.21) in (A.19) we can conclude

$$\mathbb{P}(D_n \leq \hat{m}_n + \sqrt{2\hat{v}_n^+ \log(1/\delta)} + \log(1/\delta)/n) \geq 1 - 4\delta,$$

which gives (A.17).

Now we focus on the derivation of (A.18), first of all recall that by Proposition A.3, $-D_n$ is a sub–Gamma random variable, hence

$$\mathbb{P}(-D_n > -\mathbb{E}[D_n] + \sqrt{2v_n^- s} + s/(n-1)) \leq e^{-s}$$

for any $s \geq 0$. Define $e^{-s} = \delta$, then the previous bound implies that

$$\mathbb{P}(D_n \geq \mathbb{E}[D_n] - \sqrt{2v_n^- \log(1/\delta)} - \log(1/\delta)/(n-1)) \geq 1 - \delta$$

since $\mathbb{E}[D_n] \geq 0$, we get

$$\mathbb{P}(D_n \geq -\sqrt{2v_n^- \log(1/\delta)} - \log(1/\delta)/(n-1)) \geq 1 - \delta. \tag{A.22}$$

One can now apply the bound contained in [12, Lemma 1]

$$|\mathbb{E}[K(n)] - \mathbb{E}(K_n)| \leq \frac{2}{n}\mathbb{E}[K_2(n)] \leq \frac{2}{n}\mathbb{E}[K(n)],$$

which entails that $\mathbb{E}[K(n)] \leq \mathbb{E}(K_n)/(1 - 2/n)$, and hence

$$v_n^- \leq \frac{4}{(n-1)^2(1-2/n)}\mathbb{E}(K_n).$$

Thanks to Remark A.3.2, with probability bigger than $1 - \delta$, the following holds

$$v_n^- \leq \frac{4}{(n-1)^2(1-2/n)}\left(\sqrt{\frac{1}{2}\log(1/\delta)} + \sqrt{\frac{7}{6}\log(1/\delta) + K_n}\right)^2 =: \hat{v}_n^-. \tag{A.23}$$

Using (A.23) in (A.22), we get

$$\mathbb{P}(D_n \geq -\sqrt{2\hat{v}_n^- \log(1/\delta)} - \log(1/\delta)/(n-1)) \geq 1 - 2\delta,$$

which in turn implies (A.18). Putting together (A.17)–(A.18), the thesis now follows. □

Acknowledgements

References

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

[2] 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. MR3556773

[3] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities.* Oxford University Press. MR3185193

[4] 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. MR3665906

[5] 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.

[6] 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.

[7] Daley, D.J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes. Vol. II.* Springer, New York. MR2371524

[8] Efron, B. (1987). *The jackknife, the bootstrap, and other resampling plans.* CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics. MR0659849

[9] 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. MR0388597

[10] Good, I.J. (1953). On the population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264. MR0061330

[11] Görür, D., Jäkel, F. and Rasmussen, C.E. (2006). A choice model with infinitely many latent features. *23rd International Conference on Machine Learning.*

[12] 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. MR2318403

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

[14] Hjort, N. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *The Annals of Statistics* **18**, 1259–1294. MR1062708

[15] 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* **106**, 5008–5013. MR2496537

[16] James, L.F. (2017). Bayesian Poisson calculus for latent feature modeling

via generalized Indian buffet process priors. *The Annals of Statistics* **45**, 2016–2045. MR3718160

[17] KARLIN, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, **17**, 373–401. MR0216548

[18] MEEDS, E., GHAHRAMANI, Z., NEAL, R. AND ROWIES, S.T. (2007). Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems.*

[19] MILLER, K.T., GRIFFITHS, T.L. AND JORDAN, M.I. (2010). Nonparametric latent feature models for link predictions. *Advances in Neural Information Processing Systems.* MR3022515

[20] NAVARRO, D.J. AND GRIFFITHS, T.L. (2010). A nonparametric Bayesian model for inferring features from similarity judgments. *Advances in Neural Information Processing Systems.* MR2463044

[21] QUENOUILLE, M.H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353–360. MR0081040

[22] RAJARAMAN, N., THANGARAJ, A. AND SURESH, A.T. (2017) Minimax risk for missing mass estimation. *Proceedings of the IEEE International Symposium on Information Theory.*

[23] ROBBINS, H. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *The Annals of Mathematical Statistics* **39**, 256–257. MR0221695

[24] TEH, Y.W. AND GÖRÜR, D. (2009). Indian buffet processes with power–law behavior. *Advances in Neural Information Processing Systems.*

[25] TUKEY, J.W. (1958). Bias and confidence in not-quite large samples. *The Annals of Mathematical Statistics* **29**, 614.

[26] WOOD, F. AND GRIFFITHS, T.L. (2007). Particle filtering for nonparametric Bayesian matrix factorization. *Advances in Neural Information Processing Systems.*

[27] WOOD, F., GRIFFITHS, T.L. AND GHAHRAMANI, Z. (2006). A nonparametric Bayesian method for inferring hidden causes. *22nd Conference in Uncertainty in Artificial Intelligence.*

[28] 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**.