

Contributions to Bayesian Nonparametric methods with applications



Mohamed Fadhel Ayed
Somerville College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2020

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. My personal contributions are as outlined in the authorship forms at the end of each chapter. This dissertation is my own work except as specified in the text, acknowledgements, forms and papers.

Mohamed Fadhel Ayed
Trinity 2020

This thesis is dedicated to
my family

Acknowledgements

I want to thank first my supervisor François Caron for his great support and patience. This thesis wouldn't have been possible without his guidance. As any of his students would attest, he is a dependable source of inspiration, and the best supervisor one can have.

Thanks to Xenia Miscouridou, Marco Battiston, Paul Vanetti and Juho Lee; to the Badlads: Claire Huang, Philippe Gagnon, Chris Carmona, Jotun Hein, Judith Rousseau, Giuseppe Di Benedetto, and Jonida Popi; to the Somervillains: Alex Kaltenboeck, Anna Gluck, Tadej Pahor, Grace Mason; and to all the colleagues and friends from the Oxford Statistics Department, who made these experience so rich and enjoyable.

Thanks to M. Cabau and M. Lassoued for putting me on the track that led to this thesis.

I am grateful to the Clarendon Fund for providing me with the opportunity to study at Oxford.

Finally, I would like to thank above all my family and Emma, for their constant support and care, and to whom I owe everything.

Abstract

The Bayesian methodology provides a compelling framework that offers an intuitive yet practical approach to model the world, allowing refinement of the statistician's prior knowledge in light of incoming empirical data. Any unknown quantity is regarded as random, and its updated knowledge, called the posterior distribution, quantifies the remaining uncertainty therein. The choice of the prior distribution is at the core of Bayesian modeling and plays a primary role in the model's performance. Bayesian Nonparametric (BNP) modeling utilizes infinitely many parameters to capture intricate patterns of real-world phenomena and adapts to the complexity of the data.

Bayesian Nonparametrics is, therefore, an appealing tool that has found many successful applications in a large variety of domains such as machine learning, ecology, biology, finance, etc. Our research focus is on modeling and analyzing discrete structures, with a particular interest in the context of clustering and feature allocation. In the first introductory chapter we review the necessary theoretical background and present the critical building blocks for the constructions proposed in the following chapters. The first part of this dissertation is dedicated to modeling complex data with Bayesian Nonparametrics. Our objective is to overcome the limitations of existing approaches and describe models that can capture real-world characteristics. We pay careful attention to power-law properties, which seem to be ubiquitous in real-world applications, and propose models for networks with community structures, clustering, and financial time-series. In the second part, we move beyond modelling purposes and show that Bayesian nonparametric can offer powerful tools for the theoretical analysis of (non-Bayesian) estimation problems. The part is dedicated to the question of rare events and missing mass estimation in the species sampling and feature allocation frameworks. Using carefully chosen Nonparametric priors, we show the nonexistence of universally consistent estimators of the missing mass. We also derive lower bounds on rates of estimation of rare events under appropriate assumptions.

Contents

1	Introduction	1
1.1	Frequentist species sampling and feature allocation models	2
1.2	Bayesian Nonparametric modeling of discrete random structures . . .	6
1.3	Pure Jump Lévy processes	11
1.4	Thesis outline	13
1.4.1	BNP models with power-law properties:	14
1.4.2	Estimation of rare events and missing mass in species sampling and feature allocation models	15
I	Bayesian Nonparametric models with power-law prop- erties	18
2	BNP factor models with completely random measures	20
3	Statistical Models with double power-law behavior	52
4	The Normal-Generalised Gamma-Pareto process	64
II	Estimation of rare events and missing mass in species sampling and feature allocation models	97
5	On consistent and rate optimal estimation of the missing mass in species sampling	99
6	A Good-Turing estimator for feature allocation models	127
7	Consistent estimation of rare masses in feature allocation models	159
8	Discussion	189
8.1	Summary	189
8.2	Future directions	190
	Bibliography	191

1

Introduction

Contents

1.1	Frequentist species sampling and feature allocation models . . .	2
1.2	Bayesian Nonparametric modeling of discrete random structures	6
1.3	Pure Jump Lévy processes	11
1.4	Thesis outline	13
1.4.1	BNP models with power-law properties:	14
1.4.2	Estimation of rare events and missing mass in species sampling and feature allocation models	15

This dissertation follows an integrated format and contains a collection of 6 papers presented in Parts I and II. Each article includes a review of the specific literature, a bibliography, and supplementary material. This first chapter serves as a broad introduction, and the last one develops a discussion on potential future work. Part I is dedicated to modeling complex data with Bayesian Nonparametrics. It is composed of three articles, one published in the *Proceedings of the International Conference on Machine learning (2019)*, one under revision in the journal *Statistics and Computing* and the last one has been submitted to the *Journal of Business and Economic Statistics* and is under review. Part II tackles the problem of rare events and missing mass estimation in the species sampling and feature allocation frameworks. It is also composed of three papers; The first one is published in the

Electronic Journal of Statistics. The second one has been submitted to the *Journal of Machine Learning Research* and is now under revision. The last paper is under review in the journal *Annales de l'Institut Henri Poincaré, Series B*.

1.1 Frequentist species sampling and feature allocation models

Clustering and feature allocation tasks naturally arise in various domains of mathematical modeling, with remarkable outcomes in machine learning, ecology, biology, or physics. In this paradigm, the statistician assumes that the data can be subdivided into groups with common properties. In a *feature allocation model*, the groups can overlap. On the other hand, in a *cluster model*, the groups are disjoint and hence define a partition of the data.

In the following, we denote the data points $X_1, \dots, X_n \in \mathbb{X}$ where $(\mathbb{X}, \mathcal{X})$ is a measurable space. In the context of this dissertation, one can consider that $\mathbb{X} \subset \mathbb{R}^d$, which is equipped with its Borel σ -algebra.

Cluster models:

In a cluster model, we suppose that there are K potential distinct groups. Any group k – called a cluster – is represented by a distinct label $\theta_k \in \Theta$. For a data point i , we denote $a_i \in \{\theta_1, \dots, \theta_K\}$ its cluster allocation. If we further assume that the cluster allocations are independent and identically distributed (iid), then we can define $p_k = \mathbb{P}(c_1 = \theta_k)$, which gives the probability of any data point to belong to cluster k . All clusters information $(p_k, \theta_k)_{k=1:K}$ can be conveniently encoded with a single discrete probability measure π defined by

$$\pi = \sum_{k=1}^K p_k \delta_{\theta_k},$$

where δ_{θ_k} is the Dirac measure at θ_k , which leads to the simple relation $a_i \stackrel{iid}{\sim} \pi$. Typically, the measure π is unknown to the statistician, her task is to estimate it using the observations. Besides, she can rarely observe the cluster allocations directly, since the basic notion of clusters described previously only constitutes a

building block of typically more complex models. In many settings, the cluster labels are uninformative, and we are only interested in the induced partition of the n observations. In such cases, instead of the cluster allocation a_i , we use the cluster index $c_i \in \{1, \dots, K\}$, defined by the relation $a_i = \theta_{c_i}$, which allows to marginalize out the labels if need be.

In the next paragraph, we will detail the instructive example of distribution estimation, which will serve as a motivation for using Bayesian Nonparametrics. But before that, let us consider the simplest setting where we can directly observe the cluster affiliations. We will call such a model a *frequentist species sampling model*. This model was first introduced for populations of animals from different species almost a century ago ([Fisher et al., 1943], [Good, 1953]). And even though the problem formulation is overly simplistic, it is still an active subject of research. For example, a question that has become a classical statistics problem is whether we can consistently estimate M_n , the probability of discovering a new species with the next observation X_{n+1} . This problem is referred to as the missing mass problem. Arguably the most famous estimator of the missing mass M_n is the Good-Turing estimator ([Good, 1953]), defined by $\widehat{M}_n = \frac{K_{n,1}}{n}$, where $K_{n,1}$ is the number of species observed once. Following [Ohannessian and Dahleh, 2012], the second part of this dissertation addresses the problem of the missing mass estimation under the so-called *multiplicative loss*, defined by

$$L(\widehat{M}_n, M_n) = \left| \frac{\widehat{M}_n}{M_n} - 1 \right|,$$

which is adequate for estimating small values.

Let us now consider the instructive application to distribution estimation. Committing to the basic intuition behind clustering, we assume that all observations of a cluster k are sampled from the same distribution denoted $F(\cdot \mid \theta_k)$ and sometimes called the kernel. The process generating the data can, therefore, be seen as a hierarchical procedure:

$$\begin{aligned} a_i &\sim \pi \\ X_i \mid c_i &\sim F(\cdot \mid a_i) \end{aligned}$$

A Mixture Model can equivalently represent the distribution of the data as follows

$$X_i \sim F = \sum_{k=1}^K p_k F(\cdot \mid \theta_k).$$

In this context, the probability measure π is called the *mixing measure*.

The implicit hypothesis is that the distributions $F(\cdot \mid \theta_k)$ take a simple form. The most widely used Mixture model is presumably the *Gaussian Mixture Model* (GMM), where the kernel is Gaussian. For example, in the univariate case, the label of the cluster k is the tuple $\theta_k = (\mu_k, \sigma_k^2)$, the parameters of the Gaussian kernel. These models aim to approximate the true distribution of the data. Even though GMMs may appear too simplistic, they have proven to be practical and hence are widely used. In order to recover the unknown mixing measure π and the cluster allocations c_i , several approaches have been designed. Building on the Expectation-Maximization algorithm ([Dempster et al., 1977]), Frequentist approaches can be used to compute a maximum likelihood estimator. Due to their desirable properties, Bayesian models have also been proposed for GMMs. The mixing measure π is treated as a random variable and endowed with a prior distribution Π . One can recover the posterior distribution of π using Gibbs samplers ([Diebolt and Robert, 1994]).

The drawback of these methods is the assumption that K , the number of potential clusters, is finite. Under such condition, we understand that the GMM will not be able to approximate arbitrarily well the distribution of the data. Instead, we would want K to be infinite, to allow the model to adapt to the increasing knowledge gathered from the observations as n grows. Bayesian Nonparametric methods are precisely designed to deliver on this promise. Further details are given in the next section.

Feature allocation models:

Feature allocation models are very similar to cluster models, with the notable difference that observations may belong to several groups. In this context, we interpret the K groups as features that the data points may display. A feature k is still represented using a label $\theta_k \in \Theta$, but here the feature allocation a_i is

a subset of the set of potential features $\{\theta_1, \dots, \theta_K\}$. A more convenient way to represent the feature allocation is to use a binary vector $c_i = (c_{i1}, \dots, c_{iK})$ with the convention that $c_{ik} = 1$ if the observation displays the feature k . If we further assume that the allocation vectors are iid with independent entries, we recover the so-called *Bernoulli product model*, or *binary independence model*, which is arguably the most popular feature allocation model. It represents the allocation vector as a sequence of independent Bernoulli random variables with unknown success probabilities $(p_1, \dots, p_K)$, satisfying $p_k \in [0, 1]$. Therefore, the features can be represented with a discrete measure:

$$\mu = \sum_{k=1}^K p_k \delta_{\theta_k}.$$

The difference with the measure π of the cluster models is that $\mu(\Theta)$ is not necessarily equal to one. The Bernoulli product model boils down to $c_{ik} | \mu \stackrel{\text{indep}}{\sim} \text{Bernoulli}(\mu(\theta_k))$.

The feature allocation framework often serves as a block upon which more complex models are built, with remarkable applications to numerous fields such as genetics ([Bernardo et al., 2003]) or psychology ([Görür et al., 2006]), for example. Similar to the frequentist species sampling model, the Bernoulli product model, where we assume that we can observe the allocation vectors directly, is still often used in practice. For example, it has been applied extensively in ecology when animals are captured using traps. Each observation is an incidence vector collecting the presence or absence of each species in the traps ([Colwell et al., 2012]). This model still raises many challenging questions such as the missing mass problem, which, transposed to this setting, consists of estimating the expected number of hitherto unseen features that would be observed if we were provided with one additional data point.

In clustering or feature allocation models, the number of groups formed by the data points quantifies the complexity of the models. By allowing K – the number of potential groups – to be infinite, Bayesian nonparametric priors enable the model to adapt its complexity to the data. The number of apparent clusters (or features) can grow unboundedly as the number of observations available increases,

with flexibility on the growth rate of the number of groups (see for example [Lijoi et al., 2010]) or the group sizes (see for instance [Di Benedetto et al., 2020] for a slightly more involved feature allocation model). In the following section, we introduce some fundamental notions of Bayesian Nonparametrics and give two examples of widely used BNP models for clustering and feature allocation, namely the Chinese Restaurant Process (CRP) and the Indian Buffet Process (IBP).

1.2 Bayesian Nonparametric modeling of discrete random structures

BNP is a branch of Statistics and Machine Learning that enables the design of priors for infinite dimensional parameters. It is a rich field of research from which we only review some basic notions. We refer the reader to the books [Ghosal and Van der Vaart, 2017, Hjort et al., 2010, Daley and Vere-Jones, 2007] for more details. Random measures, i.e., measure-valued random objects, are the cornerstone of many BNP constructions; the clustering probability measure π and the feature allocation measure μ described previously – when endowed with a prior – are examples of discrete random measures. In the following, we suppose that $(\mathbb{X}, \mathcal{X})$ is a separable complete metric space equipped with its Borel σ -algebra. In the scope of this dissertation, one can take $\mathbb{X} = \mathbb{R}^d$.

Completely random measures (CRMs, [Kingman, 1967, Daley and Vere-Jones, 2007, Lijoi et al., 2010]) form a subset of the random measures on $(\mathbb{X}, \mathcal{X})$ and are ubiquitous in BNP modeling. They are defined as follows:

Definition 1 *A random measure ξ on $(\mathbb{X}, \mathcal{X})$ is completely random if for any bounded and disjoint subsets $A_1, \dots, A_m \in \mathcal{X}$, the random variables $\{\xi(A_1), \dots, \xi(A_m)\}$ are mutually independent.*

We pay special attention to discrete random measures, which can be written as a countable sum of weighted Dirac measures $\mu(dX) = \sum_k w_k \delta_{X_k}(dX)$, where the w_k and X_k are respectively called the *weights* and *atoms* (or *locations*) of the measure. *Poisson point processes*, which we define hereafter, are a simple but powerful tool to

model discrete random sets. They also allow defining the so-called *Poisson random measures*, which are arguably the most simple non-trivial discrete random measures.

Definition 2 *Let ν be a measure on $(\mathbb{X}, \mathcal{X})$. A Poisson point process on $\mathbb{X}$ is defined as the random countable subset Π of $\mathbb{X}$ such that*

1. *for $A \in \mathcal{X}$, the counting measure defined as $N(A) := \#\{\Pi \cap A\}$ follows a Poisson distribution with rate $\nu(A)$*
2. *for any disjoint subsets $A_1, \dots, A_m \in \mathcal{X}$, the random variables $N(A_1), \dots, N(A_m)$ are independent.*

We call ν the mean measure of the Poisson process Π . The random measure N is called a Poisson random measure.

However, these random measures lack flexibility since their realizations are integer-valued measures. They can be generalized to what we will call *Poisson completely random measures*, which we construct as follows; Let N be a Poisson random measure on $\mathbb{R}_+ \times \mathbb{X}$ with mean measure $\nu(dw, dX)$. Let

$$\mu(dX) = \int_{\mathbb{R}_+} w N(dw, dX).$$

We assume that for any bounded $A \in \mathcal{X}$, $\int_{\mathbb{R}_+} (1 - e^{-w}) \nu(dw, A) < +\infty$. This condition enforces that for any bounded set A , $\mu(A)$ is almost surely finite. More precisely, the number of large weights – larger than any positive epsilon – is finite, while the small weights may be in infinite numbers but need to be almost surely summable. We further assume that $\forall x \in \mathbb{X}$, $\nu(\mathbb{R}_+, \{x\}) = 0$, such that μ has no fixed atoms, i.e., if we take $x \in \mathbb{X}$ fixed, then almost surely $\mu(\{x\}) = 0$. Under these conditions, $\mu = \sum_k w_k \delta_{X_k}$ is a discrete CRM that we call a Poisson CRM (note that this is not a common denomination; we use it here for clarity of exposition). Not only do Poisson CRMs offer a simple way to construct completely random measures, but they are also inevitable. Indeed, CRMs admit a representation

theorem ([Kingman, 1967, Daley and Vere-Jones, 2007]) that states that any CRM ξ can be decomposed into a sum of three independent components:

$$\xi = \xi_d + \xi_f + \mu,$$

where ξ_d is a deterministic atomless measure, ξ_f is a discrete random measure with fixed atoms – only the weights are random – and μ is a Poisson CRM. Therefore, any discrete CRM with no fixed atoms is necessarily a Poisson CRM. In particular, we will focus on homogeneous Poisson CRMs, which mean measure factorizes into $\nu(dw, dX) = \rho(dw)\alpha(dX)$, and that we will denote $\text{CRM}(\rho, \alpha)$. Under the defining conditions of CRMs, ρ is a Lévy measure on $\mathbb{R}_+$. α is called the *base measure* of the CRM, and is often a probability measure. Besides, when $\rho(\mathbb{R}_+) = +\infty$, we say that the measure is infinite activity: the CRM will have an infinite number of atoms within each set A of strictly positive measure under α .

CRMs have been extensively used to design BNP models for clustering and feature allocation. Here we present two of the most successful such models. To be consistent with the first section, we denote the location space $\mathbb{X}$: Θ , and the points X_k : θ_k . Besides, from now on we assume that α is a probability measure on Θ .

Feature allocation models: The Indian Buffet Process (IBP)

Application to the setting of feature allocation models is straightforward. Indeed, we only need the CRM $\mu = \sum_k w_k \delta_{\theta_k}$ to have weights in $[0, 1]$. It is hence enough to choose a Lévy ρ with support on this segment. Given such a random measure, we assume that any observation displays feature k with probability w_k ; in other words, we can describe the model by the following hierarchical construction:

$$\begin{aligned} \mu &\sim \text{CRM}(\rho, \alpha), \\ c_{ik} \mid \mu &\stackrel{\text{indep}}{\sim} \text{Bernoulli}(w_k). \end{aligned}$$

If we further assume that $\rho((0, 1)) = +\infty$, then the number of latent features is infinite. The *Beta process* ([Hjort et al., 1990]), defined by the Lévy measure:

$$\rho_B(dw; \eta, c) = \eta w^{-1} (1 - w)^{c-1} \mathbb{I}_{w \in (0, 1)} dw,$$

with $\eta > 0$ and $c > 0$, is such a measure. The obtained process is called the *Indian Buffet Process* (IBP) ([Ghahramani and Griffiths, 2006, Thibaux and Jordan, 2007]). This model allows for an infinite number of latent features. However, for a finite number n of observations, only $K_n < +\infty$ features will be active (displayed by at least one observation). K_n grows unboundedly as more data points are available, effectively adapting the model's complexity to the amount of observed data. However, this model implicitly imposes a logarithmic growth of K_n . A first generalization called the three-parameter IBP ([Teh and Gorur, 2009]) allows for a power-law growth of $K_n \sim Cn^\sigma$ for $C > 0$ and $\sigma \in (0, 1)$. In [James et al., 2017], the author generalizes, even more, the IBP and shows how to derive the posterior distribution for an arbitrary CRM and likelihoods not limited to Bernoulli distributions.

Cluster models: The Chinese Restaurant Process (CRP)

In a cluster model, we need the random measure to have support on the set of probability measures. We can construct a random probability measure π by normalizing an infinite activity CRM $\mu \sim \text{CRM}(\rho, \alpha)$ ([Regazzini et al., 2003, Lijoi et al., 2007]):

$$\pi = \frac{\mu}{\mu(\Theta)}.$$

We call π a *Normalized completely random measure* (NCRM) and write $\pi \sim \text{NCRM}(\rho, \alpha)$. Using this construction, we do not need the Lévy measure ρ to have support in $[0, 1]$. Arguably the most used CRM is the so called *Generalized Gamma Process* (GGP, [Hougaard, 1986, Hjort et al., 1990]), with Lévy intensity

$$\rho_{\text{GGP}}(dw; \eta, \sigma, \tau) = \frac{\eta}{\Gamma(1 - \sigma)} w^{-1-\sigma} e^{-\tau w} dw,$$

with $\eta > 0$, $\sigma \in (-\infty, 1)$, and $\tau \geq 0$ ($\tau > 0$ if $\sigma \leq 0$)

The GGP class admits three popular special cases: the *Stable Process* ($\tau = 0$ and $\sigma > 0$), the *Gamma Process* ($\sigma = 0$ and $\tau > 0$) and the *Inverse-Gaussian Process* ($\sigma = 1/2$ and $\tau > 0$). The Gamma Process (GP) is particularly renowned for its convenient computational properties. Indeed, for any partition $\{A_1, \dots, A_M\}$ of Θ , the random variables $\mu(A_1), \dots, \mu(A_M)$ are mutually independent Gamma random

variables. $\mu(A_m)$ has shape parameter $\eta\alpha(A_m)$ and inverse scale τ . Therefore, if π is a normalized Gamma Process, then for any partition $\{A_1, \dots, A_M\}$ of Θ , the vector $(\pi(A_1), \dots, \pi(A_M))$ follows a Dirichlet distribution with concentration vector $(\eta\alpha(A_1), \dots, \eta\alpha(A_M))$. Consequently, a normalized Gamma Process is called a *Dirichlet Process* (DP, [Ferguson, 1973]). If we endow π with a DP prior, the predictive rule of the cluster allocations $a_1, a_2, \dots, a_i, \dots$ is called the *Blackwell-MacQueen urn scheme* ([Blackwell et al., 1973]). The random partitions induced by the sequence of cluster indexes $c_1, c_2, \dots, c_i, \dots$, follow a *Chinese Restaurant Process* (CRP), which is arguably the most famous BNP partition model. We can summarize the CRP as the partition model induced by the following hierarchical construction:

$$\begin{aligned} \pi &\sim \text{DP}(\eta, \alpha(\cdot)) \\ a_i &\stackrel{iid}{\sim} \pi \end{aligned}$$

The CRP is convenient considering its computational properties. However, it implicitly yields restrictive asymptotic properties: for instance, the number of clusters K_n grows logarithmically with n ([Korwar et al., 1973]). Besides, the proportion of clusters of a given size s , denoted $p_n(s)$, converges to zero as n grows to infinity ([Gnedin et al., 2007]). Several empirical studies refute this last behavior since the asymptotic proportion of clusters often displays a power-law behavior for small sizes s . As an illustration, we consider the case of BNP cluster models for language modeling. In this setting, the probability measure π represents a vocabulary: the atoms label every word of the dictionary while the weights give their frequency. We then observe n words of a corpus, each cluster represents a unique word and the cluster size equals the number of occurrences of the word. Figure 1 shows the word frequencies of about 300,000 words from the American National Corpus¹ in the log-log scale. We can see that in the small clusters regime, corresponding to the left part, the cluster sizes display a power-law behavior, materialized by the red line. A generalization of the CRP called the *Pitman-Yor Process* (PY, [Pitman, 2003]) can capture such a property. However, the existing

¹<http://www.anc.org/data/anc-second-release/frequency-data/>

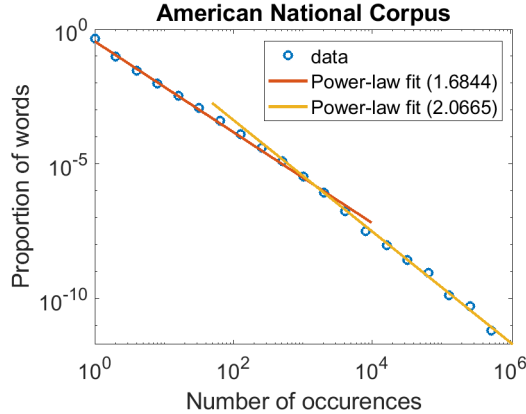


Figure 1.1: Proportion of words with a given number of occurrences for the same dataset (circles) and power-law fit (straight lines)

models do not focus on the the large clusters regime. The available priors either impose an exponential cut off for large s , or a power-law behavior with the same index as in the small clusters regime. In Figure 1.1, we see that the cluster sizes also display a power-law behavior in the large clusters regime, materialized by the yellow line. We also understand that the slope is different from the one shown in the small clusters regime. In Part I, we propose BNP models – that we call double power-law – that capture such phenomena.

1.3 Pure Jump Lévy processes

Beyond clustering and feature allocation models, BNP can be used in the context of functional modeling. *Pure-jump Lévy processes* are a flexible class of stochastic processes that have found a wide range of applications. Finance is probably the primary domain of application, as it is widely accepted that asset prices contain jumps. Such models have been used as building blocks of complex dynamic models of asset or option prices ([Tankov, 2003]). In this section, we define some of Lévy processes’ basic notions. For more details, we refer the reader to [Bertoin, 1996] or [Barndorff-Nielsen et al., 2012], for example. We start by defining the class of Lévy processes:

Definition 3 *A stochastic process $(X_t)_{t \geq 0}$ on $\mathbb{R}$ is called a Lévy process if it satisfies:*

1. $X_0 = 0$ *almost surely*
2. *The increments are independent, i.e. for $t_1 < \dots < t_m$, the increments $X_{t_1}, X_{t_2} - X_{t_1}, \dots, X_{t_m} - X_{t_{m-1}}$ are mutually independent*
3. *The increments are stationary, i.e. the distribution of $X_{t+s} - X_t$ does not depend on t*
4. *The process is stochastically continuous, i.e. $\forall t \geq 0$ and $\epsilon > 0$, $\lim_{s \rightarrow t} \mathbb{P}(|X_s - X_t| > \epsilon) = 0$*

In the context of time-series, we consider the index $t \in \mathbb{R}_+$ to be the time index. Brownian motions or Poisson processes are examples of Lévy processes. This class of stochastic processes is vast, and without further assumptions, a generic Lévy process can be a complex object. Consequently, in practice, we restrict the study to carefully-chosen sub-classes. Consider first that X_t is a non-decreasing Lévy process on $\mathbb{R}$, called a *Subordinator*. The non-decreasing property of the subordinator allows defining a random measure

$$\xi([0, t]) = X_t, \quad \forall t \geq 0.$$

The independence of the increments of X_t implies that the resulting random measure is completely random. Using the representation theorem of CRMs, we deduce that $\xi = \xi_d + \xi_f + \mu$, where ξ_d is a deterministic atomless measure, ξ_f is a discrete random measure with fixed atoms and μ is a Poisson CRM. The stationarity condition on the subordinator further implies that $\xi_f = 0$ and that $\xi_d([0, t]) = at$ with $a \geq 0$ and the mean measure of μ satisfies $\nu(dw, d\theta) = \rho(dw)d\theta$ ([Kingman, 1993]). The reciprocal is also true. This simple characterization allows constructing subordinators flexibly, using, for instance, the priors introduced in the previous section. However, the non-decreasing assumption is too restrictive for many applications. Therefore, we present the richer class of *subordinated Brownian motions*, which are characterized by:

$$X_t = B_{S_t},$$

where B and T are independent, the former is a Brownian motion, and the latter is a subordinator. This class was first introduced in [Clark, 1973] to model the logarithmic asset prices and is still extensively used in finance. The subordinator S acts as a time dilatation; it is referred to as *time change* or *business time*. The rationale for this model is that the variance of X_t should not depend on the calendar time but on an *operational time*. For instance, in a financial context, the operational time could capture the frequency of the trades; the variance of a price within a given minute should be higher if more trades happen.

In this dissertation, we will focus on the so-called *Pure-jump Lévy processes*, which we define henceforth. Let $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ be a right continuous with left limits (càdlàg) – deterministic – function. Denote $\Delta f(t) = f(t_+) - f(t_-)$, the jump size from the left-limit to the right limit of f . Let $D = \{t \in \mathbb{R}_+ : \Delta f(t) \neq 0\}$ be the discontinuity set of f . This function is called a *pure jump function* if it satisfies

$$f(t) = \sum_{s \in D \cup [0, t]} \Delta f(s).$$

A Lévy process is said to be a *Pure Jump Lévy process* if its realizations are almost surely pure jump functions. A Pure Jump subordinator has no drift. And a subordinated Brownian motion is a Pure Jump Lévy process if its subordinator is. Many celebrated models can be constructed this way; the *Variance-Gamma* ([Madan and Seneta, 1990]), *Normal-Inverse Gaussian* ([Barndorff-Nielsen et al., 2012]) and *Normal-Stable* ([Cont et al., 1997]) are obtained by taking the corresponding subordinators, which are all particular cases of the Generalized Gamma subordinator.

1.4 Thesis outline

This dissertation follows an integrated thesis format. It is organized into two main parts, formed of three publications each. In Part I, we propose BNP models for complex data. Part II tackles the problem of rare events and missing mass estimation in the species sampling and feature allocation frameworks. The papers are in their complete form, including the main article, the bibliography, and the supplementary material. I am the first author of all the documents, and I have

contributed to the theoretical, methodological, and experimental aspects. Following each publication is a signed statement of authorship detailing my contributions.

1.4.1 BNP models with power-law properties:

Part I is dedicated to modeling complex data with BNP. Our objective is to overcome the limitations of the current literature and describe models that can capture real-world characteristics. We pay careful attention to power-law properties, which seem to be ubiquitous in real-world applications.

In Chapter 2, we present a BNP factorization model for modeling dense network data with an unknown number of overlapping communities. The construction is based on completely random measures. It is a generalization of the Infinite edge-partition model proposed by [Zhou, 2015] that allows the number of communities to increase with the number of nodes at a specified logarithmic or polynomial rate, or be bounded, where the original model imposed a logarithmic growth. We develop asymptotics for the number and size of the network communities. We further derive a Markov chain Monte Carlo algorithm for targeting the exact posterior distribution for this model, where the original paper proposed an approximate inference technique. The usefulness of the approach is illustrated on various real data.

This paper is under major revision in the journal *Statistics and Computing* and is referenced as follows:

- **Ayed, F.**, and Caron, F. (2019). Nonnegative Bayesian nonparametric factor models with completely random measures for community detection. *arXiv preprint arXiv:1902.10693*.

In Chapters 3 and 4, we introduce a novel class of completely random measures which are doubly regularly-varying. These models address the growing empirical evidence suggesting that some datasets exhibit a two-regime power-law behavior: one regime for small frequencies, and a second regime, with a different exponent, for high frequencies. In Chapter 3, we demonstrate the performance of these CRMs in the context of cluster models with applications to text and network data. We

derive efficient Markov chain Monte Carlo algorithms and show that the proposed models provide a better fit than the Pitman-Yor process on various datasets.

This paper is published in the *Proceedings of the International Conference on Machine learning (2019)*, and is referenced as follows:

- **Ayed, F.**, Lee, J., and Caron, F. (2019). Beyond the Chinese Restaurant and Pitman-Yor processes: Statistical Models with double power-law behavior. *arXiv preprint arXiv:1902.04714*.

In Chapter 4, we construct a novel family of self-decomposable Lévy processes. One can control the tail behavior and the jump activity of the process separately, via two different parameters. We use this novel process in Lévy-based stochastic volatility models to predict the returns of stock market data. We show that the proposed class of models leads to superior predictive performances compared to classical alternatives.

This paper has been submitted to the *Journal of Business and Economic Statistics* and is under review. It is referenced as follows:

- **Ayed, F.**, Lee, J., and Caron, F. (2020). The Normal-Generalised Gamma-Pareto process: A novel pure-jump Lévy process with flexible tail and jump-activity properties. *arXiv preprint arXiv:2006.10968*.

1.4.2 Estimation of rare events and missing mass in species sampling and feature allocation models

The part is dedicated to the question of rare events and missing mass estimation in the species sampling and feature allocation frameworks. It is composed of three papers, two of which concern the feature allocation model, the other concerns the species sampling model.

Chapter 5 addresses the missing mass problem in its most classical context, i.e., in the species sampling context. This problem has been extensively studied. In particular, using a multiplicative loss, recent results have shown:

1. The impossibility of estimating the missing mass without imposing further assumptions on type's proportions ([Mossel and Ohannessian, 2019]).
2. The convergence rate for the Good-Turing estimator if the true proportions decay as a power-law ([Grabchak and Zhang, 2017]).

In this paper, we introduce an alternative and shorter proof of the impossibility of estimation of the missing mass using a BNP construction. The proof idea is of independent interest and will be adapted in Chapter 7 to show the impossibility of estimation of the missing mass in the feature allocation model, which is a new result. Based on BNP techniques, we further prove that, under the power-law assumption, the rate of convergence of the Good-Turing estimator is optimal.

This work has been submitted to the journal *Annales de l'Institut Henri Poincaré, Series B* and is under review. It is referenced as follows:

- **Ayed, F.**, Battiston, M., Camerlenghi, F., and Favaro, S. (2018). On consistent estimation of the missing mass. *arXiv preprint arXiv:1806.09712*.

In Chapters 6 and 7, we focus on the context of feature allocation models. In Chapter 6, we propose a nonparametric estimator of the missing mass with the same analytic form of the Good-Turing estimator for species sampling models. We show that the estimator admits a natural interpretation both as a jackknife estimator and as a nonparametric empirical Bayes estimator. Furthermore, we give provable guarantees for the estimator's performance for the L_2 norm in terms of minimax rate optimality and non-asymptotic confidence intervals. We provide a connection between our estimator and the Good-Turing estimator for species sampling.

The paper has been published in the *Electronic Journal of Statistics* and is referenced as follows:

- **Ayed, F.**, Battiston, M., Camerlenghi, F., and Favaro, S. (2019). A Good-Turing estimator for feature allocation models. *Electronic Journal of Statistics*, 13(2), 3775-3804.

In Chapter 7, we consider the problem of estimating the missing mass and rare events in a feature allocation model. The difference with the previous chapter is that we now perform the analysis with the more appropriate multiplicative loss. Using a BNP approach, we first show that there are no universally consistent estimators of the missing mass in the feature allocation context. Then, we show that the Good-Turing estimator for the feature allocation model introduced in the previous chapter is strongly consistent, in the multiplicative sense, under the assumption power-law assumption on the proportions.

The paper has been submitted to the *Journal of Machine Learning Research*, and is now under major revision. It can be referenced as follows:

- **Ayed, F.**, Battiston, M., Camerlenghi, F., and Favaro, S. (2019). Consistent estimation of the missing mass for feature models. *arXiv preprint arXiv:1902.10530*.

Aside note

During my DPhil, I also worked on privacy-related problems, which resulted in two papers, one published in the journal *Statistics and Computing*, the other has been accepted to the proceedings of the *Privacy in Statistical Datasets conference*.

- **Ayed, F.**, Battiston, M., and Camerlenghi, F. (2020). An Information Theoretic approach to Post Randomization Methods under Differential Privacy. *Statistics and Computing*.
- **Ayed, F.**, Di Benedetto, G. and Battiston, M. (2020). A Bayesian Nonparametric approach to Differentially Private data. *Privacy in Statistical Datasets conference 2020*, to appear.

Part I

Bayesian Nonparametric models with power-law properties

Abstract

This part is composed of three independent papers; The first one has been submitted to the journal *Statistics and Computing*, and is under revision. The second one has been accepted for publication in the Proceedings of the *International Conference on Machine Learning 2019*. The last paper has been submitted to the *Journal of Business and Economic Statistics* and is under review.

Table of Contents

2	BNP factor models with completely random measures	20
3	Statistical Models with double power-law behavior	52
4	The Normal-Generalised Gamma-Pareto process	64

2

BNP factor models with completely random measures

Nonnegative Bayesian nonparametric factor models with completely random measures

Fadhel Ayed · François Caron

Received: date / Accepted: date

Abstract We present a Bayesian nonparametric Poisson factorization model for modeling dense network data with an unknown and potentially growing number of overlapping communities. The construction is based on completely random measures and allows the number of communities to either increase with the number of nodes at a specified logarithmic or polynomial rate, or be bounded. We develop asymptotics for the number and size of the communities of the network and derive a Markov chain Monte Carlo algorithm for targeting the exact posterior distribution for this model. The usefulness of the approach is illustrated on various real networks.

Keywords Poisson factorization · Generalized gamma process · Community detection

1 Introduction

Non-negative matrix factorization (NMF) methods (Lee and Seung, 2001; Paatero and Tapper, 1994) aim to find a latent representation of a positive $n \times m$ matrix A as a sum of K non-negative factors. For integer-valued data, Poisson factorization models (Dunson and Herring, 2005) offer a flexible probabilistic framework for non-negative matrix factorization, and have found wide applicability in signal processing (Cemgil, 2009; Virtanen et al., 2008) or recommender systems (Gopalan et al., 2015; Ma et al., 2011). In this paper, we focus on the application to network analysis, where $m = n$ and the $n \times n$ count matrix A , the adjacency matrix, represents the number of directed or undirected interactions between n individuals; the latent factors may be interpreted as latent and potentially overlapping communities (Ball et al., 2011), such as sport team members or other social activities circles. We also consider binary data where the matrix represents the existence or absence of a directed or undirected link between individuals. The estimated latent factors can be used for the prediction of missing links/interactions, or for interpretation of the uncovered latent community structure.

Poisson factorization approaches require the user to set the number K of latent factors, which is typically assumed to be independent of the sample size n . To address this problem, Zhou et al. (2012), Gopalan et al. (2014) and Zhou (2015) proposed Bayesian nonparametric approaches that allow the number of latent factors to be estimated from the data, and to grow unboundedly with the size n of the matrix. In particular, Gopalan et al. (2014) and Zhou

Fadhel Ayed
Department of Statistics, University of Oxford
24-29 St Giles', Oxford OX1 3LB, UK.
E-mail: fadhel.ayed@gmail.com

François Caron
Department of Statistics, University of Oxford
24-29 St Giles', Oxford OX1 3LB, UK.
E-mail: caron@stats.ox.ac.uk

(2015), considered a Poisson factorization model

$$A_{ij} \sim \text{Poisson} \left(\sum_{k=1}^{+\infty} r_k v_{ik} v_{jk} \right), \quad 1 \leq i, j \leq n \quad (1)$$

where the positive weights $(r_k)_{k \geq 1}$ represent the importance of community k , and $v_{ik} > 0$ represents the level of affiliation of individual i to community k . Gopalan et al. (2014) and Zhou (2015), extending work from Titsias (2008), assume that the weights (r_k) are the jumps of a gamma process, ensuring the sum in equation (1) is almost surely finite. Using properties of Poisson random variables, the model (1) can be equivalently represented as

$$A_{ij} = \sum_{k=1}^{+\infty} Z_{ijk} \quad (2)$$

$$Z_{ijk} \sim \text{Poisson}(r_k v_{ik} v_{jk}), \quad k = 1, 2, \dots \quad (3)$$

for $1 \leq i, j \leq n$. The latent count variables Z_{ijk} may be interpreted as the number of latent interactions between two individuals i and j via community k , the overall number A_{ij} of interactions being the sum of those community interactions. For example, two members of the same company who also play sport together may meet five times at the company, and twice at the sport center, resulting in seven interactions overall. The overall number

$$K_n = \sum_{k=1}^{+\infty} \mathbb{1}_{\sum_{1 \leq i, j \leq n} Z_{ijk} > 0} \quad (4)$$

of communities k that generated at least one interaction between the n individuals is termed the number of *active communities*. For the gamma process Poisson factor model (Zhou, 2015), the number of active communities K_n grows logarithmically with the number n of individuals. The logarithmic growth assumption may be too restrictive. For example, the number of active communities may actually be unknown but bounded above; alternatively, it may increase at a rate faster or slower than logarithmic.

In this paper, we consider generalizations of the gamma process Poisson factorization model, using completely random measures (CRM) (Kingman, 1967). CRMs offer a flexible and tractable modeling framework (Lijoi and Prünster, 2010). The proposed models fit in the class of multivariate generalized Indian Buffet process priors recently developed by James (2017) and are also related to compound completely random measures (Griffin and Leisen, 2017). We consider that (r_k) are the points of Poisson point process with mean measure ρ . Depending on the properties of this measure, the number of active communities K_n is either (i) bounded, with a random upper bound, (ii) unbounded and grows sub-polynomially (e.g. $\log n$ or $\log \log n$) or (iii) unbounded and grows as $n^{2\sigma}$, for some $\sigma \in (0, 1)$. For the implementation, we focus in particular on the generalized gamma process (Brix, 1999) where a single parameter flexibly controls all three behaviors.

The article is organized as follows. In Section 2, we describe the statistical model for count and binary matrices. The asymptotic properties of the model are derived in Section 3. In particular, we relate the asymptotic growth of the number of active features to the regular variation properties of the measure ρ . In Section 4 we derive a Markov chain Monte Carlo algorithm for posterior inference that does not require any approximation to the original model. In Section 5 we consider applications of our approach to overlapping community detection and link detection in networks, considering real network data with up to tens of thousands of nodes.

2 Statistical model for count and binary data

2.1 General construction

We present here the model for directed count or binary observations, but the model can be straightforwardly adapted to undirected interactions. Let $(r_k)_{k=1,2,\dots}$, be the points of

a Poisson point process with σ -finite mean measure ρ on $(0, +\infty)$, and assume that v_{ik} , $i = 1, \dots, n$, $k \geq 1$, are independent and identically distributed from some probability distribution F on $\mathbb{R}_+ = [0, +\infty)$. The variable v_{ik} can be interpreted as the level of affiliation of an individual i to community k , and r_k to the importance of that community.

For count data (A_{ij}) , where A_{ij} denotes the number of directed interactions from node i to node j , we consider the Poisson factorization model

$$A_{ij} \mid (r_k, v_{ik}) \sim \text{Poisson} \left(\sum_{k=1}^{+\infty} r_k v_{ik} v_{jk} \right), \quad 1 \leq i, j \leq n. \quad (5)$$

Denoting $\Lambda_{ij} = \sum_{k=1}^{+\infty} r_k v_{ik} v_{jk}$ the Poisson rate for A_{ij} , the $n \times n$ rate matrix $\Lambda^{(n)} = (\Lambda_{ij})_{1 \leq i, j \leq n}$ admits the following factorization as an infinite sum of rank-1 matrices

$$\Lambda^{(n)} = \sum_{k=1}^{+\infty} r_k v_{1:n,k} v_{1:n,k}^\top$$

where $v_{1:n,k} = (v_{1k}, \dots, v_{nk})^\top$. For the model to be well specified, the sum in the right-handside of Equation (5) needs to be almost surely finite. A necessary and sufficient condition is

$$\iint (1 - e^{-rv^2}) \rho(dr) F(dv) < +\infty \quad \text{and} \quad \iint (1 - e^{-rv_1 v_2}) \rho(dr) F(dv_1) F(dv_2) < +\infty. \quad (6)$$

A sufficient set of conditions¹, which we will assume to hold in the rest of this article, is that ρ is a Lévy measure and F has finite second moment, that is

$$\int_0^{+\infty} (1 - e^{-r}) \rho(dr) < +\infty \quad \text{and} \quad (A1)$$

$$\int_0^{+\infty} v^2 F(dv) < +\infty. \quad (A2)$$

In this case, denoting δ_v the Dirac measure at vector v , the community affiliations and weights for n nodes can be conveniently represented by a completely random measure

$$G = \sum_{k \geq 1} r_k \delta_{v_{1:n,k}} \quad (7)$$

on $\mathbb{R}_+^n$ with mean measure $\rho(dr) F^{\otimes n}(dv_1, \dots, dv_n)$ where $F^{\otimes n}$ denotes the n th product measure of F ; see Kingman (1967) and Lijoi and Prünster (2010) for background on CRMs and their applications. If the Lévy measure is finite, that is, if

$$\int_0^{+\infty} \rho(dr) < +\infty$$

then the number of points (r_k) , and therefore the number of communities, is almost surely finite. Otherwise, when $\int \rho(dr) = +\infty$, the number of communities is infinite.

When we have binary observation (Y_{ij}) , we treat the count matrix (A_{ij}) as a latent variable, and consider that $Y_{ij} = \mathbb{1}_{A_{ij} > 0}$ as in (Caron and Fox, 2017; Zhou, 2015), where $\mathbb{1}$ is the indicator function. Integrating out (A_{ij}) , this leads to the following model for binary observations

$$Y_{ij} \mid (r_k, v_{ik}) \sim \text{Ber} \left(1 - \exp \left[- \sum_{k=1}^{+\infty} r_k v_{ik} v_{jk} \right] \right), \quad 1 \leq i, j \leq n. \quad (8)$$

¹ The sufficientness follows from the bound (34) given in Appendix.

2.2 Specific model

In the inference and experimental part, we use the following choice for the ρ and F . The Lévy measure ρ is taken to be that of a generalized gamma process (GGP, see Hougaard (1986), Brix (1999), James (2002), Pitman (2003))

$$\rho(dr) = \frac{\kappa}{\Gamma(1 - \sigma_0)} r^{-1-\sigma} e^{-\tau r} dr \quad (9)$$

where $\sigma_0 \in (-\infty, 1)$, $\kappa > 0$ and $\tau > 0$. When $\sigma_0 = 0$, we obtain a gamma process, and the model corresponds to that of Zhou (2015). When $\sigma_0 < 0$, the Lévy measure is finite, while when $\sigma_0 \geq 0$, the Lévy measure is infinite.

Concerning the affiliations, we will assume that F is a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. That is, the probability density function (pdf) f is given by

$$f(v) = \frac{\beta^\alpha}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v}$$

where Γ denotes the usual gamma function. The hyperparameters $(\kappa, \sigma_0, \tau, \alpha, \beta)$ and $(\kappa' = \kappa/\beta^{2\sigma_0}, \sigma_0, \tau' = \tau\beta^2, \alpha, 1)$ induce the same distribution for the latent factors (Λ_{ij}) . In order to guarantee the identifiability of the hyperparameters, we therefore set $\beta = 1$.

2.3 Related work

Several network models building on latent factors have been proposed in the last years and have proven to be very useful tools (Airoldi et al., 2008; Durante and Dunson, 2014; Hoff, 2009; Hoff et al., 2002). In general these models differ from Poisson factor models since they use a different likelihood for the connections. However, they share a similar approach: every node i is embedded in $\mathbb{R}_+^K$ (where K is the number of latent factors or communities), resulting in a latent representation X_i quantifying the affiliation of node i to each latent factor. Then the probability of an edge (i, j) is function of the similarity between X_i and X_j .

The model introduced in this section can be seen from different perspectives that nicely connect it to the existing literature. First, the model can be seen as obtained from a functional of a CRM. Recall the definition of the CRM G in Eq. (7). Define the $n \times n$ matrix $\Lambda^{(n)}$ as the following functional of G

$$\Lambda^{(n)} = \int_{(0, +\infty)^n} h(u) G(du) = \sum_{k \geq 1} r_k v_{1:n, k} v_{1:n, k}^\top$$

where $h(u) = uu^\top$. Alternatively, this can be interpreted in the framework of compound completely random measures (Griffin and Leisen, 2017). For each $1 \leq i, j \leq n$, denote $G_{ij} = \sum_{k \geq 1} r_k v_{ik} v_{jk} \delta_{\zeta_k}$ where ζ_k are some community locations in some domain Θ , iid from some distribution H , irrelevant here. Then $(G_{ij})_{1 \leq i, j \leq n}$ are compound CRMs on Θ and $\Lambda_{ij} = G_{ij}(\Theta)$. In the same vein, the model can also be interpreted as an instance of the class of Generalized Indian Buffet Processes introduced by James (2017), where the Bernoulli likelihood (of the classical IBP) is replaced by any likelihood as long as the observation can take the value 0 with strictly positive probability. More precisely, if we denote $Z_k^{(n)}$ the $n \times n$ matrix with entries Z_{ijk} , then the matrix-valued process $\sum_{k \geq 1} Z_k^{(n)} \delta_{\zeta_k}$ is a draw from a generalized multivariate Indian buffet process.

Finally, as mentioned in the introduction, the model admits as a special case the Poisson factorization based on the gamma process of Zhou (2015).

3 Asymptotic Properties

In this section we study the asymptotic properties of the proposed class of models, and in particular the growth rate of the number of active communities as the sample size n grows, and the asymptotic proportion of communities of a given size. For a given sequence $(r_k)_{k \geq 1}$ and

$(v_{ik})_{i \geq 1, k \geq 1}$, denote $A_{ij}^{(n)}$ and $Z_{ijk}^{(n)}$ where $n \geq 1$, $1 \leq i, j \leq n$, $k \geq 1$ respectively the number of directed interactions and the number of community directed interactions distributed from Equations (2) and (3). We consider two different asymptotic settings

- Unconstrained setting. This setting is more general, and we only assume that $A_{ij}^{(n)}$ and $Z_{ijk}^{(n)}$ are marginally sampled from Equations (2) and (3).
- Constrained setting. For any $1 \leq m \leq n$, and $1 \leq i, j \leq m$, $A_{ij}^{(n)} = A_{ij}^{(m)}$. In this setting, we suppose that the connections between the already observed nodes remain unchanged. It is equivalent to assuming that there is an infinite but fixed graph and $A^{(n)}$ represents the connections between the n first nodes of that graph.

All the results of this section, otherwise stated, hold for the unconstrained setting. We indicate when a stronger result holds in the constrained setting. All proofs are given in Appendix A.

3.1 General model

Let $d_k^{(n)}$ be the degree of the community/feature k , corresponding to the number of interactions amongst n individuals due to community k , and defined as

$$d_k^{(n)} = \sum_{1 \leq i, j \leq n} Z_{ijk}^{(n)}. \quad (10)$$

A community is active if $d_k^{(n)} \geq 1$. The number of active communities is therefore defined as

$$K_n = \sum_{k=1}^{+\infty} \mathbb{1}_{d_k^{(n)} \geq 1} \quad (11)$$

Denote $K_{n,j}$ the number of communities with degree $j \geq 1$

$$K_{n,j} = \sum_{k=1}^{+\infty} \mathbb{1}_{d_k^{(n)} = j}$$

Note that under the constrained setting, $d_k^{(n)}$, K_n and $\sum_{\ell \geq j} K_{n,\ell}$ are all almost surely non-decreasing with the sample size n , whereas this is not necessarily the case for the unconstrained setting.

Proposition 1 *Under Assumptions (A1) and (A2), the number of active communities K_n is a Poisson random variable with mean*

$$\Psi(n) = \int \left(1 - e^{-r(\sum_{i=1}^n v_i)^2}\right) \left[\prod_{i=1}^n F(dv_i) \right] \rho(dr) < +\infty. \quad (12)$$

The number $K_{n,j}$ of communities with degree j is also Poisson distributed, with mean

$$\Psi_j(n) = \frac{1}{j!} \int r^j \left(\sum_{i=1}^n v_i \right)^{2j} e^{-r(\sum_{i=1}^n v_i)^2} \left[\prod_{i=1}^n F(dv_i) \right] \rho(dr). \quad (13)$$

Finally, for $j \geq 1$, $\sum_{k \geq j} K_{n,k}$, the number of communities with degree at least j , is also Poisson distributed with mean $\sum_{k \geq j} \Psi_k(n)$.

In the rest of the section we relate the asymptotic behavior of quantities of interest to the properties of the mean measure ρ . Let consider the tail Lévy intensity defined as

$$\forall x > 0, \bar{\rho}(x) = \int_x^{+\infty} \rho(dr).$$

We assume that $\bar{\rho}$ is a regularly varying function at 0, that is

$$\bar{\rho}(x) \asymp x^{-\sigma} \ell(1/x) \text{ as } x \rightarrow 0 \quad (\text{A4})$$

where $\sigma \in [0, 1)$ and ℓ is a slowly varying function verifying $\lim_{t \rightarrow +\infty} \ell(at)/\ell(t) = 1$ for all $a > 0$. Besides we write $a(x) \asymp b(x)$ if $\lim a(x)/b(x) = 1$. Examples of slowly varying functions include functions converging to a constant, $\log^a t$ for any t , $\log \log t$, etc. Note that the CRM is finite activity if and only if $\sigma = 0$ and $\ell(t) \rightarrow C < +\infty$.

Now, let us consider the asymptotic behavior of the number of active communities K_n .

Proposition 2 *Let K_n be the number of active communities. Then for $0 \leq \sigma < 1$,*

$$\mathbb{E}[K_n] \asymp \Gamma(1 - \sigma) m_f^{2\sigma} n^{2\sigma} \ell(n^2) \quad (14)$$

as n tends to infinity, where $m_f = \int vF(dv)$. Additionally, for $0 < \sigma < 1$,

$$K_n \asymp \mathbb{E}[K_n] \text{ a.s.} \quad (15)$$

If we further assume that the sequence $(K_n)_{n \geq 1}$ is almost surely non-decreasing (as in the constrained setting), then (15) holds for $\sigma = 0$ and $\ell(t) \rightarrow +\infty$ as well. In the finite activity case, that is $\sigma = 0$ and $\ell(t) \rightarrow \bar{\rho}(0) = \int_0^{+\infty} \rho(dr) < \infty$, we have

$$K_n \rightarrow K_\infty$$

as n tends to infinity, where K_∞ is a Poisson random variable with mean $\bar{\rho}(0)$. The above convergence holds in distribution for the unconstrained setting and almost surely for the constrained setting.

Proposition 3 *Let $K_{n,j}$ be the number of communities of degree j . Then for $0 < \sigma < 1$ and any $j \geq 1$,*

$$K_{n,j} \asymp \frac{\sigma \Gamma(j - \sigma)}{j!} m_f^{2\sigma} n^{2\sigma} \ell(n^2) \text{ a.s.} \quad (16)$$

as n tends to infinity. Therefore,

$$\frac{K_{n,j}}{K_n} \rightarrow \frac{\sigma \Gamma(j - \sigma)}{\Gamma(1 - \sigma) j!} \text{ a.s.} \quad (17)$$

as n tends to infinity. This corresponds to a power-law behavior as

$$\frac{\sigma \Gamma(j - \sigma)}{\Gamma(1 - \sigma) j!} \asymp \frac{\sigma}{j^{\sigma+1}}$$

for large j . If we further assume that for all $k \geq 1$, $\left(\sum_{j \geq k} K_{n,j} \right)_{n \geq 1}$ is non-decreasing (constrained setting), then (17) holds also for $\sigma = 0$ and $\ell(t) \rightarrow +\infty$.

It has been observed empirically that in many networks, the distribution of the sizes of the communities displays a power-law behavior $f_S(s) \sim s^{-1-\sigma}$ where f_S is the distribution of community sizes and $\sigma > 0$ (see for example Arenas et al. (2004); Clauset et al. (2004); Radicchi et al. (2004); Stegehuis et al. (2016)). As stated in Proposition 17, this property cannot be captured in the framework of Zhou (2015) for example where $\sigma = 0$ is constant. These empirical observations seem to indicate that models with flexible σ are needed.

Finally, let $c^{(n)}(k, k')$ denote the cosine between the corresponding affiliation vectors

$$c^{(n)}(k, k') = \frac{\sum_{i=1}^n v_{ik} v_{ik'}}{\sqrt{\sum_i v_{ik}^2} \sqrt{\sum_i v_{ik'}^2}}.$$

This coefficient gives a measure of the overlap between two communities k and k' . By the law of large numbers, for any $k \neq k'$,

$$c^{(n)}(k, k') \asymp \frac{(\int vF(dv))^2}{\int v^2 F(dv)} \text{ a.s. as } n \rightarrow +\infty.$$

3.2 Specific case of the GGP

In the case of the GGP, we have

$$\bar{\rho}(x) = \frac{\kappa \tau^{\sigma_0} \Gamma(-\sigma_0, \tau x)}{\Gamma(1 - \sigma_0)} \asymp \begin{cases} -\frac{\kappa \tau^{\sigma_0}}{\sigma_0} & \text{if } \sigma_0 < 0 \\ \kappa \log(1/x) & \text{if } \sigma_0 = 0 \\ \frac{\kappa x^{-\sigma_0}}{\sigma_0 \Gamma(1 - \sigma_0)} & \text{if } \sigma_0 > 0 \end{cases}$$

as x tends to 0, where $\Gamma(a, x)$ is the incomplete gamma function. Note that $\bar{\rho}(x)$ is of the form $x^{-\sigma} \ell(1/x)$ where $\sigma = \max(0, \sigma_0)$ and

$$\ell(t) = \begin{cases} -\frac{\kappa \tau^{\sigma_0}}{\sigma_0} & \text{if } \sigma_0 < 0 \\ \kappa \log(t) & \text{if } \sigma_0 = 0 \\ \frac{\kappa}{\sigma_0 \Gamma(1 - \sigma_0)} & \text{if } \sigma_0 > 0 \end{cases}$$

is a slowly varying function at infinity. The results of the previous subsection therefore apply. For simplicity, we state the results for the constrained setting. We have, almost surely as $n \rightarrow +\infty$

$$K_n \asymp \begin{cases} K_\infty & \text{if } \sigma_0 < 0 \\ 2\kappa \log(n) & \text{if } \sigma_0 = 0 \\ \kappa \alpha^{2\sigma_0} n^{2\sigma_0} / \sigma_0 & \text{if } \sigma_0 > 0 \end{cases}$$

where $K_\infty \sim \text{Poisson}(-\kappa \tau^{\sigma_0} / \sigma_0)$. Additionally, for $\sigma \geq 0$,

$$\frac{K_{n,j}}{K_n} \rightarrow \frac{\sigma_0 \Gamma(j - \sigma_0)}{\Gamma(1 - \sigma_0) j!}$$

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

$$c^{(n)}(k, k') \rightarrow \frac{\alpha}{\alpha + 1}.$$

Therefore, σ_0 governs the asymptotic behavior of the number of active communities. K_n is bounded with a random upper bound ($\sigma_0 < 0$), increases logarithmically ($\sigma_0 = 0$) or polynomially ($\sigma_0 > 0$). In the polynomial case, σ_0 also controls the power-law exponent of the proportion of communities of a given size. The parameter κ is an overall linear scaling parameter. Finally, the parameter α governs the amount of overlapping between two communities.

4 Simulation, posterior characterization and inference

In this section we describe the marginal distribution and conditional characterization of the model. Building on these, we derive an exact sampler for simulating from the model, and a Markov chain Monte Carlo algorithm to approximate the posterior distribution. Importantly, the sampler targets the distribution of interest and does not require any truncation or approximation. For simplicity of exposition, we assume that ρ and F are absolutely continuous with respect to the Lebesgue measure, with $\rho(dr) = \rho(r)dr$ and $F(dx) = f(x)dx$.

4.1 Marginal distribution and simulation

For a fixed n , recall that K_n denotes the number of active communities. Let $((\tilde{r}_1, \tilde{v}_{1:n,1}), \dots, (\tilde{r}_{K_n}, \tilde{v}_{1:n,K_n}))$ be the subsequence of $(r_k, v_{1:n,k})$ such that community k is active, meaning that $\sum_{1 \leq i, j \leq n} Z_{ijk} \geq 1$, arranged in random order. Let $\tilde{Z}_{ijk}$ be the number of community interactions corresponding to the active community $(\tilde{r}_k, \tilde{v}_{1:n,k})$. Note that

$$A_{ij} = \sum_{k=1}^{K_n} \tilde{Z}_{ijk}. \quad (18)$$

Let $\tilde{Z}_k = (\tilde{Z}_{ijk})_{1 \leq i, j \leq n}$. Using Proposition 5.2 of James (2017), we obtain the following lemma.

Lemma 1 (Marginal distribution) *The joint distribution of $(K_n, (\tilde{r}_{1:K_n}, \tilde{v}_{1:n,1:K_n}), (\tilde{Z}_k)_{k=1,\dots,K_n})$ is given by*

$$K_n \sim \text{Poisson}(\Psi(n)) \quad (19)$$

where $\Psi(n)$ is defined in Eq.(12), and

$$p((\tilde{r}_{1:K_n}, \tilde{v}_{1:n,1:K_n}) | K_n) = \prod_{k=1}^{K_n} p(\tilde{r}_k, \tilde{v}_{1:n,k} | K_n)$$

where

$$p(\tilde{r}_k, \tilde{v}_{1:n,k} | K_n) \propto (1 - e^{-\tilde{r}_k(\sum_{i=1}^n \tilde{v}_{ik})^2}) \rho(\tilde{r}_k) \prod_{i=1}^n f(\tilde{v}_{ik}). \quad (20)$$

Finally, for each $k = 1, \dots, K_n$,

$$\tilde{Z}_k | (\tilde{r}_{1:K_n}, \tilde{v}_{1:n,1:K_n}) \sim \text{tPoisson}(\tilde{r}_k \tilde{v}_{1:n,k} \tilde{v}_{1:n,k}^\top) \quad (21)$$

where $\text{tPoisson}(\Lambda)$ denotes the distribution of a integer-valued matrix with Poisson entries with mean values Λ_{ij} , conditionally on the sum of the entries being strictly positive. This has probability mass function

$$p(A) = \begin{cases} (1 - e^{-\sum_{i,j} \Lambda_{ij}})^{-1} \prod_{1 \leq i,j \leq n} \frac{\Lambda_{ij}^{A_{ij}} e^{-\Lambda_{ij}}}{A_{ij}!} & \text{if } \sum_{i,j} \Lambda_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

The model has an infinite number of parameters, but Lemma 1 allows us to derive an algorithm to exactly sample from it, by successively simulating $K_n, (\tilde{r}_{1:K_n}, \tilde{v}_{1:n,1:K_n}), (\tilde{Z}_k)_{k=1,\dots,K_n}$ and A using Equations (19), (20), (21) and (18).

Sampling from the conditional distribution (21) can be done efficiently by first sampling the number of multiedges $\sum_{i,j} \tilde{Z}_{i,j,k}$ from a truncated Poisson with mean $\tilde{r}_k(\sum_i \tilde{v}_{i,k})^2$, then sampling iid the end nodes of the edges proportionally to the affiliation vector. Simulating from the conditional distribution (20) can be more challenging since it requires sampling a $n+1$ dimensional vector. However, if we suppose that the affiliations are Gamma distributed, the problem reduces to sampling $(\tilde{r}_k, \sum_i \tilde{v}_{i,k})$, which is a two dimensional vector, and independently sample the normalized affiliations from a Dirichlet distribution. Indeed, if the affiliations are Gamma distributed, we consider the following change of variable.

$$\tilde{\varsigma}_k = \sum_{i=1}^n \tilde{v}_{ik}, \quad k = 1, \dots, K_n \quad (22)$$

$$\tilde{\varphi}_{ik} = \frac{\tilde{v}_{ik}}{\tilde{\varsigma}_k}, \quad k = 1, \dots, K_n; i = 1, \dots, n \quad (23)$$

This gives the following algorithm for exact simulation from the model.

1. Sample K_n from Eq. (19)
2. For $k = 1, \dots, K_n$
 - (a) Sample $(\tilde{\varphi}_{1k}, \dots, \tilde{\varphi}_{nk}) \sim \text{Dirichlet}(\alpha, \dots, \alpha)$
 - (b) Sample $\tilde{\varsigma}_k$ from

$$p(\tilde{\varsigma}) \propto \psi(\tilde{\varsigma}^2) \text{Gamma}(\tilde{\varsigma}; n\alpha, \beta) \quad (24)$$

- (c) Sample $\tilde{r}_k | \tilde{\varsigma}_k$ from

$$p(\tilde{r} | \tilde{\varsigma}) \propto (1 - e^{-\tilde{r}\tilde{\varsigma}^2}) \rho(\tilde{r}) \quad (25)$$

- (d) Sample $\tilde{Z}_k^{(n)}$ from Eq. (21)

3. For $1 \leq i, j \leq n$, set $A_{ij} = \sum_{k=1}^{K_n} \tilde{Z}_{ijk}$

where $\psi(t) = \int_0^{+\infty} (1 - e^{-wt}) \rho(dw)$ is the Laplace exponent, $\text{Dirichlet}(\alpha, \dots, \alpha)$ denotes the standard Dirichlet distribution and $\text{Gamma}(x; a, b)$ denotes the probability density function of a Gamma random variable with parameters a and b , evaluated at x . In the case of the GGP, the Laplace exponent is

$$\psi(t) = \frac{\kappa}{\sigma} ((t + \tau)^\sigma - \tau^\sigma). \quad (26)$$

One can sample from Eq. (24) and (25) using rejection.

4.2 Posterior characterization

Using Proposition 5.1 in (James, 2017), one can characterize the conditional distribution of the CRM G given the latent community counts $\tilde{Z}_{ijk}$.

Lemma 2 *Conditionally on $(\tilde{Z}_k^{(n)})_{k=1,\dots,K_n}$, the CRM G has the same distribution as*

$$G' + \sum_{k=1}^{K_n} \tilde{r}_k \delta_{\tilde{v}_{1:n,k}}$$

where G' is an inhomogeneous CRM on $\mathbb{R}_+^n$ with mean intensity

$$e^{-r(\sum_{i=1}^n v_i)^2} \rho(r) \prod_{i=1}^n f(v_i)$$

and $(\tilde{r}_k, \tilde{v}_{1:n,k})_{k=1,\dots,K_n}$ are independent of G' and iid with density

$$p(\tilde{r}_k, \tilde{v}_{1:n,k} | \tilde{Z}_k^{(n)}) = e^{-\tilde{r}_k(\sum_i \tilde{v}_{ik})^2} \tilde{r}_k^{\tilde{d}_k} \rho(\tilde{r}_k) \prod_{i=1}^n \tilde{v}_{ik}^{\tilde{m}_{ik}} f(\tilde{v}_{ik}) \quad (27)$$

where $\tilde{m}_{ik} = \sum_j \tilde{Z}_{ijk} + \tilde{Z}_{jik}$ and $\tilde{d}_k = \sum_{i,j} \tilde{Z}_{ijk}$.

In the case where f is a gamma pdf, we can use the same reparameterization as in the previous subsection with $(\tilde{\varsigma}_k, \tilde{\varphi}_{1:n,k})$ in place of $\tilde{v}_{1:n,k}$. This leads to the following conditional distributions.

$$\begin{aligned} \tilde{\varphi}_{1:n,k} | \tilde{Z}_k^{(n)} &\sim \text{Dirichlet}(\alpha + \tilde{m}_{1k}, \dots, \alpha + \tilde{m}_{nk}) \\ p(\tilde{\varsigma}_k | \tilde{Z}_k^{(n)}) &\propto \varkappa(\tilde{d}_k, \tilde{\varsigma}_k^2) \text{Gamma}(\tilde{\varsigma}_k; n\alpha + 2\tilde{d}_k, \beta) \\ p(\tilde{r}_k | \tilde{\varsigma}_k, \tilde{Z}_k^{(n)}) &\propto e^{-\tilde{r}_k \tilde{\varsigma}_k^2} \tilde{r}_k^{\tilde{d}_k} \rho(\tilde{r}_k) \end{aligned}$$

where $\varkappa(m, t) = \int_0^{+\infty} r^m e^{-rt} \rho(r) dr$. In the GGP case, we have

$$\varkappa(m, t) = \kappa \frac{\Gamma(m - \sigma)}{\Gamma(1 - \sigma)} (t + \tau)^{\sigma - m}$$

and

$$\tilde{r}_k | \tilde{\varsigma}_k, \tilde{Z}_k^{(n)} \sim \text{Gamma}(\tilde{d}_k - \sigma, \tilde{\varsigma}_k^2 + \tau).$$

4.3 Slice sampler for posterior inference

We recall that θ denote the set of hyperparameters of the mean measure ρ and pdf f . To simplify the presentation, here we suppose that we observe the complete adjacency matrix A , which means that we observe a directed and weighted graph with no missing (hidden) edge. The objective is obtain samples distributed from the conditional distribution

$$p(K_n, (\tilde{r}_k, \tilde{v}_{1:n,k})_{k=1,\dots,K_n}, \theta | A).$$

In the Appendix, we show how to do inference when we only observe a partial graph (with missing edges to predict) that can be directed or undirected, weighted or binary. In order to leverage the Poisson factorization construction, we augment the model with the latent community counts $\tilde{Z}_k$. Additionally, to deal with the unknown number of active communities K_n , we use auxiliary slice variables, similarly to other Gibbs sampler for Bayesian nonparametric models (Favaro and Teh, 2013; Kalli et al., 2011; Walker, 2007). For each directed pair (i, j) such that $A_{ij} \geq 1$ consider the scalar latent variable

$$s_{ij} | (\tilde{r}_k, \tilde{Z}_{ijk})_{k=1,\dots,K_n} \sim \text{Unif} \left(0, \min_{\{k | \tilde{Z}_{ijk} \geq 1\}} \tilde{r}_k \right) \quad (28)$$

and denote $s = \min_{ij} s_{ij}$. Note that by definition, $\tilde{r}_k \geq s$ for all $k = 1, \dots, K_n$. In the following, we will consider the communities k , active or inactive, which r is higher than s . We use the notation $\bar{\cdot}$ to denote the corresponding variables. More precisely, let

$$\bar{G} = \sum_k r_k \delta_{v_{1:n}, k} \mathbb{1}_{r_k \geq s} := \sum_{k=1}^{\bar{K}_n} \bar{r}_k \delta_{\bar{v}_{1:n}, k}$$

be the CRM corresponding to the set of active or inactive communities with weight $r_k \geq s$, of (almost surely finite) cardinality $\bar{K}_n \geq K_n$. Denote $\bar{Z}_{ijk} \geq 0$ the associated community interactions, and $\bar{Z}_k = (\bar{Z}_{ijk})$. The data augmented slice sampler draws samples asymptotically distributed from

$$p((\bar{Z}_k)_{k=1, \dots, \bar{K}_n}, \bar{G}, \theta, s \mid A).$$

The main steps of the algorithm are as follows.

1. For each directed pair (i, j) such that $A_{ij} \geq 1$, Update $(\bar{Z}_{ijk})_{k=1, \dots, \bar{K}_n}$ given the rest of the variables,
2. Update the hyperparameters θ given the rest of the variables,
3. Update $(\bar{G}, s)$ given the rest of the variables.

The details of each step are given in Appendix B. Each iteration of the Gibbs sampler has a time complexity scaling in $\bar{K}_n S$ where S is the number of nonzero entries of the matrix. Therefore, the algorithm takes advantage of the sparsity of the networks. Additionally, each entry of the sparse graph can be dealt with independently, making the algorithm straightforwardly parallelizable.

5 Experiments

We implement the algorithm described in the previous section with the GGP-Gamma scores model. We assign Gamma priors on the hyperparameters κ, τ, α with parameters $(0.1, 0.1)$. We fix $\beta = 1$. We allow up to a linear growth of the number of communities, corresponding to $\sigma < 0.5$, for small datasets and use a Gamma prior with parameter $(0.1, 0.1)$ on $1 - 2\sigma$. For larger datasets, we restrict $\sigma < 0.25$, meaning that the number of communities cannot grow at a faster rate than $\sqrt{n}$. This is obtained by using a Gamma prior with parameter $(0.1, 0.1)$ on $1 - 4\sigma$.

The model allows overlapping communities but, for visualization purposes for example, it is useful to obtain an associated partition of the nodes. For each iteration, one can cluster the nodes by assigning each node to the community where it is most active. That is, at iteration t of the MCMC algorithm, define for $i = 1, \dots, n$

$$c_i^{(t)} = \operatorname{argmax}_k \{ \sqrt{r_k^{(t)}} v_{ik}^{(t)} \} \quad (29)$$

the cluster membership of node i . We then compute an approximate Bayesian point estimate $\hat{c} = (\hat{c}_1, \dots, \hat{c}_n)$ of the partition of the nodes, using Binder's loss function (Lau and Green, 2007).

5.1 Synthetic datasets

Data generated from the GGP-gamma model. We first run the algorithm on a synthetic dataset simulated from our model, to check that the algorithm can recover the true parameters. We sample a directed and unweighted graph from the GGP-gamma model with size $n = 800$ and $\sigma = 0.2, \kappa = 1, \tau = 0.15, \alpha = 0.05, \beta = 0.2$. The number of edges of the obtained graph is 20198, and the true number of active communities is 42. We run three chains in parallel with 500,000 iterations, with 250,000 iterations for burn-in. We show in Figure 1 trace plots of the number of active communities K_n and parameter σ showing the MCMC algorithm can recover these parameters.

Statistic	$n = 500$	$n = 1000$	$n = 1500$	$n = 3000$
$\sigma = 0$	6.55 (± 1.36)	6.11 (± 1.16)	6.17 (± 1.26)	6.59 (± 1.20)
σ free	6.14 (± 1.18)	5.40 (± 0.85)	5.21 (± 0.57)	5.12 (± 0.43)
$\mathbb{P}(\sigma < 0)$	0.69	0.93	0.97	0.99

Table 1 Concentration of the posterior of K_n when increasing n on synthetic dataset with a number of communities $K = 5$ fixed (independent of n). We report the mean and the standard deviation (between parenthesis) of the posterior of K_n for two models: first line $\sigma = 0$ is constant, recovering the Dirichlet model second line σ is free and inferred from the data, in the third line we report the posterior probability of $\sigma < 0$ when σ is considered unknown.

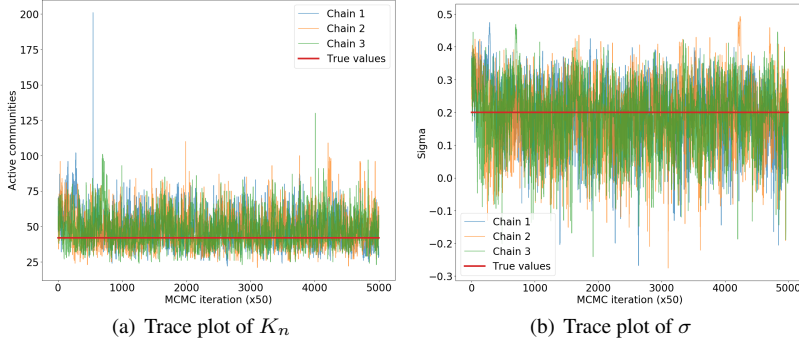


Fig. 1 Trace plots of (a) the number of active communities K_n and (b) σ , on a synthetic example.

Data generated from a Poisson factor model with a fixed number of communities. Following Miller and Harrison (2013), we know that the Dirichlet process is inconsistent for estimating the number of clusters if the number of clusters is fixed and does not increase with n . Similarly, we conjecture that when $\sigma = 0$ is fixed, corresponding to the setting of Zhou (2015), the model will fail recovering the right number of communities, whereas if σ is free, we can expect it to concentrate on negative values. The model should then recover the correct number of communities if the data is generated from a Poisson factor model. A proper posterior consistency analysis is beyond the scope of this article. However we design a simple numerical experiment to support this conjecture. We take $K = 5$ true communities, the affiliations $v_{i,k}$ are iid Gamma(0.1, 1) and the five communities importance $(r_1, \dots, r_5)$ are iid Gamma(2, 0.2). Networks of increasing sizes $n = 500, 1000, 1500, 3000$ are then generated from the Poisson factor model (3).

We estimate the number of active communities K_n under the model with $\sigma = 0$ and with σ unknown. In Table 1 we report the posterior mean and variance of the recovered number of communities. We can see that the gamma process model ($\sigma = 0$) does not seem to concentrate around the right value. On the contrary, when σ is free, we see that the posterior of K_n seems to concentrate around $K = 5$. For $n = 500$, the posterior on σ ranges from negative to positive values, which explains the higher variance. However, from $n = 1000$ onward, the posterior of σ concentrates on negative values, which translates in a significant decrease of variance for the posterior of K_n .

Data generated from the Stochastic Block Model (SBM). From our experimentation, it seems that Poisson factorization models do not capture well the generating process of the SBM. Both when $\sigma = 0$ and σ free, the model creates many very small communities. However, the model is able to recover with high precision the true communities once we cluster the nodes using (29).

We generate a synthetic dataset from a SBM with $n = 600$ with three communities of size 200 each. The probability of an edge between nodes of a same community is $p_{in} = 0.1$, the probability of an edge between nodes from different communities is $p_{out} = 0.01$. The resulting dataset is hence undirected and unweighted with 7239 edges.

The GGP-Gamma model does not capture well the SBM generating process and the posterior mean of the number of communities is $K_n = 38.6$, with σ positive. We obtain 3 main communities and the rest are very small (composed of a few edges each). As explained pre-

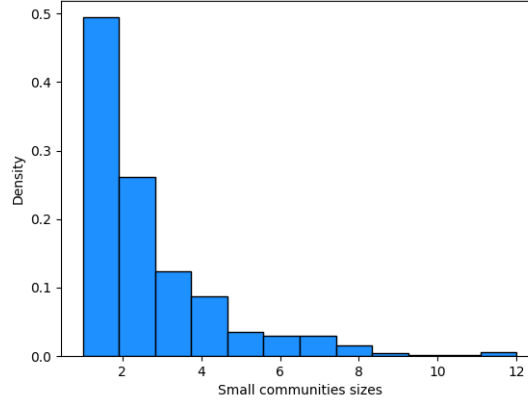


Fig. 2 Posterior density of the sizes of the small communities for the dataset generated from a Stochastic Block Model.

viously, we cluster the nodes by assigning each one of them to the community to which it has the highest scaled affiliation. We then get an average of 8.35 communities, 3 of which are on average composed of 195.5 nodes each. We plot the distribution of the sizes of the small communities in Figure 2. In Table 2 we report the contingency table of the posterior clusters.

We also use the model with $\sigma = 0$ on this dataset and find very similar results. After clustering the nodes, we find that the average number of communities is 7.66, which slightly less than previously, but the small communities are slightly larger on average, giving at the end an average size for each large cluster of 195.5, which is exactly the same. We do not report here the contingency matrix as it is very similar to the one we obtain with σ free.

	Estimated communities			
	$\hat{C}_1$	$\hat{C}_2$	$\hat{C}_3$	Others
C_1	97.67	0.09	0.09	2.15
C_2	0.01	97.94	0.0	1.95
C_3	0.14	0.0	97.43	2.43

Table 2 Contingency table of the posterior communities in per cent of the true communities size.

5.2 Political blogs

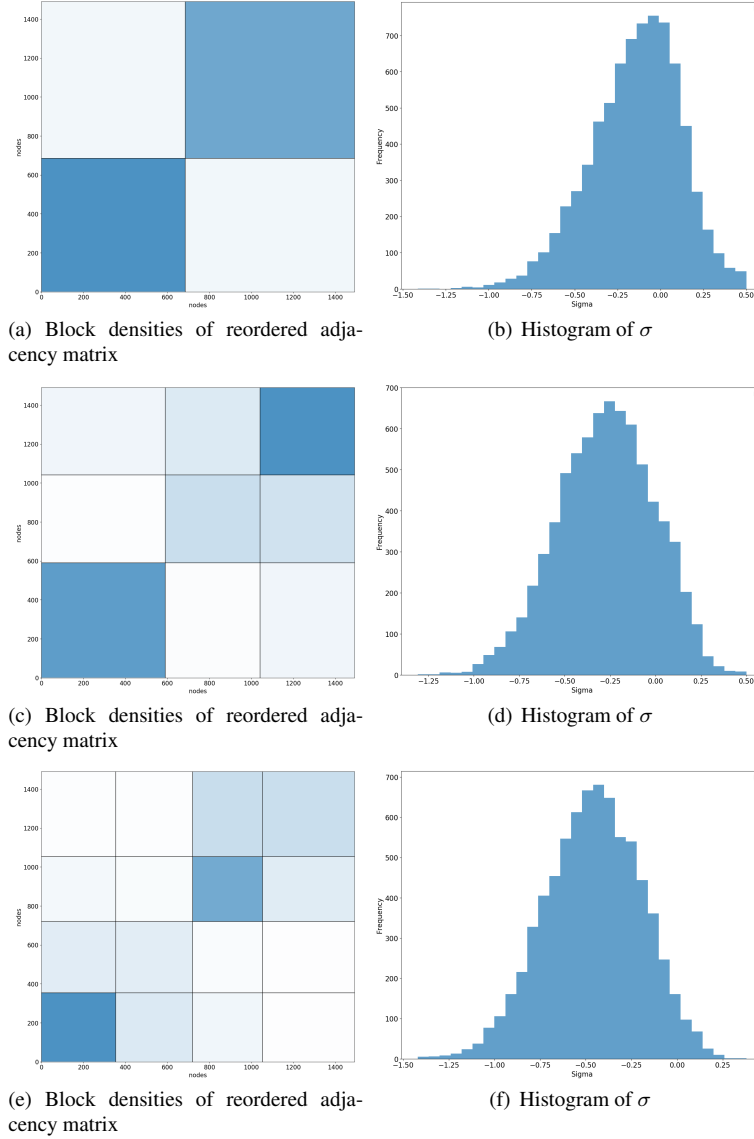
The polblogs network (Adamic and Glance, 2005) is the network of the American political blogosphere in February 2005. It is a directed unweighted graph, where there is an edge (i, j) if blog i cites blog j . It is composed of 1490 nodes and 19025 edges. For each node, some ground truth information about its political affiliation (republican/democrat) is known.

We will use this dataset in order to illustrate the role of the parameter α in the model. As indicated in Section 3 this parameter tunes the amount of overlapping between the communities. A smaller value enforces less overlap between communities. We run three chains with 500,000 iterations. The posterior samples of σ for three different values of α are in also shown in Figure 3. Nodes are reordered according to their estimated membership $\hat{c}$ (through (29)), and Figure 3 shows the densities of connection between and within clusters for three different values of α . Depending on the amount of overlapping, we obtain two ($\alpha = 0.8$), three ($\alpha = 0.4$) or four ($\alpha = 0.2$) communities. In order to interpret those communities, we calculate in Table 3 for each community the proportion of interactions between democrat blogs, between a democrat and a republican blog, and between two republican blogs. For $\alpha = 0.8$, there are two estimated communities which can clearly be identified as democrat (community #1) and republican (community 2). For $\alpha = 0.4$, we have three communities.

Table 3 Proportion of the interactions of the features in each block for different values of overlapping

	$\alpha = 0.8$		$\alpha = 0.4$			$\alpha = 0.2$			
	1	2	1	2	3	1	2	3	4
Dem/Dem	91.5	0.7	93.2	1.1	0.8	95	84	1.5	0.1
Dem/Rep	8.0	9.5	6.5	13	5.1	5	14.5	15.5	4.8
Rep/Rep	0.5	89.8	0.3	85.9	94.1	0	1.5	83	95.1

One is mostly associated to democrat blogs (#1) while the other two correspond to a split of the republican blogs into right (#2) and center-right (#3) groups. For $\alpha = 0.2$, we obtain a further split of the democrat blogs into left (#1) and center-left (#2) groups. Increasing the value of α therefore leads to a finer and finer partition of the nodes.

**Fig. 3** (Left) Estimated communities and (right) posterior on σ for the Polblogs dataset with (top row) $\alpha = 0.8$, (middle row) $\alpha = 0.4$ and (bottom row) $\alpha = 0.2$.

5.3 Wikipedia topcast

The network is a partial web graph of Wikipedia hyperlinks collected in September 2011 (Klymko et al., 2014). It is a directed unweighted graph where an edge (i, j) corresponds to a citation from a page i to page j . We restrict it to the first 3000 nodes, and the associated 5687 edges. We run three MCMC chains for 200,000 iterations. Trace plots of the number of active communities and parameter σ are given in Figure 4. Figure 5 shows the adjacency matrix re-ordered by communities, as explained in the previous section. In order to check that the learnt communities/features are meaningful, we report in Figures the proportions of webpages associated to a given category within a given community/feature (note that a webpage can be associated to multiple categories hence the proportion do not sum to 1).

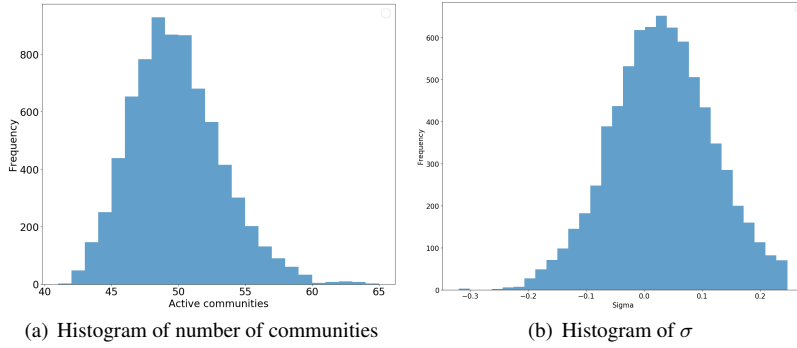


Fig. 4 Posterior of K_n and σ for the Wiki-topcats dataset

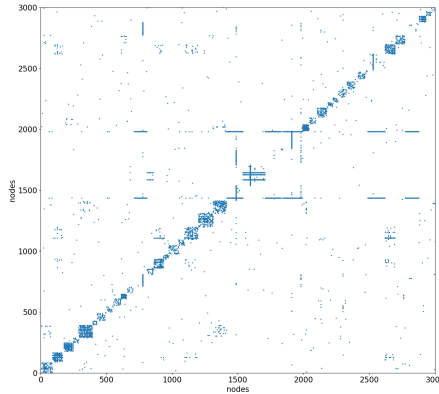


Fig. 5 Reordered adjacency matrix of the Wikipedia topcats dataset

Note that, while the approach is able to estimate the latent block-structure, this dataset has the particularity of having star nodes, a feature that is not captured by our model.

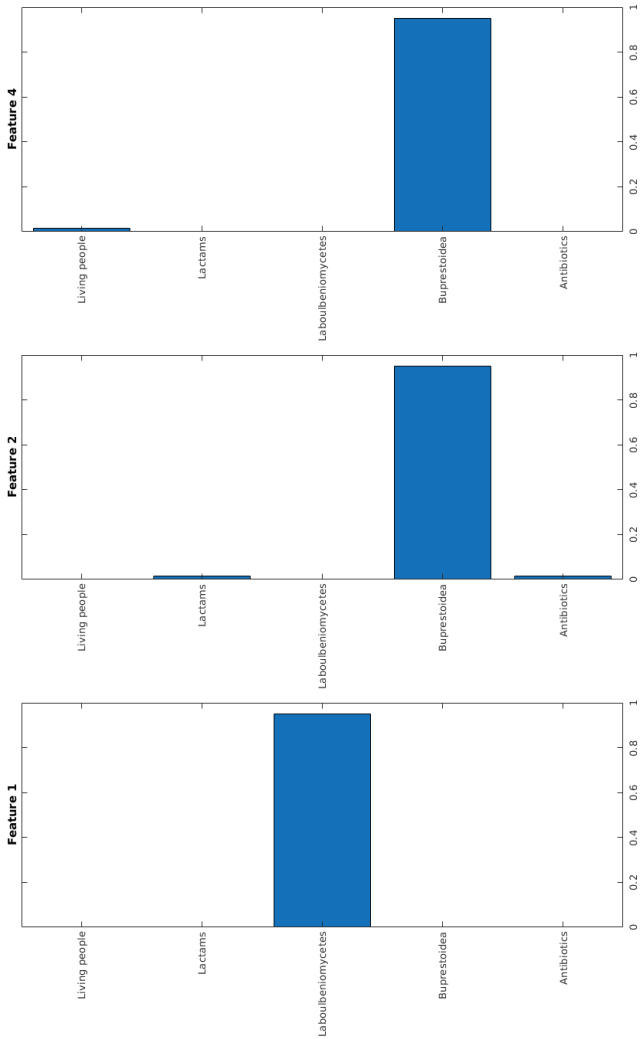


Fig. 6 Features compared to categories for the wikipedia dataset

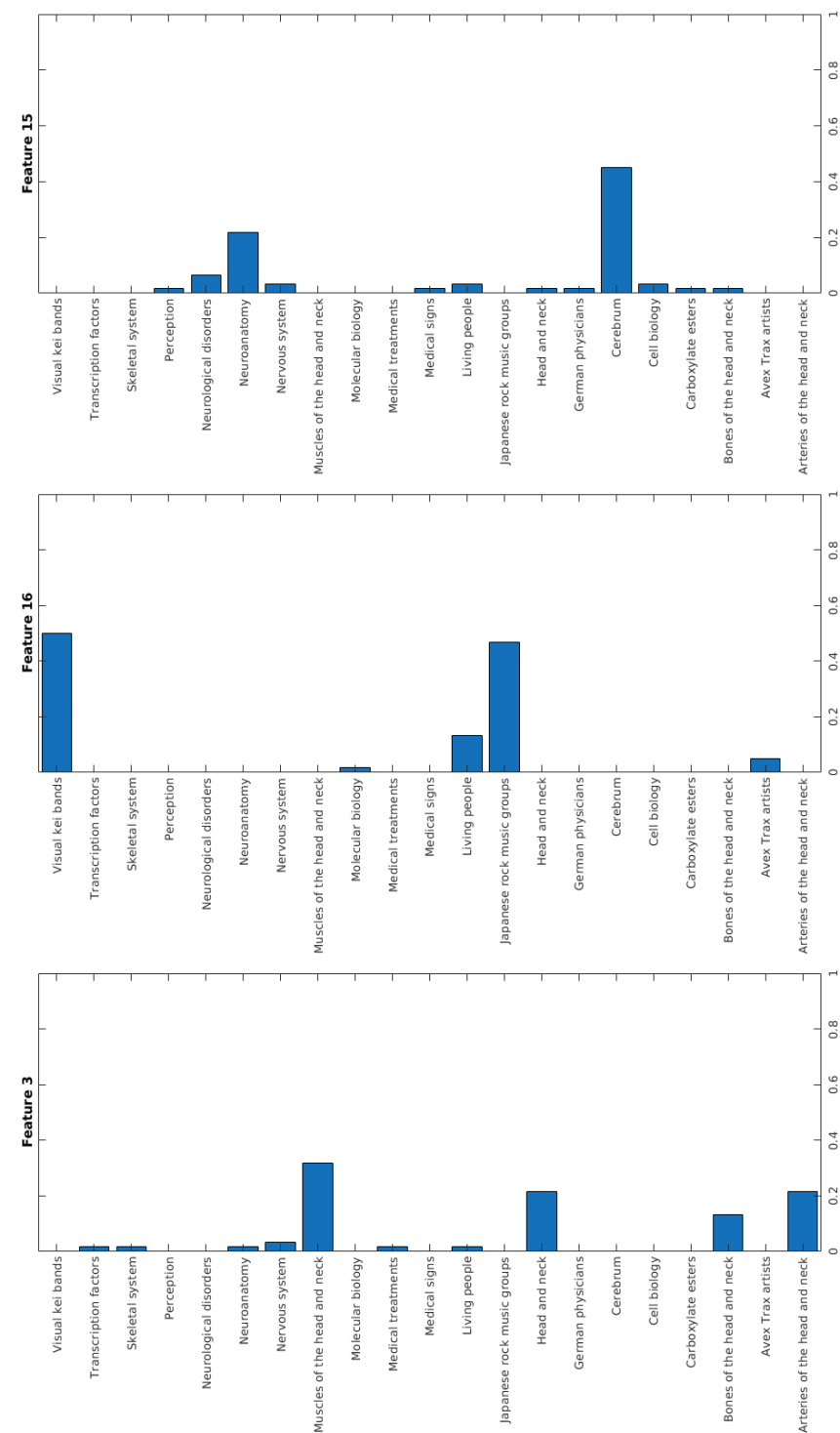


Fig. 7 Features compared to categories for the wikipedia dataset

5.4 Deezer

The dataset was collected from the music streaming service Deezer in November 2017 (Rozem-berczki et al., 2018). It represents the friendship network of a subset of Deezer users from

Romania. It is an undirected unweighted graph where nodes represent the users and edges are the mutual friendships. There are 41773 nodes and 125826 edges. We run three chains with 100000 iterations each. Posterior histograms of the number of active communities and σ are given in Figure 8. The algorithm finds around 45 communities/features for this dataset. The reordered adjacency matrix and block densities based on the point estimate of the partition are given in Figure 9.

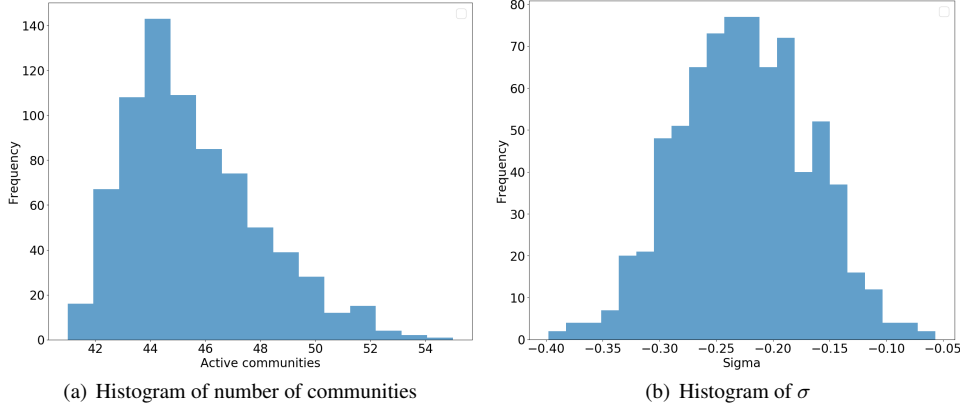


Fig. 8 Posterior of K_n and σ on Deezer's dataset

Now we can reorder the nodes using approximate MAP clustering as previously. We obtain the following adjacency matrix

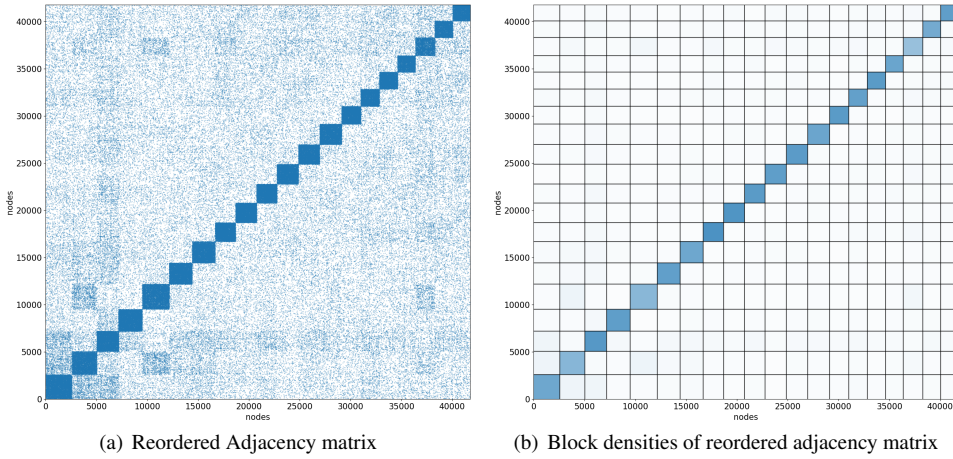


Fig. 9 Reordered adjacency matrix and block densities for Deezer's dataset.

For each individual in the network, a list of musical genres liked by that person are available. There are in total 84 distinct genres. We represent in Figure 10 the proportion of individuals who liked a subset of the 84 genres for three different communities where the interpretation in terms of genres is quite clear. The overall proportion of individuals liking a given genre is shown at the bottom of Figure 10. If the bar is red, this indicates that the proportion is 10% higher in the community than in the population. If the bar is blue, this means the proportion is 10% lower. Community 11 can be interpreted as *R&B*, Community 8 as Dance, and Community 3 as Rock music. For some of the communities, not reported here, the interpretation in terms of the liked genres is less clear, and may be due to other covariates.

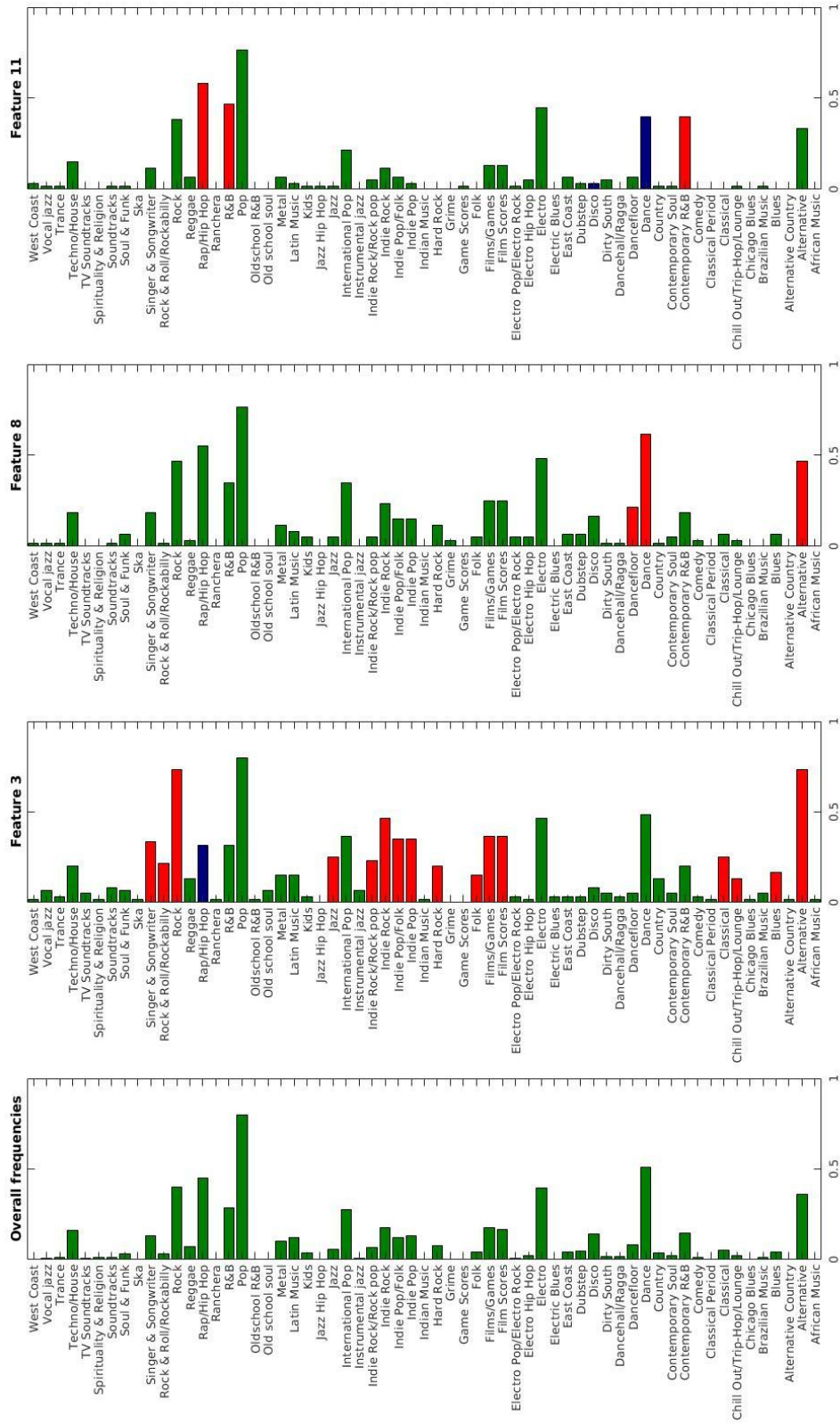


Fig. 10 Features compared to genres for Deezer's dataset.

6 Discussion

The model presented in this paper assumed the same parameter β for each node. We can also consider a degree corrected version of the model, similarly to Zhou (2015), where each node is assigned a different parameter $\beta_i > 0$ and then defining $Z_{ijk} \sim \text{Poisson}(\frac{r_k v_{ik} v_{jk}}{\beta_i \beta_j})$. It is unclear however if a MCMC sampler targeting the exact posterior distribution could be implemented, and one may need to resort to some truncation approximation as in Zhou (2015).

The count matrix (A_{ij}) is infinitely exchangeable, hence the model presented in this article lead to asymptotically dense graphs. That is, $\sum_{1 \leq i, j \leq n} A_{ij} \asymp n^2$ as n tends to infinity. In order to obtain sparse graphs, we could consider two different strategies. The first solution consists in dropping the infinite exchangeability property and take $\beta_i^{(n)} \rightarrow +\infty$ with n , then the number of edges will behave as $(n/\beta^{(n)})^2$ (we can for instance take $\beta_i^{(n)} = \sqrt{n}$ for any node i to obtain a linear growth of the number of edges). The model would still be finitely exchangeable for any fixed n , but not projective anymore. The second solution would be to consider the different notion of infinite exchangeability developed in (Caron and Fox, 2017) and consider $(\beta_i)_i$ as a realization of a Poisson point process.

Finally, we presented a model for count (and binary) data. The results build on the additive contributions of the communities, which is why we chose the Poisson distribution on the entries of the adjacency matrix (A_{ij}) . We can generalize to non count data using other probability distributions which are closed under convolution. For example, one could consider $A_{ij} \sim \text{Gamma}(\sum_k r_k v_{ik} v_{jk}, 1)$ for $A_{ij} \in \mathbb{R}_+$ or $A_{ij} \sim \mathcal{N}(\sum_k r_k v_{ik} v_{jk}, 1)$ for $A_{ij} \in \mathbb{R}$.

References

- L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.
- Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.
- Alex Arenas, Leon Danon, Albert Diaz-Guilera, Pablo M Gleiser, and Roger Guimera. Community analysis in social networks. *The European Physical Journal B*, 38(2):373–380, 2004.
- B. Ball, B. Karrer, and M. E. J. Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.
- A. Brix. Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31(4):929–953, 1999.
- F. Caron and E. B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society B*, 79(5), 2017.
- A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational intelligence and neuroscience*, 2009, 2009.
- Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
- D. B. Dunson and A. H. Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11–25, 2005.
- Daniele Durante and David B Dunson. Nonparametric bayes dynamic modelling of relational data. *Biometrika*, 101(4):883–898, 2014.
- S. Favaro and Y. W. Teh. MCMC for normalized random measure mixture models. *Statistical Science*, pages 335–359, 2013.
- A. Gneden, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability surveys*, 4:146–171, 2007.
- P. Gopalan, F. J. Ruiz, R. Ranganath, and D. M. Blei. Bayesian nonparametric Poisson factorization for recommendation systems. In *AISTATS*, pages 275–283, 2014.
- P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with hierarchical Poisson factorization. In *UAI*, pages 326–335, 2015.

- J. E. Griffin and F. Leisen. Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):525–545, 2017.
- P. D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, 15(4):261, 2009.
- P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.
- P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986.
- L. F. James. Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093*, 2002.
- L. F. James. Poisson latent feature calculus for generalized Indian buffet processes. *arXiv preprint arXiv:1411.2936*, 2014.
- L. F. James. Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors. *Ann. Statist.*, 45(5):2016–2045, 10 2017. doi: 10.1214/16-AOS1517. URL <https://doi.org/10.1214/16-AOS1517>.
- M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.
- J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.
- J.F.C. Kingman. *Poisson processes*, volume 3. Oxford University Press, USA, 1993.
- Christine Klymko, David Gleich, and Tamara G Kolda. Using triangles to improve community detection in directed networks. *arXiv preprint arXiv:1404.5874*, 2014.
- J. W. Lau and P. J. Green. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558, 2007.
- D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.
- A. Lijoi and I. Prünster. Models beyond the Dirichlet process. *Bayesian nonparametrics*, 28(80):3, 2010.
- H. Ma, C. Liu, I. King, and M. R. Lyu. Probabilistic factor models for web site recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 265–274. ACM, 2011.
- Jeffrey W Miller and Matthew T Harrison. A simple example of dirichlet process mixture inconsistency for the number of components. In *Advances in neural information processing systems*, pages 199–206, 2013.
- P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.
- J. Pitman. Poisson-kingman partitions. *Lecture Notes-Monograph Series*, pages 1–34, 2003.
- D. Pollard. Mini empirical. *Manuscript*. <http://www.stat.yale.edu/pollard/Books/Mini>, 2015.
- Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the national academy of sciences*, 101(9):2658–2663, 2004.
- B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton. Gemsec: Graph embedding with self clustering. *ArXiv e-prints*, feb 2018.
- Clara Stegehuis, Remco Van Der Hofstad, and Johan SH Van Leeuwen. Power-law relations in random networks with communities. *Physical Review E*, 94(1):012302, 2016.
- M. K. Titsias. The infinite gamma-Poisson feature model. In *Advances in Neural Information Processing Systems*, pages 1513–1520, 2008.
- T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1825–1828. IEEE, 2008.
- S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, 2007.
- M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Artificial Intelligence and Statistics*, pages 1135–1143, 2015.
- M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the*

Fifteenth International Conference on Artificial Intelligence and Statistics, volume 22 of *Proceedings of Machine Learning Research*, pages 1462–1471, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

A Proofs

A.1 Technical Lemmas

Lemma 3 (Gnedin et al. (2007), Propositions 17 and 19) Let ρ be a Lévy measure, let $\bar{\rho}(x) = \int_x^\infty \rho(r)dr$ be the tail Lévy intensity and $\psi(t) = \int (1 - e^{-rt})\rho(dr)$ its Laplace exponent. Then the two following conditions are equivalent:

$$\bar{\rho}(x) \underset{x \rightarrow 0^+}{\asymp} \ell(1/x)x^{-\sigma} \quad (30)$$

$$\psi(t) \underset{x \rightarrow +\infty}{\asymp} \Gamma(1 - \sigma)t^\sigma \ell(t) \quad (31)$$

with ℓ a slowly varying function and $0 \leq \sigma < 1$.

Besides, if we let $\psi_d(t) = \frac{t^d}{d!} \int r^d e^{-rt} \rho(dr)$

1. if $\sigma > 0$, then (30) implies that $\psi_d(t) \underset{t \rightarrow +\infty}{\asymp} \frac{\sigma \Gamma(d - \sigma)}{d!} t^\sigma \ell(t)$
2. if $\sigma = 0$, then (30) implies that $\psi_d(t) = o(\ell(t))$

Lemma 4 (Pollard, 2015, Exercise 15) Let X be a Poisson random variable with parameter λ . For any $t > 0$

$$\mathbb{P}(|X - \lambda| \geq \lambda t) \leq 2e^{-\frac{\lambda t^2}{2(1+t)}}. \quad (32)$$

Lemma 5 Let $(X_n)_{n \geq 1}$ be a sequence of Poisson random variables with mean $(\mu_n)_{n \geq 1}$. If $\log n = o(\mu_n)$ then $X_n \asymp \mu_n$ almost surely as n tends to infinity.

Proof Let $0 < \epsilon < 1/2$. Using Lemma 4, we have

$$\begin{aligned} \mathbb{P}\left(\left|\frac{X_n}{\mu_n} - 1\right| \geq \epsilon\right) &\leq 2e^{-\frac{\epsilon^2 \mu_n}{4}} \\ &= 2n^{-\frac{\epsilon^2 \mu_n}{4 \log n}} \end{aligned} \quad (33)$$

Using the assumption, we have that $-\frac{\epsilon^2 \mu_n}{4 \log n} \rightarrow -\infty$. Therefore, the RHS of (33) is summable. The almost sure result follows from Borel-Cantelli lemma.

Lemma 6 For any $x, y \geq 0$, we have the following bound

$$1 - e^{-xy} \leq \max(1, y)(1 - e^{-x}) \quad (34)$$

Proof The bound is trivial when $y \leq 1$. Consider the case $y > 1$. For all x , the function $y \rightarrow e^{-xy}$ is convex hence $f_x(y) = \frac{e^{-xy} - 1}{y}$ is a monotonically non-decreasing function of y therefore

$$\frac{e^{-xy} - 1}{y} \geq e^{-x} - 1$$

for $y \geq 1$.

A.2 Proofs of Section 3

Proof (Proof of Proposition 1)

The result for K_n is proved in James (2014) in the general context of GIBP. We provide here the details of the proof for K_n , which can be straightforwardly adapted to $K_{n,j}$.

First, let us remark that the bound (34) together with assumptions (A1) and (A2) imply $\Psi(n) < \infty$. For $s < 0$,

$$\begin{aligned} \mathbb{E}[e^{sK_n}] &= \mathbb{E}\left[\prod_{k=1}^{\infty} \mathbb{E}\left[e^{\sum_{1 \leq i, j \leq n} s \mathbb{1}_{Z_{ijk} \geq 1}} \middle| G\right]\right] \\ &= \mathbb{E}\left[\prod_{k=1}^{\infty} \left[e^s + (1 - e^s)e^{-r_k \left(\sum_{i=1}^n v_{ik}\right)^2}\right]\right] \\ &= \mathbb{E}\left[e^{\sum_k \log[e^s + (1 - e^s)e^{-r_k \left(\sum_{i=1}^n v_{ik}\right)^2}]} \right] \end{aligned}$$

Then, since

$$\begin{aligned} \log[e^s + (1 - e^s)e^{-r_k(\sum_{i=1}^n v_{ik})^2}] &= \log[e^{-r_k(\sum_{i=1}^n v_{ik})^2} + e^s(1 - e^{-r_k(\sum_{i=1}^n v_{ik})^2})] \\ &\leq \log[1 + e^s(1 - e^{-r_k(\sum_{i=1}^n v_{ik})^2})] \\ &\leq 1 - e^{-r_k(\sum_{i=1}^n v_{ik})^2} \end{aligned}$$

and the last part is integrable, we can use Campbell's theorem (Kingman, 1993) to get:

$$\begin{aligned} \mathbb{E}[e^{sK_n}] &= \exp\left[\int (e^{\log[e^s + (1 - e^s)e^{-r(\sum_{i=1}^n v_i)^2}]} - 1) \prod_{i=1}^n F(dv_i) \rho(dr)\right] \\ &= \exp[(e^s - 1)\Psi(n)] \end{aligned}$$

We can prove similarly that $K_{n,d}$ is a Poisson random variable with mean $\Psi_d(n)$ and that $\sum_{d \geq D} K_{n,d}$ is Poisson distributed with mean $\sum_{d \geq D} \Psi_d(n)$. The assumption $\Psi_d(n) < \infty$ is also sufficient in this case to apply Campbell's theorem.

Proof (Proof of Proposition 2) From Proposition 1, we get that

$$\mathbb{E}(K_n) = \int (1 - e^{-r(\sum_{i=1}^n v_i)^2}) \prod_{i=1}^n F(dv_i) \rho(dr) = \Psi(n) \quad (35)$$

Let $(V_i)_{i \in \mathbb{N}}$ be i.i.d random variables with distribution F . By assumption, $0 < \mathbb{E}[V_i] = \mu < +\infty$ and $\text{Var}[V_i] = \tau^2 < +\infty$. Let $\epsilon > 0$. Let $A(r)$ be defined for $r > 0$ by

$$A(r) = \mathbb{E}[1 - e^{-r(\sum_{i=1}^n V_i)^2}]$$

Since $v \mapsto 1 - e^{-rv}$ is concave, using successively Jensen's inequality and the independence of (V_i) , we obtain

$$\begin{aligned} A(r) &\leq 1 - e^{-r\mathbb{E}[(\sum_{i=1}^n V_i)^2]} \\ &\leq 1 - e^{-r(n^2\mu^2 + n\tau^2)} \\ &\leq 1 - e^{-(1+\epsilon)rn^2\mu^2} \end{aligned}$$

where the last inequality holds for any $\epsilon > 0$ when $n > \frac{\tau^2}{\epsilon\mu^2}$. Therefore, for $\epsilon > 0$ and $n > \frac{\tau^2}{\epsilon\mu^2}$

$$\Psi(n) \leq \int (1 - e^{-(1+\epsilon)rn^2\mu^2}) \rho(dr).$$

Besides, since $v \mapsto 1 - e^{-rv}$ is increasing, by Markov's inequality we have for any $\epsilon > 0$

$$\begin{aligned} A(r) &= \mathbb{E}\left[1 - e^{-rn^2(\frac{1}{n}\sum_{i=1}^n V_i)^2}\right] \\ &\geq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n V_i \geq \frac{\mu}{(1+\epsilon)}\right) \left(1 - e^{-\frac{n^2\mu^2}{(1+\epsilon)^2}r}\right) \end{aligned}$$

Hence

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n V_i \geq \frac{\mu}{(1+\epsilon)}\right) \psi\left(\frac{n^2\mu^2}{(1+\epsilon)^2}\right) \leq \Psi(n) \leq \psi((1+\epsilon)n^2\mu^2)$$

where $\psi(t) = \int (1 - e^{-rt}) \rho(dr)$ is the Laplace exponent. Furthermore, by the law of large numbers,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n V_i \geq \frac{\mu}{(1+\epsilon)}\right) \rightarrow 1.$$

Therefore, under Assumption (A4), Lemma 3 implies

$$\mathbb{E}(K_n) \asymp \Gamma(1 - \sigma) \mu^{2\sigma} n^{2\sigma} \ell(n^2)$$

as n tends to infinity.

In the finite-activity case, that is $\sigma = 0$ and $\ell(t) \rightarrow \bar{\rho}(0) < \infty$, we have $\mathbb{E}(K_n) \rightarrow \bar{\rho}(0)$ hence K_n tends in distribution to $\text{Poisson}(\bar{\rho}(0))$.

Now, for $\sigma > 0$, the almost sure result (15) follows from Lemma 5 and the fact that for every slowly varying function ℓ_0 and every $\epsilon > 0$

$$\lim_{x \rightarrow \infty} \ell_0(x) x^{-\epsilon} = 0.$$

Finally, assume that $(K_n)_{n \geq 1}$ is non-decreasing. We only need to prove the asymptotic behavior for $\sigma = 0$. In that setting, $\Psi(n) \asymp \ell(n^2)$. Using the assumption that $\int \rho(dr) = \infty$, we therefore have $\lim_{n \rightarrow \infty} \Psi(n) = \infty$. Let $n \geq 1$,

$$\Psi(n+1) - \Psi(n) = \int \mathbb{E} \left[e^{-r(\sum_{i=1}^n V_i)^2} - e^{-r(\sum_{i=1}^{n+1} V_i)^2} \right] \rho(r) dr$$

Since $V_i \geq 0$ a.s, it comes that $(\Psi(n))_n$ is non-decreasing. Now extend the sequence $(\Psi(n))_n$ to a non-decreasing and continuous function Ψ on $\mathbb{R}_+$ (by linear interpolation for instance). Let $t > 1$, then

$$1 \leq \frac{\Psi(t+1)}{\Psi(t)} \leq \frac{\Psi(3\lfloor t \rfloor)}{\Psi(\lfloor t \rfloor)} \asymp \frac{\ell(9\lfloor t \rfloor^2)}{\ell(\lfloor t \rfloor^2)} \rightarrow 1$$

Hence $\lim_{t \rightarrow \infty} \frac{\Psi(t+1)}{\Psi(t)} = 1$

Now, for every integer $m \geq \Psi(0)$, choose t_m such that $\Psi(t_m) = m$. We have that (t_m) is non-decreasing and diverges. Since Ψ is increasing, it comes

$$\Psi(t_{m+1} - 1) \leq \Psi(\lfloor t_{m+1} \rfloor - 1) \leq \Psi(t_{m+1}) = m + 1$$

Hence, $\Psi(\lfloor t_{m+1} \rfloor - 1) \asymp m$. Then, using Lemma 5, we get that

$$K_{\lfloor t_{m+1} \rfloor - 1} \asymp \Psi(\lfloor t_{m+1} \rfloor - 1) \asymp m \quad a.s.$$

Finally, let $n \geq \Psi(0)$, let $m_n = \min\{m \mid n \in \{\lfloor t_m \rfloor, \dots, \lfloor t_{m+1} \rfloor - 1\}\}$,

$$\frac{K_{\lfloor t_{m_n} \rfloor}}{\Psi(\lfloor t_{m_n+1} \rfloor - 1)} \leq \frac{K_n}{\Psi(n)} \leq \frac{K_{\lfloor t_{m_n+1} \rfloor - 1}}{\Psi(\lfloor t_{m_n} \rfloor)}$$

Since $t_{m_n} \rightarrow \infty$, both bounds converge to 1 almost surely, which gives the result.

Proof (Proof of Proposition 3) As for Proposition 2, we only need to show that for $d \geq 1$,

$$\Psi_d(n) \asymp \frac{\sigma \Gamma(d - \sigma)}{d!} n^{2\sigma} \mu^{2\sigma} \ell(n^2)$$

Therefore the proof is very similar to the one of Proposition 2. However, there are some technicalities we need to address since here $v \mapsto \frac{v^d}{d!} e^{-rv}$ is neither convex nor decreasing. Like previously, we will lower bound and upper bound $\Psi_d(n)$ by two quantities that are equivalent to $\frac{\sigma \Gamma(d - \sigma)}{d!} n^{2\sigma} \mu^{2\sigma} \ell(n^2)$.

Let us first introduce some notations. Let $(V_i)_{i \in \mathbb{N}}$ i.i.d variables with distribution F . Let $S_n = \sum_{i=1}^n V_i$ and $\varphi_d(r, v)$ defined as

$$\varphi_d(r, v) = \frac{r^d v^d}{d!} e^{-rv}.$$

Now, define $A_d(r, n)$ for $r > 0$ by

$$A_d(r, n) = \mathbb{E}[\varphi_d(r, S_n^2)]$$

Suppose that $0 < \sigma < 1$, let $\epsilon > 0$, recalling that $\mathbb{E}V_i = \mu$, define $B_\epsilon = [\frac{n\mu}{\epsilon+1}, (1+\epsilon)n\mu]$. Let us notice that the law of large numbers gives us that $\mathbb{P}(S_n \in B_\epsilon) \rightarrow 1$.

1. Lower bound: We have that

$$\mathbb{E}[\mathbb{1}_{S_n \in B_\epsilon} \varphi_d(r, S_n^2)] \leq A_d(r, n)$$

Besides, for all $v \in B_\epsilon$ and all $r > 0$

$$\phi_d(r, v) \geq \frac{r^d \mu^{2d} n^{2d}}{(1+\epsilon)^{2d} d!} e^{-rn^2(1+\epsilon)^2 \mu^2}$$

hence

$$\mathbb{P}(S_n \in B_\epsilon) \frac{r^d \mu^{2d} n^{2d}}{(1+\epsilon)^{2d} d!} e^{-rn^2(1+\epsilon)^2 \mu^2} \leq \mathbb{E}[\mathbb{1}_{S_n \in B_\epsilon} \varphi_d(r, S_n^2)].$$

Therefore, using Lemma 3 and since $\mathbb{P}(S_n \in B_\epsilon) \rightarrow 1$, we have for n large enough

$$\frac{\sigma \Gamma(d - \sigma)}{d! (1 + \epsilon)^{4d+1-2\sigma}} n^{2\sigma} \mu^{2\sigma} \ell(n^2) \leq \int A_d(r, n) \rho(dr).$$

2. Upper bound: We have that

$$A_d(r, n) = \mathbb{E}[\mathbb{1}_{S_n \in B_\epsilon} \varphi_d(r, S_n^2)] + \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \varphi_d(r, S_n^2)]$$

Like previously, since

$$\mathbb{E}[\mathbb{1}_{S_n \in B_\epsilon} \varphi_d(r, nS_n)] \leq \mathbb{P}(S_n \in B_\epsilon) \frac{r^d (1+\epsilon)^{2d} \mu^{2d} n^{2d}}{d!} e^{-r \frac{n^2 \mu^2}{(1+\epsilon)^2}}$$

We find that for n large enough,

$$\int A_d(r, n) \rho(dr) \leq \frac{\sigma \Gamma(d-\sigma)(1+\epsilon)^{4d+1-2\sigma}}{d!} n^{2\sigma} \mu^{2\sigma} \ell(n^2) + \int \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \varphi_d(r, S_n^2)] \rho(dr)$$

Therefore, we only need to prove that

$$\int \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \varphi_d(r, S_n^2)] \rho(r) dr = o(n^{2\sigma} \ell(n^2)).$$

In order to do so, we split the integral with respect to r in two parts, an integral over $(0, \frac{1}{n^2})$ and an integral over $(\frac{1}{n^2}, \infty)$ and show that both are $o(n^{2\sigma} \ell(n^2))$. Since $\varphi_d(r, v) \leq 1$,

$$\begin{aligned} \int_{1/n^2}^{\infty} \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \varphi_d(r, S_n^2)] \rho(dr) &\leq \mathbb{P}(S_n \notin B_\epsilon) \int_{1/n^2}^{\infty} \rho(dr) \\ &= \mathbb{P}(S_n \notin B_\epsilon) \bar{\rho}(1/n^2) \\ &= o(n^{2\sigma} \ell(n^2)) \end{aligned}$$

where the last line follows from the law of large numbers and Assumption (A4). Besides,

$$\begin{aligned} \int_0^{1/n^2} \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \varphi_d(r, S_n^2)] \rho(dr) &= \int_0^{1/n^2} \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \frac{r S_n^2}{d} \varphi_{d-1}(r, S_n^2)] \rho(dr) \\ &\leq \int_0^{1/n^2} \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} r S_n^2] \rho(dr) \\ &= \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \frac{S_n^2}{n^2}] \int_0^{1/n^2} r n^2 \rho(dr) \\ &\leq \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \frac{S_n^2}{n^2}] e \int_0^{1/n^2} r n^2 e^{-r n^2} \rho(dr) \\ &\leq 8 \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \frac{S_n^2}{n^2}] \frac{\sigma \Gamma(d-\sigma)}{d!} n^{2\sigma} \ell(n^2) \end{aligned}$$

where the last inequality holds for n large enough by Assumption (A4) and Lemma 3. Now, we have that

$$\frac{S_n^2}{n^2} < \frac{1}{n} \sum_{i=1}^n V_i^2$$

Since $(V_i^2)_i$ are i.i.d random variables in $\mathcal{L}_1$, we know that $(\frac{1}{n} \sum_{i=1}^n V_i^2)_{n \geq 1}$ is uniformly integrable. Therefore, $(\mathbb{1}_{S_n \notin B_\epsilon} \frac{S_n^2}{n^2})_{n \geq 1}$ is uniformly integrable. Besides, using the law of large numbers, the sequence converges almost surely, and hence in probability, to 0. Therefore, $\lim_n \mathbb{E}[\mathbb{1}_{S_n \notin B_\epsilon} \frac{S_n^2}{n^2}] = 0$, which concludes the proof.

For $\sigma = 0$, the previous computations for the upper bound give that almost surely, $\Psi_d(n) = o(\ell(n^2)) = o(\Psi(n))$. Now let $D > 1$,

$$\mathbb{E} \sum_{d \geq D} K_{n,d} = \Psi(n) - \sum_{d=1}^{D-1} \Psi_d(n) \asymp \ell(n^2)$$

And since $x \mapsto \sum_{d \geq D} \varphi_d(1, x)$ is non-decreasing, $(\mathbb{E} \sum_{d \geq D} K_{n,d})_n$ is non-decreasing, therefore, similarly to the proof for $\sigma = 0$ for $(K_n)_n$, we find that

$$\sum_{d \geq D} K_{n,d} \asymp \mathbb{E} \sum_{d \geq D} K_{n,d} \asymp \ell(n^2) \quad a.s$$

Therefore, we finally find that

$$\frac{K_{n,D}}{K_n} = \frac{\sum_{d \geq D} K_{n,d} - \sum_{d \geq D+1} K_{n,d}}{K_n} \rightarrow 0 \quad a.s$$

B Gibbs sampler

As mentioned in the main text, the observed graph can be directed or undirected, binary or count, and can have missing entries we would like to predict. Denote by B the observed graph. Here we describes the steps of a Gibbs algorithm with stationary distribution

$$p(K_n, (\tilde{r}_k, \tilde{v}_{1:n,k})_{k=1,\dots,K_n}, \theta \mid B).$$

Notice that observing the full matrix $B = A$ corresponds to a weighted and directed graph with no missing entry. Let $\mathcal{I}$ denote the set of all possible edges. In the directed case, $\mathcal{I} = \{(i, j) \mid 1 \leq i, j \leq n\}$ and on the undirected case $\mathcal{I} = \{(i, j) \mid 1 \leq i \leq j \leq n\}$. We say that (i, j) is not observed if we don't know the value of $A_{i,j}$. Remark that (i, j) can be observed and still $A_{i,j} = 0$. Denote $\mathcal{O}$ the set of all observed entries and $\mathcal{O}^c = \mathcal{I} \setminus \mathcal{O}$, the set on unobserved entry. For all unobserved entry $(i, j) \in \mathcal{O}^c$, set $B_{i,j} = -1$

Additionally, to deal with the unknown number of active communities K_n , we use auxiliary slice variables $s_{i,j}$ for all $(i, j) \in \mathcal{I}$, details are given in the following paragraphs. Denote s the smallest non-zero slice variable $s_{i,j}$ for $(i, j) \in \mathcal{I}$. By definition of the slice variables, $\tilde{r}_k \geq s$ for all $k = 1, \dots, K_n$. Let

$$\bar{G} = \sum_k r_k \delta_{v_{1:n,k}} \mathbb{1}_{r_k \geq s} := \sum_{k=1}^{\bar{K}_n} \bar{r}_k \delta_{\bar{v}_{1:n,k}}$$

be the CRM corresponding to the set of active or inactive communities with weight $r_k \geq s$, of (almost surely finite) cardinality $\bar{K}_n \geq K_n$. Denote $\bar{Z}_{ijk} \geq 0$ the associated community interactions, and $\bar{Z}_k = (\bar{Z}_{ijk})$.

B.1 Directed graph

For each observed pair $(i, j) \in \mathcal{O}$, we define the slice variable as

$$s_{ij} | (\tilde{r}_k, \bar{Z}_{ijk})_{k=1,\dots,K_n} \sim \text{Unif} \left(0, \min_{\{k \mid \bar{Z}_{ijk} \geq 1\}} \tilde{r}_k \right) \quad (36)$$

if $A_{ij} > 0$ and $s_{ij} = 0$ otherwise. For each non observed entry $(i, j) \in \mathcal{O}^c$, we define $s_{i,j}$ by (38) if $\{k \mid \bar{Z}_{ijk} \geq 1\} \neq \emptyset$ and

$$s_{ij} | (\tilde{r}_k, \bar{Z}_{ijk})_{k=1,\dots,K_n} \sim \text{Unif}(0, 1) \quad (37)$$

otherwise.

B.1.1 Gibbs sampler step 1 for weighted graph on observed entries

Updating $(\bar{Z}_k)_{k=1,\dots,K_n} \mid (s, \bar{G}), \theta, B$ on observed entries indexes

We sample $(\bar{Z}_l)_{l=1,\dots,\bar{K}_n}$ associated to all atoms of $\bar{G}$ and keep only the non empty communities. For every $(i, j) \in \mathcal{O}$ such that $A_{i,j} > 0$, define the random variable $m_{ij} = \min_{\{l \mid \bar{Z}_{ijl} \geq 1\}} \tilde{r}_l$. Then, writing the joint distribution it comes that independently for every such (i, j) ,

$$\mathbb{P}((\bar{Z}_{ijl})_{l=1,\dots,\bar{K}_n} \mid (s, G^+), \theta, B_{ij}) \propto \prod_{i,j} \frac{1}{m_{ij}} \mathbb{1}_{s_{ij} < m_{ij}} \text{Mult}((\bar{Z}_{ijl})_l; B_{ij}, (\bar{p}_{ijl})_l)$$

where Mult is the multinomial distribution and $\bar{p}_{ijl} = \frac{r_l v_{il} v_{jl}}{\sum_{t=1}^{\bar{K}_n} r_t v_{it} v_{jt}}$. Let $p_{ijl} = r_l v_{il} v_{jl}$. To simplify the notations,

let us suppose that the atoms of $\bar{G}$ are in decreasing order. Remark that the indexing of $\bar{Z}$ is different from the one of $\bar{Z}$, the second corresponding to the one of the truncated random measure. For each observed edge (i, j) independently, we can proceed in 4 phases for this step.

1. Sample m_{ij} from the locations of $\bar{G}$ such that $\mathbb{P}(m_{ij} = r_L) \propto \frac{(\sum_{l=1}^L p_{ijl})^{B_{ij}} - (\sum_{l=1}^{L-1} p_{ijl})^{B_{ij}}}{r_L} \mathbb{1}_{s_{ij} < r_L}$.
2. For $l > L$, set $\bar{Z}_{ijl} = 0$
3. Sample $\bar{Z}_{ijL} \sim \text{tBin}(B_{ij}, \frac{p_{ijL}}{\sum_{l=1}^L p_{ijl}})$, where tBin is the zero truncated binomial distribution
4. Sample $(\bar{Z}_{ij1}, \dots, \bar{Z}_{ijL-1}) \sim \text{Mult}(B_{ij} - \bar{Z}_{ijL}, (\frac{p_{ijl}}{\sum_{t=1}^{L-1} p_{ijt}})_{l \leq L-1})$

B.1.2 Gibbs sampler step 1 for unweighted graph on observed entries

In this setting we observe a binary matrix $B_{ij} = 1_{A_{ij} > 0}$. Then the first step of the Gibbs sampler is modified and becomes:

Updating $(\tilde{Z}_k)_{k=1, \dots, K_n} | (s, \bar{G}), \theta, B$ on observed entries indexes

For each observed edge $(i, j) \in \mathcal{O}$ independently do

1. Sample m_{ij} from the locations of $\bar{G}$ such that

$$\mathbb{P}(m_{ij} = r_L) \propto \frac{e^{\sum_{k=1}^L p_{ijl}} - e^{\sum_{k=1}^{L-1} p_{ijk}}}{r_L} \mathbb{1}_{s_{ij} < r_L}$$

Suppose $m_{ij} = r_L$

2. For $l > L$, set $\tilde{Z}_{ijl} = 0$
3. Sample $\tilde{Z}_{ijL} \sim \text{tPoisson}(p_{ijL})$, where tPoisson is the zero truncated Poisson distribution
4. For $l < L$, sample $\tilde{Z}_{ijl} \sim \text{Poiss}(p_{ijl})$

B.1.3 Gibbs sampler step 1 on unobserved entries

For each unobserved entry $(i, j) \in \mathcal{O}^c$, knowing s_{ij} , we define $L_0 = \max\{k \mid r_k > s_{ij}\}$.

1. Draw $1_{A_{ij}=0}$, which is a Bernoulli with parameter

$$p = \frac{1}{1 + \sum_{L=1}^{L_0} \frac{e^{\sum_{k=1}^L p_{ijk}} - e^{\sum_{k=1}^{L-1} p_{ijk}}}{r_L}}$$

2. If $A_{i,j} \neq 0$, then use subsection B.1.2. Otherwise, set all counts of that entry to zero

B.2 Undirected graph

In the undirected graph, we suppose that for $i \neq j$, $B_{ij} = A_{ij} + A_{ji}$ and $B_{ii} = A_{ii}$. Besides, in this setting we actually don't need to sample $\tilde{Z}_{ijk}$ for all (i, j, k) but only $\tilde{Z}_{ijk} + \tilde{Z}_{jik}$. For each observed pair $(i, j) \in \mathcal{O}$, we define the slice variable as

$$s_{ij} | (\tilde{r}_k, \tilde{Z}_{ijk} + \tilde{Z}_{jik})_{k=1, \dots, K_n} \sim \text{Unif} \left(0, \min_{\{k \mid \tilde{Z}_{ijk} + \tilde{Z}_{jik} \geq 1\}} \tilde{r}_k \right) \quad (38)$$

if $B_{ij} > 0$ and $s_{ij} = 0$ otherwise. For each non observed entry $(i, j) \in \mathcal{O}^c$, we define s_{ij} by (38) if $\{k \mid \tilde{Z}_{ijk} + \tilde{Z}_{jik} \geq 1\} \neq \emptyset$ and

$$s_{ij} | (\tilde{r}_k, \tilde{Z}_{ijk} + \tilde{Z}_{jik})_{k=1, \dots, K_n} \sim \text{Unif}(0, 1) \quad (39)$$

otherwise. Then Step 2 and 3 remain unchanged. For step 1, simply replace p_{ijk} by $2p_{ijk}$ for $i \neq j$.

B.3 Proofs for the Gibbs sampler step 1

B.3.1 Weighted graph

Here will give the posterior distribution of the count matrices and show that

$$(\tilde{v}_k, \tilde{Z}_k)_{k=1, \dots, K_n} | (s, \bar{G}), \theta, B \stackrel{\text{dist}}{=} (\tilde{v}_k, \tilde{Z}_k)_{k=1, \dots, K_n} | s, G, \theta, B$$

In order to do so, we derive the RHS posterior distribution. Let us first notice that given G , sampling the non zero counts and corresponding locations is equivalent to sampling $(Z_k)_k$ for $k \in \mathbb{N}$. As stated previously, we can treat each edge (i, j) independently. Therefore, we sample the sequence $(Z_{ijk})_k$. Here we suppose that the communities come with decreasing activity order. Let the random variable $L = \max\{k \mid Z_{ijk} > 0\}$ (supposing that the $(r_k)_k$ are decreasing). And let $p_{ijk} = r_k v_{ik} v_{jk}$

$$\begin{aligned}
\mathbb{P}((Z_{ijk})_k | s, G, \theta, A_{ij}) &\propto \mathbb{P}((Z_{ijk})_k | G, \theta, A_{ij}) \times \mathbb{P}(s_{ij} | (Z_{ijk})_k, G, \theta, A_{ij}) \\
&\propto \mathbb{1}_{\sum_k Z_{ijk} = A_{ij}} \frac{A_{ij}!}{\prod_{k=1}^L Z_{ijk}!} \prod_{k=1}^L p_{ijk}^{Z_{ijk}} \times \frac{1}{r_L} \mathbb{1}_{s_{ij} < r_L} \\
&\propto \frac{(\sum_{k=1}^L p_{ijk})^{A_{ij}} (1 - (\frac{\sum_{k=1}^{L-1} p_{ijk}}{\sum_{k=1}^L p_{ijk}})^{A_{ij}})}{r_L} \mathbb{1}_{s_{ij} < r_L} \\
&\times \mathbb{1}_{1 \leq Z_{ijL} \leq A_{ij}} \frac{A_{ij}!}{Z_{ijL}! (A_{ij} - Z_{ijL})!} \frac{(\frac{p_{ijL}}{\sum_{k=1}^L p_{ijk}})^{Z_{ijL}} (\frac{\sum_{k=1}^{L-1} p_{ijk}}{\sum_{k=1}^L p_{ijk}})^{A_{ij} - Z_{ijL}}}{1 - (\frac{\sum_{k=1}^{L-1} p_{ijk}}{\sum_{k=1}^L p_{ijk}})^{A_{ij}}} \\
&\times \mathbb{1}_{\sum_{k=1}^{L-1} Z_{ijk} = A_{ij} - Z_{ijL}} \frac{(A_{ij} - Z_{ijL})!}{\prod_{k=1}^{L-1} Z_{ijk}!} \prod_{k=1}^{L-1} (\frac{p_{ijk}}{\sum_{k=1}^{L-1} p_{ijk}})^{Z_{ijk}}
\end{aligned}$$

This shows how we can sample in three steps these variables. Let us remark that the second part corresponds to the distribution of a zero truncated binomial and that the third part corresponds to the distribution of a multinomial. We also notice that only the elements of $\bar{G}$ are actually needed.

B.3.2 Unweighted Graph

We proceed similarly for the unweighted graph

$$\begin{aligned}
\mathbb{P}((Z_{ijk})_k | s, G, \theta, B_{ij} = 1) &\propto \mathbb{P}((Z_{ijk})_k | G, \theta, B_{ij} = 1) \times \mathbb{P}(s_{ij} | (Z_{ijk})_k, G, \theta, B_{ij} = 1) \\
&\propto \mathbb{1}_{Z_{ijL} \neq 0} \prod_{k=1}^L \frac{p_{ijk}^{Z_{ijk}}}{Z_{ijk}!} e^{-p_{ijk}} \times e^{\sum_{k=1}^L p_{ijk}} \times \frac{1}{r_L} \mathbb{1}_{s_{ij} < r_L} \\
&\propto \frac{e^{\sum_{k=1}^L p_{ijk}} - e^{\sum_{k=1}^{L-1} p_{ijk}}}{r_L} \mathbb{1}_{s_{ij} < r_L} \\
&\times \mathbb{1}_{Z_{ijL} \neq 0} \frac{1}{1 - e^{-p_{ijL}}} \frac{p_{ijL}^{Z_{ijL}}}{Z_{ijL}!} e^{-p_{ijL}} \\
&\times \mathbb{1}_{Z_{ij(L+1)}, \dots = 0} \prod_{k=1}^{L-1} \frac{p_{ijk}^{Z_{ijk}}}{Z_{ijk}!} e^{-p_{ijk}}
\end{aligned}$$

B.3.3 Prediction

Here we show how to update the missing entries we try to predict. Let us recall that for a predicted count, if it is positive, we define the slice variable as previously. However, if the count is equal to zero, then the slice variable is simply uniform over $[0, 1]$. Now let $L_0 = \max\{k \mid r_k > s_{ij}\}$

$$\begin{aligned}
\mathbb{P}((Z_{ijk})_k | s, G, \theta) &\propto \mathbb{P}((Z_{ijk})_k | G, \theta) \times \mathbb{P}(s_{ij} | (Z_{ijk})_k, G, \theta) \\
&\propto \mathbb{1}_{Z_{ij(L_0+1)}, \dots = 0} e^{\sum_{k=1}^{L_0} p_{ijk}} \prod_{k=1}^{L_0} \frac{p_{ijk}^{Z_{ijk}}}{Z_{ijk}!} e^{-p_{ijk}} \times (\mathbb{1}_{A_{ij} \neq 0} \frac{1}{r_L} \mathbb{1}_{s_{ij} < r_L} + \mathbb{1}_{A_{ij} = 0})
\end{aligned}$$

Now let

$$f((Z_{ijk})_k) = \mathbb{1}_{Z_{ij(L_0+1)}, \dots = 0} e^{\sum_{k=1}^{L_0} p_{ijk}} \prod_{k=1}^{L_0} \frac{p_{ijk}^{Z_{ijk}}}{Z_{ijk}!} e^{-p_{ijk}} \times (\mathbb{1}_{A_{ij} \neq 0} \frac{1}{r_L} \mathbb{1}_{s_{ij} < r_L} + \mathbb{1}_{A_{ij} = 0})$$

Using B.3.2, it comes that

$$\mathbb{E}(f((Z_{ijk})_k) \mid A_{ij} \neq 0) = e^{\sum_{k=1}^{L_0} p_{ijk}} \sum_{L=1}^{L_0} \frac{e^{\sum_{k=1}^L p_{ijk}} - e^{\sum_{k=1}^{L-1} p_{ijk}}}{r_L}$$

Besides,

$$\mathbb{E}(f((Z_{ijk})_k) \mid A_{ij} = 0) = f(0) = 1$$

Therefore, here we proceed in two steps, first we sample the binomial $\mathbb{1}_{A_{ij}=0}$ with parameter

$$p = \frac{1}{1 + \sum_{k=1}^{L_0} p_{ijk} \sum_{L=1}^{L_0} \frac{e^{\sum_{k=1}^L p_{ijk}} - e^{\sum_{k=1}^{L-1} p_{ijk}}}{r_L}}$$

Then, conditioning on the event $A_{ij} \neq 0$, we use B.3.2 to proceed.

B.4 Proof for the Gibbs step 2

Here we show how we can update the parameters $\theta = (\kappa, \sigma, \tau, \alpha, \beta)$ using a Metropolis-Hastings update. First, let us derive the posterior distribution of the hyperparameters.

We write $G = G' + \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}$ where G' is the non observed part. And we note $\overline{G'}$ the restriction of G' to the locations which intensity is larger than $\min s$.

$$\begin{aligned} p(\theta \mid (s, \overline{G}), \tilde{v}, \tilde{Z}) &\propto p(\theta \mid \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}, s, \overline{G'}, \tilde{Z}) \\ &\propto p(\theta, \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}, s, \overline{G'}, \tilde{Z}) \\ &\propto p(\theta, \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}, \tilde{Z}) p(s, \overline{G'} \mid \theta, \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}, \tilde{Z}) \end{aligned}$$

Now let us derive consider the first part

$$\begin{aligned} p(\theta, \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}, \tilde{Z}) &\propto p(\theta) p(\sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c} \mid \theta) p(\tilde{Z} \mid \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}) \\ &\propto p(\theta) p(\sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c} \mid \theta) \\ &\propto p(\theta) e^{-\Psi(n)} \prod_{c=1}^K \rho_{\kappa, \sigma, \tau}(\tilde{r}_c) f_{\alpha, \beta}(\tilde{v}_c) \end{aligned}$$

Now let us consider the second part

$$\begin{aligned} p(s, \overline{G'} \mid \theta, \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}, \tilde{Z}) &\propto p(s \mid \sum_{c=1}^K \tilde{r}_c \delta_{\tilde{v}_c}, \tilde{Z}) p(\overline{G'} \mid \theta, s) \\ &\propto p(\overline{G'} \mid \theta, s) \\ &\propto e^{-\Psi'(\min s, n)} \prod_k \rho_{\kappa, \sigma, \tau}(r'_k) f_{\alpha, \beta}(v'_k) \end{aligned}$$

where (r'_k) and (v'_k) are respectively the intensities and locations of $\overline{G'}$

Let

$$\pi(\theta) = e^{-\Psi(n) - \Psi'(\min s, n)} \prod_{t=1}^T \mu_\theta(r_t, v_t),$$

where we are taking the product over the T atoms and jumps of $\overline{G}$ and

$$\Psi'(s, n) = \int_{r>s, v} e^{-r|v|^2} \prod_{i=1}^n (f(v_i) dv_i) \rho(r) dr.$$

The posterior satisfies $p(\theta \mid (s, \overline{G})) \propto p(\theta) \pi(\theta)$. With our particular choice of distribution of the CRM, the multivariate integrals are reduced to one dimensional integrals, which makes the algorithm tractable. Indeed, we find that

$$\begin{aligned} \Psi(n) + \Psi'(\min s, n) &= \frac{\kappa}{\sigma} \int_0^{+\infty} (\tau + \varsigma^2)^\sigma \left[\frac{\sigma}{\Gamma(1-\sigma)} \Gamma(-\sigma, (\tau + \varsigma^2) \min s) + 1 \right] f_{n\alpha, \beta}(\varsigma) d\varsigma - \frac{\kappa \tau^\sigma}{\sigma} \\ \Psi(n) &= \begin{cases} \frac{\kappa}{\sigma} \int_0^{+\infty} (\tau + \varsigma^2)^\sigma f_{n\alpha, \beta}(\varsigma) d\varsigma - \frac{\kappa \tau^\sigma}{\sigma}, & \text{if } \sigma > 0 \text{ or } \sigma < 0 \\ \kappa \int_0^{+\infty} \log(\tau + \varsigma^2) f_{n\alpha, \beta}(\varsigma) d\varsigma - \kappa \log \tau, & \text{if } \sigma = 0 \end{cases} \end{aligned}$$

and

$$\Psi'(\min s, n) = \begin{cases} \frac{\kappa}{\Gamma(1-\sigma)} \int_0^{+\infty} (\tau + \varsigma^2)^\sigma \Gamma(-\sigma, (\tau + \varsigma^2) \min s) f_{n\alpha, \beta}(\varsigma) d\varsigma, & \text{if } \sigma > 0 \text{ or } \sigma < 0 \\ \kappa \int_0^{+\infty} \int_{(\tau + \varsigma^2) \min s}^{+\infty} \frac{e^{-r}}{r} dr f_{n\alpha, \beta}(\varsigma) d\varsigma, & \text{if } \sigma = 0 \end{cases}$$

We use the following priors:

$$\begin{aligned} 1 - 2\sigma &\sim \text{Gamma}(a_\sigma, b_\sigma) \\ \kappa &\sim \text{Gamma}(a_\kappa, b_\kappa) \\ \tau &\sim \text{Gamma}(a_\tau, b_\tau) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\ \beta &\sim \text{Gamma}(a_\beta, b_\beta) \end{aligned}$$

And proposals

$$\begin{aligned} 1 - 2\tilde{\sigma}|\sigma &\sim \text{Lognormal}(\log(1 - 2\sigma), \Sigma_\sigma) \\ \tilde{\kappa}|\kappa &\sim \text{Lognormal}(\log \kappa, \Sigma_\kappa) \\ \tilde{\tau}|\tau &\sim \text{Lognormal}(\log \tau, \Sigma_\tau) \\ \tilde{\alpha}|\alpha &\sim \text{Lognormal}(\log \alpha, \Sigma_\alpha) \\ \tilde{\beta}|\beta &\sim \text{Lognormal}(\log \beta, \Sigma_\beta) \end{aligned}$$

We find that

$$\begin{aligned} \log \frac{p(\tilde{\theta})q(\theta|\tilde{\theta})}{p(\theta)q(\tilde{\theta}|\theta)} &= a_\sigma \log \frac{1 - 2\tilde{\sigma}}{1 - 2\sigma} + 2b_\sigma(\tilde{\sigma} - \sigma) \\ &\quad + a_\kappa \log \frac{\tilde{\kappa}}{\kappa} - b_\kappa(\tilde{\kappa} - \kappa) \\ &\quad + a_\tau \log \frac{\tilde{\tau}}{\tau} - b_\tau(\tilde{\tau} - \tau) \\ &\quad + a_\alpha \log \frac{\tilde{\alpha}}{\alpha} - b_\alpha(\tilde{\alpha} - \alpha) \\ &\quad + a_\beta \log \frac{\tilde{\beta}}{\beta} - b_\beta(\tilde{\beta} - \beta) \end{aligned}$$

And

$$\begin{aligned} \log \frac{\pi(\tilde{\theta})}{\pi(\theta)} &= \Psi_\theta(n) + \Psi'_\theta(\min s, n) - \Psi_{\tilde{\theta}}(n) - \Psi'_{\tilde{\theta}}(\min s, n) \\ &\quad + T \log \frac{\tilde{\kappa}}{\kappa} - T \log \frac{\Gamma(1 - \tilde{\sigma})}{\Gamma(1 - \sigma)} - nT \log \frac{\Gamma(\tilde{\alpha})}{\Gamma(\alpha)} \\ &\quad - (\tilde{\sigma} - \sigma) \sum_t \log r_t - (\tilde{\tau} - \tau) \sum_t r_t \\ &\quad + (\tilde{\alpha} - \alpha) \sum_{t,i} \log v_{t,i} - (\tilde{\beta} - \beta) \sum_{t,i} v_{t,i} \\ &\quad + nT(\tilde{\alpha} \log \tilde{\beta} - \alpha \log \beta) \end{aligned}$$

B.5 Sampling from the inhomogeneous CRM

In this section we show how we can sample from the inhomogeneous CRM G' with measure:

$$\mu'(dr, dv) = e^{-r(\sum_{i=1}^n v_i)^2} \rho_{\kappa, \tau, \sigma}(r) \left[\prod_{i=1}^n f_{\alpha, \beta}(v_i) dv_i \right] dr$$

Let us recall that $f_{\alpha, \beta}$ is the gamma pdf and $\rho_{\kappa, \tau, \sigma}$ the GGP intensity. From Section 4, we know that if we make the following change of variables $(v_1, \dots, v_n) \mapsto (\varsigma = \sum_i v_i, \nu_1 = v_1/\varsigma, \dots, \nu_n = v_n/\varsigma)$, we get

$$\mu'(dr, d\varsigma, d\nu) = e^{-r\varsigma^2} \rho_{\kappa, \tau, \sigma}(r) f_{n\alpha, \beta}(\varsigma) d\varsigma dr \text{Dir}(d\nu_1, \dots, d\nu_n; \boldsymbol{\alpha})$$

Hence, we can sample independently (r, ς) and ν . From one hand, $(\nu_1, \dots, \nu_n)$ is sampled from a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha, \dots, \alpha)$. On the other hand, the total sum ς and the intensity r are sampled from

$$\begin{aligned} \mu(r, \varsigma) &= e^{-r\varsigma^2} \rho_{\kappa, \tau, \sigma}(r) f_{n\alpha, \beta}(\varsigma) \\ &= \frac{\kappa}{\Gamma(1 - \sigma)} e^{-r(\varsigma^2 + \tau)} r^{-1 - \sigma} f_{n\alpha, \beta}(\varsigma) \end{aligned}$$

Now, to reduce the problem to sampling from a homogeneous CRM, let us consider the change of variable $(r, \varsigma) \mapsto (\bar{r} = r[\tau + \varsigma^2], s)$ which determinant is $\tau + \varsigma^2$. We find finally that

$$\mu(\bar{r}, \varsigma) = \frac{\kappa}{\Gamma(1-\sigma)} e^{-\bar{r}} \bar{r}^{-1-\sigma} (\varsigma^2 + \tau)^\sigma f_{n\alpha, \beta}(\varsigma)$$

Besides, since $\bar{r} \geq r\tau \ \forall(i, j)$, we only need to sample the points such that $\bar{r} \geq \tau \min s$. Therefore, since $\tau > 0$, we sample a finite number of atoms. Then, we only keep the points such as $r > \min s$. Finally, let us notice that in our setting, even with $\sigma \leq 0$

$$\int_{\varsigma} (\varsigma^2 + \tau)^\sigma f_{n\alpha, \beta}(\varsigma) d\varsigma = \frac{\sigma}{\kappa} \Psi(n) + \tau^\sigma$$

Therefore, we first sample the jumps from the levy measure

$$\rho(\bar{r}) = \frac{\sigma \Psi(n) + \kappa \tau^\sigma}{\Gamma(1-\sigma)} e^{-\bar{r}} \bar{r}^{-1-\sigma} \mathbb{1}_{\bar{r} \geq \tau \min s}$$

using adaptive thinning (Favaro and Teh, 2013). Then, we sample ς with pdf $\propto (\varsigma^2 + \tau)^\sigma f_{n\alpha, \beta}(\varsigma)$ using rejection sampling.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Nonnegative Bayesian nonparametric factor models with completely random measures
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Under major revision for Statistics and Computing journal

Student Confirmation

Student Name:	Fadhel Ayed		
Contribution to the Paper	I am the first author of this paper; I have contributed in deriving the theoretical results, implementing the inference scheme and running the experiments.		
Signature	Mohamed El Fadhel Ayed	Date	12/07/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor François CARON		
Supervisor comments			
Signature		Date	14/07/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

3

Statistical Models with double power-law behavior

Beyond the Chinese Restaurant and Pitman-Yor processes: Statistical Models with double power-law behavior

Fadhel Ayed^{*1} Juho Lee^{*12} François Caron¹

Abstract

Bayesian nonparametric approaches, in particular the Pitman-Yor process and the associated two-parameter Chinese Restaurant process, have been successfully used in applications where the data exhibit a power-law behavior. Examples include natural language processing, natural images or networks. There is also growing empirical evidence suggesting that some datasets exhibit a two-regime power-law behavior: one regime for small frequencies, and a second regime, with a different exponent, for high frequencies. In this paper, we introduce a class of completely random measures which are doubly regularly-varying. Contrary to the Pitman-Yor process, we show that when completely random measures in this class are normalized to obtain random probability measures and associated random partitions, such partitions exhibit a double power-law behavior. We present two general constructions and discuss in particular two models within this class: the beta prime process (Broderick et al. (2015, 2018)) and a novel process called generalized BFRY process. We derive efficient Markov chain Monte Carlo algorithms to estimate the parameters of these models. Finally, we show that the proposed models provide a better fit than the Pitman-Yor process on various datasets.

1. Introduction

Power-law distributions appear to arise in a wide range of contexts, including natural languages, natural images or networks. For example, the empirical distribution of the word frequencies in natural languages is well approximated

by a power-law distribution, an observation attributed to Zipf (1935). That is, the frequency $f_{(k)}$ of the k th most frequent word in a corpus satisfies, within some range

$$f_{(k)} \simeq Ck^{-\xi}$$

where C is some constant and $\xi > 0$ is the power-law exponent which is typically close to 1 for natural languages. These empirical findings have motivated the development of numerous generative models that can reproduce this power-law behavior; see the reviews of (Mitzenmacher, 2004) and (Newman, 2005).

Amongst these generative models, Bayesian nonparametric hierarchical models based on infinite-dimensional random measures have been successfully used to capture the power-law behavior of various datasets. Applications include natural language processing (Goldwater et al., 2006; Teh, 2006; Wood et al., 2009; Mochihashi et al., 2009; Sato & Nakagawa, 2010), natural image segmentation (Sudderth & Jordan, 2009) or network analysis (Caron, 2012; Caron & Fox, 2017; Crane & Dempsey, 2018; Cai et al., 2016). A very popular model is the Pitman-Yor (PY) process (Pitman, 1995; Pitman & Yor, 1997; Pitman, 2006), an infinite-dimensional random probability measure whose properties induce a power-law behavior. It admits two parameters ($0 \leq \alpha < 1, \theta > -\alpha$). The PY random probability measure is almost surely discrete, with weights $\pi_{(1)} \geq \pi_{(2)} \geq \dots$ following the so-called two-parameter Poisson-Dirichlet distribution $PD(\alpha, \theta)$ (Pitman & Yor, 1997). For $\alpha > 0$, the random weights satisfy

$$\pi_{(k)} \sim k^{-1/\alpha} S \text{ almost surely as } k \rightarrow \infty$$

where S is a random variable. That is, small weights asymptotically follow a power-law distribution whose exponent is controlled by the parameter α . The PY process also enjoys tractable alternative constructions via the two-parameter Chinese restaurant process or the stick-breaking construction which explains its great popularity amongst models with similar properties. Other popular infinite-dimensional random measures that have been used for their similar power-law properties include the stable Indian buffet process (Teh & Gorur, 2009) or the generalized gamma process (Hougaard, 1986; Brix, 1999).

^{*}Equal contribution ¹Department of Statistics, University of Oxford, Oxford, United Kingdom ²AITRICS, Seoul, Republic of Korea. Correspondence to: Fadhel Ayed <fadhel.ayed@stats.ox.ac.uk>.

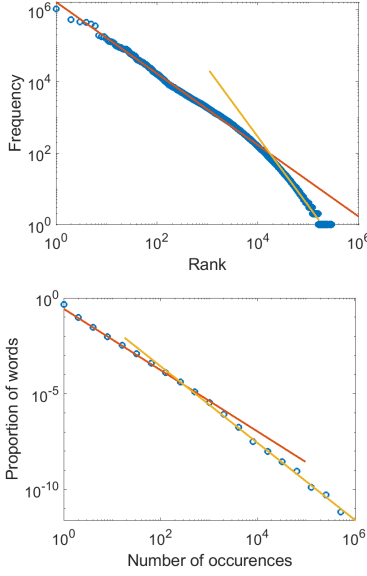


Figure 1. (Top) Ranked word frequencies from the American National Corpus (circles) and power-law fit (straight lines). (Bottom) Proportion of words with a given number of occurrences for the same dataset (circles) and power-law fit (straight lines).

Double power-law in empirical data. There is a growing empirical evidence that some datasets may exhibit a double power-law regime when the sample size is large enough. Examples include word frequencies in natural languages (Ferrer i Cancho & Solé, 2001; Montemurro, 2001; Gerlach & Altmann, 2013; Font-Clos et al., 2013), Twitter rates and retweet distributions (Bild et al., 2015), or degree distributions in social (Csányi & Szendrői, 2004), communication (Seshadri et al., 2008) or transportation networks (Paleari et al., 2010). In the case of word frequencies for example, it is conjectured that high frequency words approximately follow a power-law with Zipfian exponent approximately equal to 1, while the low frequency words follow a power-law with a higher exponent. An illustration is given in Figure 1, which shows the word frequencies of about 300,000 words from the American National Corpus¹.

In this paper, we introduce a class of completely random measures (CRMs), named doubly regularly-varying CRMs. We show that, when a random measure in this class is normalized to obtain a random probability measure P , and one repeatedly samples from P , the resulting frequencies exhibit a double power-law behavior. Informally, the ranked frequencies satisfy

$$f_{(k)} \simeq \begin{cases} C_1 k^{-1/\tau} & \text{for small rank } k \\ C_2 k^{-1/\alpha} & \text{for large rank } k \end{cases} \quad (1)$$

¹<http://www.anc.org/data/anc-second-release/frequency-data/>

where $\tau > 0$, $\alpha \in (0, 1)$ and $C_1, C_2 > 0$. The above statement is made mathematically accurate later in the article. We describe two general constructions to obtain doubly regularly varying CRMs, and consider two specific models within this class: the beta prime process of Broderick et al. (2015; 2018) and a novel process named generalized BFRY process. We show how these two CRMs can be obtained from transformations of the generalized gamma and stable beta processes. We derive Markov chain Monte Carlo inference algorithms for these models, and show that such models provide a good fit compared to a Pitman-Yor process on text and network datasets.

2. Background on (normalized) completely random measures

CRMs, introduced by Kingman (1967), are important building blocks of Bayesian nonparametric models (Lijoi & Prünster, 2010). A homogeneous CRM on a Polish space Θ , without deterministic component nor fixed atoms, is almost surely (a.s.) discrete and takes the form

$$W = \sum_{k \geq 1} w_k \delta_{\theta_k} \quad (2)$$

where $(w_k, \theta_k)_{k \geq 1}$ are the points of a Poisson point process with mean measure $\rho(dw)H(d\theta)$. H is some probability distribution on Θ , and ρ is a Lévy measure on $(0, \infty)$. We write $W \sim \text{CRM}(\rho, H)$. A popular CRM is the generalized gamma process (GGP) (Hougaard, 1986; Brix, 1999) with Lévy measure

$$\rho_{\text{GGP}}(dw; \sigma, \zeta) = \frac{1}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\zeta w} dw \quad (3)$$

where $\sigma \in (0, 1)$ and $\zeta \geq 0$ or $\sigma \leq 0$ and $\zeta > 0$. The GGP admits as special case the gamma process ($\sigma = 0, \zeta > 0$) and the stable process ($\sigma \in (0, 1), \zeta = 0$). If

$$\int_0^\infty \rho(dw) = \infty \quad (4)$$

then the CRM is said to be infinite-activity: it has an infinite number of atoms and the weights satisfy $0 < W(\Theta) = \sum_{k=1}^\infty w_k < \infty$ a.s. We can therefore construct a random probability measure P by normalizing the CRM (Regazzini et al., 2003; Lijoi et al., 2007)

$$P = \frac{W}{W(\Theta)}. \quad (5)$$

We call P a normalized CRM (NCRM) and write $P \sim \text{NCRM}(\rho, H)$. The Pitman-Yor process with parameters $\theta \geq 0$ and $\alpha \in [0, 1)$ and distribution H , written $P \sim \text{PY}(\alpha, \theta, H)$ admits a representation as a (mixture of)

CRMs (Pitman & Yor, 1997, Proposition 21). If $\theta, \alpha > 0$ it is a mixture of normalized generalized gamma processes

$$\eta \sim \text{Gamma}\left(\frac{\theta}{\alpha}, \frac{1}{\alpha}\right) \quad (6)$$

$$P \mid \eta \sim \text{NCRM}(\eta \rho_{\text{GGP}}(\cdot; \alpha, 1), H) \quad (7)$$

and for $\theta = 0$, it is a normalized stable process

$$P \sim \text{NCRM}(\rho_{\text{GGP}}(\cdot; \alpha, 0), H). \quad (8)$$

Although this representation is more complicated than the usual stick-breaking or urn constructions of the PY, it will be useful later on when we will discuss its asymptotic properties. The above construction essentially tells us that the PY has the same asymptotic properties as the normalized GGP for $\theta > 0$ and the stable process for $\theta = 0$.

3. Doubly regularly varying CRMs

3.1. General definition

We first introduce a few definitions on regularly varying functions (Bingham et al., 1989).

Definition 3.1 (Slowly varying function) A positive function ℓ on $(0, \infty)$ is slowly varying at infinity if for all $c > 0$ $\ell(ct)/\ell(t) \rightarrow 1$ as $t \rightarrow +\infty$. Examples of slowly varying functions are constant functions, functions converging to a strictly positive constant, $(\log t)^a$ for any real a , etc.

Definition 3.2 (Regularly varying function) A positive function f on $(0, \infty)$ is said to be regularly varying at infinity with exponent $\xi \in \mathbb{R}$ if $f(x) = x^\xi \ell(x)$ where ℓ is a slowly varying function. Similarly, a function f is said to be regularly varying at 0 if $f(1/x)$ is regularly varying at infinity, that is $f(x) = x^{-\xi} \ell(1/x)$ for some $\xi \in \mathbb{R}$ and some slowly varying function ℓ .

Informally, regularly varying functions with exponent $\xi \neq 0$ behave asymptotically similarly to a “pure” power-law function $g(x) = x^\xi$.

A homogeneous CRM W on Θ with mean measure $\rho(dw)H(d\theta)$ is said to be doubly regularly varying if its tail Lévy intensity

$$\bar{\rho}(x) = \int_x^\infty \rho(dw) \quad (9)$$

is regularly varying at 0 and ∞ , that is

$$\bar{\rho}(x) \sim \begin{cases} x^{-\alpha} \ell_1(1/x) & \text{as } x \rightarrow 0 \\ x^{-\tau} \ell_2(x) & \text{as } x \rightarrow \infty \end{cases} \quad (10)$$

where $\alpha \in [0, 1]$, $\tau \geq 0$ and ℓ_1 and ℓ_2 are slowly varying functions. The CRM is said to be doubly power-law if it is doubly regularly varying with exponents $\alpha > 0$ and $\tau > 0$. Note that in this case, the CRM necessarily satisfies condition (4) and is therefore infinite activity.

3.2. Properties

In the following, let $w_{(1)} \geq w_{(2)} \geq \dots$ denote the ordered weights of the CRM. The first proposition states that, if the CRM is regularly varying at 0 with exponent $\alpha > 0$, the small weights asymptotically scale as a power-law (up to a slowly varying function). Its proof is given in the Supplementary material.

Proposition 1 A CRM, regularly varying at 0 with exponent $\alpha > 0$, satisfies

$$w_{(k)} \sim k^{-1/\alpha} \ell_1^*(k) \quad \text{as } k \rightarrow \infty \quad (11)$$

where ℓ_1^* is a slowly varying function whose expression, which depends on ℓ_1 and α , is given in the supplementary material.

The next proposition states that, if the CRM is regularly varying at infinity with $\tau > 0$ and the scaling factor of the Lévy measure is large, the CRM has a power-law behavior for large weights.

Proposition 2 [Kevei & Mason (2014, Theorem 1.2)] Consider a CRM with mean measure $\eta\rho(dw)H(d\theta)$, regularly varying at ∞ with $\tau > 0$. Then, for any $k_1, k_2 \geq 1$

$$\frac{w_{(k_1+k_2)}^\tau}{w_{(k_1)}^\tau} \xrightarrow{d} \text{Beta}(k_1, k_2) \quad \text{as } \eta \rightarrow \infty. \quad (12)$$

Note that Equation (12) indicates a power-law behavior with exponent $1/\tau$, as for large η and $k \gg 1$, $w_{(k)} \simeq w_{(1)} k^{-1/\tau}$.

GGP and stable process. The GGP with parameter $\zeta > 0$ is regularly varying at 0 with exponent $\alpha = \max(0, \sigma)$. Hence, it satisfies Proposition (1). However, the exponential decay of the tails of the Lévy measure implies that it is not regularly varying at ∞ . Large weights therefore decay exponentially fast. The stable process, which is a GGP with parameter $\zeta = 0$ and $\sigma \in (0, 1)$, is doubly regularly-varying with the same power-law exponent σ at 0 and ∞ . Hence, it satisfies Proposition 1. Additionally, Pitman & Yor (1997, Proposition 8) showed that the result of Proposition 2 holds non-asymptotically for the stable process. In particular, for all $k \geq 1$, $w_{(k+1)}/w_{(k)} \sim \text{Beta}(k\sigma, 1)$.

In Section 3.3, we describe two general constructions for obtaining doubly regularly varying CRMs. Then we describe two specific processes with doubly regularly varying tail Lévy measure where one can flexibly tune both exponents. In the rest of the paper, we assume that the Lévy measure ρ is absolutely continuous with respect to the Lebesgue measure, and use the same notation for its density $\rho(dw) = \rho(w)dw$.

3.3. Construction of doubly regularly varying CRMs

Scaled-CRM. A first way of constructing a doubly regularly varying CRMs is to consider a CRM, regularly varying

at 0, and to divide its weights by independent and identically distributed (iid) random variables, whose cumulative density function (cdf) is also regularly varying at 0. More precisely, let

$$W = \sum_{k \geq 1} \frac{w_{0k}}{z_k} \delta_{\theta_k} \quad (13)$$

where $(z_1, z_2, \dots)$ are strictly positive, continuous and iid random variables with cumulative density function $F_Z(z)$ and probability density function $f_Z(z)$, and

$$W_0 = \sum_{k \geq 1} w_{0k} \delta_{\theta_k} \sim \text{CRM}(\rho_0, H)$$

where $\bar{\rho}_0(x)$ and $F_Z(z)$ are both regularly varying functions at 0, that is, for some $\alpha \in (0, 1)$ and $\tau > \alpha$,

$$\bar{\rho}_0(x) \sim x^{-\alpha} \ell_1(1/x) \quad (14)$$

$$F_Z(z) \sim z^\tau \ell_2(1/z). \quad (15)$$

The random measure W is a CRM $W \sim \text{CRM}(\rho, H)$ where

$$\rho(w) = \int_0^\infty z f_Z(z) \rho_0(wz) dz.$$

The next proposition shows that W is doubly regularly varying.

Proposition 3 *Assume that $\bar{\rho}_0$ and F_Z verify Equations (14) and (15). Additionally, suppose that ρ_0 and f_Z are bounded on an interval $[a, b]$ for some $0 < a < b$ and that there exists $\beta > \tau$, such that $\mu_\beta = \int_0^\infty w^\beta \rho_0(w) dw < \infty$. Then the CRM W defined by Equation (13) is doubly regularly varying, with*

$$\bar{\rho}(x) \sim \begin{cases} \mathbb{E}(Z^{-\alpha}) x^{-\alpha} \ell_1(1/x) & \text{as } x \rightarrow 0 \\ \mu_\tau x^{-\tau} \ell_2(x) & \text{as } x \rightarrow \infty \end{cases}.$$

where Z is a random variable with cdf F_Z .

In Sections 3.4 and 3.5 we present two specific models constructed via a scaled GGP.

Discrete Mixture. An alternative to the scaled-CRM construction is to consider that the CRM is the sum of two CRMs, one regularly varying at 0 (hence infinite activity), the second one regularly varying at infinity. More precisely, consider the Lévy density

$$\rho(w) = \rho_0(w) + \beta f(w) \quad (16)$$

where ρ_0 is a Lévy measure, regularly varying at 0, and f is the probability density function of a random variable with power-law tails. That is ρ_0 satisfies (14) and

$$\int_x^\infty f(t) dt \sim x^{-\tau} \ell_2(x) \quad \text{as } x \rightarrow \infty.$$

If we additionally assume that $\bar{\rho}_0(x)$ has light tails at infinity (e.g. exponentially decaying tails), then the resulting CRM ρ is then doubly regularly varying and satisfies Equation (10). For example, one can take for ρ_0 the Lévy density (3) of a GGP, and for f the pdf of a Pareto, generalized Pareto or inverse gamma distribution.

3.4. Generalized BFRY process

Consider the Lévy density

$$\rho(w) = \frac{1}{\Gamma(1-\sigma)} w^{-1-\tau} \gamma(\tau-\sigma, cw) \quad (17)$$

where $\gamma(\kappa, x) = \int_0^x u^{\kappa-1} e^{-u} du$ is the lower incomplete gamma function and the parameters satisfy $\sigma \in (-\infty, 1)$, $\tau > \max(0, \sigma)$ and $c > 0$. We have

$$\bar{\rho}(x) \sim \frac{\Gamma(\tau-\sigma)}{\tau \Gamma(1-\sigma)} x^{-\tau} \quad (18)$$

as x tends to infinity and, for $\sigma > 0$,

$$\bar{\rho}(x) \sim \frac{c^{\tau-\sigma}}{\sigma(\tau-\sigma)\Gamma(1-\sigma)} x^{-\sigma} \quad (19)$$

as x tends to 0. When $\sigma \leq 0$, $\bar{\rho}(x)$ is a slowly varying function, with $\lim_{x \rightarrow 0} \bar{\rho}(x) = \infty$ if $\sigma = 0$ and $\lim_{x \rightarrow 0} \bar{\rho}(x) < \infty$ if $\sigma < 0$. $\bar{\rho}(x)$ therefore satisfies Equation (10) with $\alpha = \max(\sigma, 0)$. When $\sigma > 0$, it is doubly power-law with exponent $\sigma \in (0, 1)$ and $\tau > 0$.

The Lévy density (17) admits the following latent construction as a scaled-GGP. Note that

$$\rho(w) = \frac{c^{\tau-\sigma}}{\tau} \int_0^1 z \rho_{\text{GGP}}(wz; \sigma, c) f_Z(z) dz$$

where $f_Z(z) = \tau z^{\tau-1}$ is the probability density function of a Beta($\tau, 1$) random variable. We therefore have the hierarchical construction. For $k \geq 1$,

$$w_k = \frac{w_{0k}}{\beta_k}, \beta_k \sim \text{Beta}(\tau, 1).$$

where $(w_{0k})_{k \geq 1}$ are the points of a Poisson process with mean measure $c^{\tau-\sigma} / \tau \rho_{\text{GGP}}(w_0; \sigma, c) dw_0$.

The process is somewhat related to, and can be seen as a natural generalization of the BFRY distribution (Pitman & Yor, 1997; Winkel, 2005; Bertoin et al., 2006). The name was coined by Devroye & James (2014) after the work of Bertoin, Fujita, Roynette and Yor. This distribution has recently found various applications in machine learning (Lee et al., 2016; 2017). Taking $c = 1$, $\tau \in (0, 1)$ and $\sigma = \tau - 1 < 0$, we have

$$\rho(w) \propto w^{-\tau-1} (1 - e^{-w})$$

which corresponds to the unnormalized pdf of a BFRY random variable. The BFRY random variable admits a representation as the ratio of a gamma and beta random variable, and the stochastic process introduced in this section, which admits a similar construction, can be seen as a natural generalization of the BFRY distribution, and we call this process a generalized BFRY (GBFRY) process. In Section 4 of the supplementary material, we provide more details on the BFRY distribution and its generalization.

3.5. Beta prime process

Consider the Lévy density

$$\rho(w) = \frac{\Gamma(\tau - \sigma)}{\Gamma(1 - \sigma)} w^{-1-\sigma} (c + w)^{\sigma-\tau} \quad (20)$$

where $\sigma \in (-\infty, 1)$, $\tau > 0$ and $c > 0$. This density is an extension of the beta prime (BP) process, with an additional tuning parameter. This process was introduced by Broderick et al. (2015) and generalized by Broderick et al. (2018), as a conjugate prior for odds Bernoulli process. We have

$$\bar{\rho}(x) \sim \frac{\Gamma(\tau - \sigma)}{\tau \Gamma(1 - \sigma)} x^{-\tau} \quad (21)$$

as x tends to infinity and, for $\sigma > 0$,

$$\bar{\rho}(x) \sim \frac{c^{\sigma-\tau} \Gamma(\tau - \sigma)}{\sigma \Gamma(1 - \sigma)} x^{-\sigma} \quad (22)$$

as x tends to 0. When $\sigma \leq 0$, $\bar{\rho}(x)$ is a slowly varying function, with $\lim_{x \rightarrow 0} \bar{\rho}(x) = \infty$ if $\sigma = 0$ and $\lim_{x \rightarrow 0} \bar{\rho}(x) < \infty$ if $\sigma < 0$. $\bar{\rho}(x)$ therefore satisfies Equation (10) with $\alpha = \max(\sigma, 0)$. When $\sigma > 0$, it is doubly power-law with exponent $\sigma \in (0, 1)$ and $\tau > 0$.

The BP process is related to the stable beta process (Teh & Gorur, 2009) with Lévy density

$$\frac{\alpha \Gamma(\tau - \sigma)}{c^\tau \Gamma(1 - \sigma)} u^{-1-\sigma} (1 - u)^{\tau-1} \mathbb{1}_{u \in (0,1)},$$

via the transformation $w = \frac{cu}{1-u}$. Similarly to the generalized BFRY model, the beta prime process can also be obtained via a scaled GGP. Note that

$$\rho(w) = \Gamma(\tau) c^{-\tau} \int_0^\infty y \rho_{\text{GGP}}(wy; \sigma, c) f_Y(y) dy$$

where $f_Y(y) = \frac{c^\tau y^{\tau-1} e^{-cy}}{\Gamma(\tau)}$ is the density of a $\text{Gamma}(\tau, c)$ random variable. We therefore have the following hierarchical construction, for $k \geq 1$

$$w_k = \frac{w_{0k}}{\gamma_k}, \quad \gamma_k \sim \text{Gamma}(\tau, c)$$

where $(w_{0k})_{k \geq 1}$ are the points of a Poisson process with mean measure $c^{-\tau} \Gamma(\tau) \rho_{\text{GGP}}(w_0; \sigma, 1) dw_0$.

4. Normalized CRMs with double power-law

For some probability distribution H , Lévy measure ρ satisfying Equation (4) and $\eta > 0$, let

$$P = \frac{W}{W(\Theta)} \text{ where } W \sim \text{CRM}(\eta \rho, H)$$

and for $i = 1, \dots, n$, $X_i \mid P \stackrel{i.i.d.}{\sim} P$. As P is a.s. discrete, there will be repeated values within the sequence $(X_i)_{i \geq 1}$. Let $K_n \leq n$ be the number of unique values in $(X_1, \dots, X_n)$, and $m_{n,(1)} \geq m_{n,(2)} \geq \dots \geq m_{n,(K_n)}$ their ranked multiplicities. For $k = 1, \dots, K_n$, denote $f_{n,(k)} = \frac{m_{n,(k)}}{n}$ the ranked frequencies.

4.1. Double power-law properties

The following theorem provides a precise formulation of Equation (1) and shows that the ranked frequencies have a double power-law regime when the CRM is doubly regularly varying with strictly positive exponents.

Theorem 1 *The ranked frequencies satisfy*

$$(f_{n,(1)}, f_{n,(2)}, \dots) \rightarrow \left(\frac{w_{(1)}}{W(\Theta)}, \frac{w_{(2)}}{W(\Theta)}, \dots \right) \quad (23)$$

almost surely as n tends to infinity. If the CRM is regularly varying at 0 with exponent $\alpha > 0$ we have

$$\frac{w_{(k)}}{W(\Theta)} \sim W(\Theta)^{-1} k^{-1/\alpha} \ell_1^*(k) \text{ as } k \rightarrow \infty. \quad (24)$$

If the CRM is regularly varying at ∞ with exponent $\tau > 0$ we have, for any $k_1, k_2 \geq 1$

$$\frac{w_{(k_1+k_2)}^\tau}{w_{(k_1)}^\tau} \xrightarrow{d} \text{Beta}(k_1, k_2) \text{ as } \eta \rightarrow \infty. \quad (25)$$

Equation (23) in Theorem 1 follows from (Gnedin et al., 2007, Proposition 26). Equations (24) and (25) follow from Propositions 1 and 2. Instead of expressing the power-law properties in terms of the ranked frequencies, we can alternatively look at the asymptotic behavior of the number $K_{n,j}$ of elements with multiplicity $j \geq 1$, defined by

$$K_{n,j} = \sum_{k=1}^{K_n} \mathbb{1}_{m_{n,(k)}=j}. \quad (26)$$

Let $p_{n,j} = \frac{K_{n,j}}{K_n}$. Note that $\sum_{j \geq 1} p_{n,j} = 1$. The following is a corollary of Equation (24). It follows from Proposition 23 and Corollary 21 in (Gnedin et al., 2007).

Corollary 2 *If the CRM is regularly varying at 0 with exponent α , we have*

$$p_{n,j} \rightarrow p_j \text{ a.s. as } n \rightarrow \infty \quad (27)$$

where

$$p_j = \frac{\sigma \Gamma(j - \alpha)}{j! \Gamma(1 - \alpha)} \sim \frac{\alpha}{\Gamma(1 - \alpha)} \frac{1}{j^{1+\alpha}} \quad \text{for large } j.$$

Figure 2 shows some illustration of these empirical results for the GBFRY model.

Remark 1 *The GGP with parameter $\sigma > 0$ is regularly varying at 0, but not at infinity. Hence, the normalized GGP with $\zeta > 0$ and the related Pitman-Yor process with $\theta > 0$ satisfy Equation (24) and (27) but not (25), due to the exponentially decaying tails of the Lévy measure of the GGP. The normalized GGP with $\zeta = 0$, which is the same as the Pitman-Yor with $\theta = 0$, satisfies both equations, but with the same exponent $\sigma \in (0, 1)$, lacking the flexibility of the three models presented in Section 3.*

4.2. Posterior Inference

In this subsection, we briefly discuss the inference procedure for estimating the parameters of the normalized CRMs we introduced in Section 3. Additional details are provided in the supplementary material. Assume that the Lévy measure ρ is parameterised by some parameters ϕ we want to estimate, in particular the two power-law exponents. We write $\rho(w; \phi)$ to emphasize this, and let $p(\phi)$ be the prior density. The objective is to approximate the posterior density of the parameters given the ranked counts $p(\phi \mid (m_{n,(k)})_{k=1,\dots,K_n})$.

Parametrisation. Since we are working with normalized CRMs, multiplying W by any positive constant $\xi > 0$ gives the same random probability measure P . In particular, the normalized CRMs with Lévy densities $\rho(w)$ and $\tilde{\rho}(w) = \xi \rho(\xi w)$ have the same distribution. To avoid overparameterisation we set the parameter $c = 1$ in the GBFRY and BP processes, and estimate the parameter $\phi = (\sigma, \tau, \eta)$.

We introduce a latent variable $U \mid W \sim \text{Gamma}(n, W(\Theta))$. Using Proposition 3 of James et al. (2009) (see also Pitman (2003)) and Equation (2.2) of Pitman (2006)), the joint density is written as

$$p((m_{n,(k)})_{k=1,\dots,K_n}, u, \phi) \propto p(\phi) u^{n-1} e^{-\psi(u; \phi)} \prod_{k=1}^{K_n} \kappa(m_{n,(k)}, u; \phi) \quad (28)$$

where the normalizing constant only depends on n and the ranked counts, and

$$\psi(t; \phi) = \eta \int_0^\infty (1 - e^{-tw}) \rho(w; \phi) dw, \quad (29)$$

$$\kappa(m, t; \phi) = \eta \int_0^\infty w^m e^{-tw} \rho(w; \phi) dw. \quad (30)$$

If ψ and κ have analytic forms, one can derive a MCMC sampler to approximate the posterior by successively updating U and ϕ . Unfortunately, this is not the case for our models. For instance, in the generalized BFRY process case, we have

$$\psi(t; \phi) = \frac{\eta}{\sigma} \int_0^c ((y+t)^\sigma - y^\sigma) y^{\tau-\sigma-1} dy \quad (31)$$

$$\kappa(m, t; \phi) = \frac{\eta \Gamma(m - \sigma)}{\Gamma(1 - \sigma)} \int_0^c \frac{y^{\tau-\sigma-1}}{(y+t)^{m-\sigma}} dy. \quad (32)$$

We may resort to a numerical integration algorithm to approximate ψ as only one evaluation of this function is needed at each iteration. We could do the same for κ . However, this would require K_n numerical integrations at each step of the MCMC sampler, which is computationally prohibitive for large K_n . Instead, building on the construction of the generalized BFRY as a scaled generalized gamma process described in Section 4, we introduce a set of latent variables $Y = (Y_k)_{k=1,\dots,K_n}$ whose conditional density is written as

$$p(y_k \mid u, (m_{n,(k)})_{k=1,\dots,K_n}) \propto \frac{y_k^{\tau-\sigma-1}}{(y_k + u)^{m_{n,(k)}-\sigma}} \mathbb{1}_{0 < y_k < c},$$

and this gives the joint density

$$p((m_{n,(k)})_{k=1,\dots,K_n}, u, y, \phi) \propto p(\phi) u^{n-1} e^{-\psi(u; \phi)} \prod_{k=1}^{K_n} \frac{\eta \Gamma(m_{n,(k)} - \sigma) y_k^{\tau-\sigma-1}}{\Gamma(1 - \sigma) (y_k + u)^{m_{n,(k)}-\sigma}},$$

where the normalizing constant only depends on n and the ranked counts. Then we can alternate between updating ϕ and U via Metropolis-Hastings and updating Y via Hamiltonian Monte-Carlo (HMC) (Duane et al., 1987; Neal et al., 2011) to estimate the posterior. See the supplementary material for more details. A similar strategy can be used for the beta prime process.

5. Experiments

We run the algorithms described in Section 4.2 for the GBFRY and BP models. We fix $c = 1$ to avoid overparameterisation, as explained in Section 4.2. We use standard normal prior on $\log \eta$, $\log \tau$ and $\text{logit } \sigma$. The proposed models are compared to the normalized GGP with the same priors on η and σ and fixed $\zeta = 1$, and the PY process with standard normal prior on $\log \theta$ and $\text{logit } \alpha$. We also considered the discrete mixture construction described in Section 3.3 with ρ_0 taken to be a GGP, and f a Pareto or generalized Pareto distribution. While we were able to recover the parameters on simulated data, this model was underperforming on real data, and results are not reported. The codes to replicate our experiments can be found in <https://github.com/OxCSML-BayesNP/doublepowerlaw>.

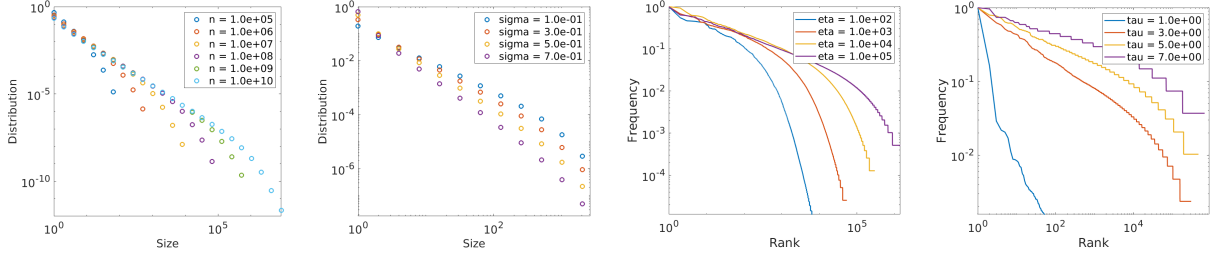


Figure 2. Simulated data from the normalized GBFRY model. Proportion of clusters of a given size for (First) $\eta = 4000, \tau = 3, \sigma = 0.2$ with varying n and (Second) $\eta = 4000, \tau = 3, n = 10^7$ with varying σ . Ordered frequencies, normalized by the largest one, for (Third) $n = 10^6, \sigma = 0.2, \tau = 3$ with varying η and (Fourth) $n = 10^6, \sigma = 0.2, \eta = 50000$ with varying τ .

We stress that the objective is to show that the proposed models provide a better fit than alternative models, not to test the double power-law assumption.

5.1. Synthetic data

We sample simulated datasets from the normalized GBFRY and the BP with parameters $\sigma = 0.1, \tau = 2, c = 1$ and $\eta = 4000$. We run the MCMC algorithm described in Section 4.2 with 100 000 iterations. The 95% credible intervals are $\sigma \in (0.09, 0.12), \tau \in (1.6, 2.2)$ for the BFRY and $\sigma \in (0.08, 0.11), \tau \in (1.8, 2.3)$ for the BP, indicating that the MCMC recovers true parameters. Trace plots are reported in the supplementary material.

5.2. Real data

We then consider five real datasets, four of which are word frequencies in natural languages, and the last is the out-degree distribution of a Twitter network. We first provide a description of the different datasets.

Word frequencies. Each dataset is composed of n words $X_1, \dots, X_n$, with $K_n \leq n$ unique words. The counts $m_{n,(k)}$ represent the number of occurrences of the k th most frequent word in the dataset. The first dataset is the written dataset of the American National Corpus² (ANC), composed of about 18 million word occurrences and 300 000 unique words. The second and third datasets are the words of a collection of most popular English books and French books, downloaded from the Project Gutenberg³. The English books dataset is composed of about 3 million words and 71 000 unique words, the French books of about 7 million words and around 135 000 unique words. The fourth dataset represents the words of a thousand papers from the NIPS conference. It contains about 2 million word occurrences and 68 000 unique words.

Twitter network. We consider a rank-1 edge-exchangeable model for directed multigraphs (Crane

Table 1. Average Kolmogorov-Smirnov divergence between the data and the posterior predictive. Lower is better.

Dataset	GBFRY	Beta Prime	GGP	PY
Englishbooks	0.072	0.041	0.12	0.12
Frenchbooks	0.064	0.032	0.11	0.11
NIPS1000	0.041	0.081	0.08	0.059
ANC	0.033	0.034	0.082	0.081
Twitter	0.10	0.047	0.25	0.26

& Dempsey, 2018; Cai et al., 2016). In this case, the atoms of W represent the nodes of the graph, and each directed edge (X_i, Y_i) from node X_i to node Y_i is sampled independently from $P \times P$. Note that when P is a Pitman-Yor process, the associated model corresponds to the urn-based Hollywood model of Crane & Dempsey (2018). Here, we only consider the out-degree distribution. Therefore, n represents the number of directed edges and $X_1, \dots, X_n$ the source nodes of the directed edges sampled from the normalized CRM P . $m_{n,(k)}$ corresponds to the k th largest out-degree in the network. We consider a subset of 25 millions tweets of August 2009 from Twitter (Yang & Leskovec, 2011). We construct a directed multigraph by adding an edge (X_i, Y_i) whenever user X_i mentions user Y_i (with @) in tweet i . The resulting graph contains about 4 millions edges and 300 000 source nodes.

Results

For each of the four models and each dataset, we approximate the posterior distribution of the parameters ϕ of the Lévy measure, and sample new datasets from the posterior predictive. The 95% credible intervals of the posterior predictive for the proportion of occurrences and ranked frequencies are reported in Figure 3 for the ANC dataset (plots for the other datasets are given in the supplementary material). As the results for the normalized GGP and PY are almost identical, we only show the plot for the PY model. As can clearly be seen from the posterior predictive plots, all models provide a good fit for low frequencies. However, the

²<http://www.anc.org/data/anc-second-release/frequency-data/>

³<http://www.gutenberg.org/>

Table 2. 95% posterior credible intervals of the power-law exponents.

Dataset	GBFRY		Beta Prime		GGP	PY
	σ	τ	σ	τ	σ	σ
Englishbooks	(0.351, 0.362)	(0.912, 0.980)	(0.345, 0.358)	(0.974, 1.078)	(0.416, 0.423)	(0.416, 0.423)
Frenchbooks	(0.368, 0.375)	(0.967, 1.039)	(0.363, 0.371)	(1.04, 1.175)	(0.407, 0.412)	(0.407, 0.412)
NIPS1000	(0.538, 0.545)	(1.338, 1.906)	(0.538, 0.545)	(1.541, 2.286)	(0.542, 0.548)	(0.542, 0.549)
ANC	(0.433, 0.438)	(0.998, 1.055)	(0.431, 0.436)	(1.09, 1.17)	(0.461, 0.465)	(0.461, 0.465)
Twitter	(0.282, 0.287)	(1.590, 1.600)	(0.099, 0.116)	(1.336, 1.411)	(0.272, 0.277)	(0.272, 0.277)

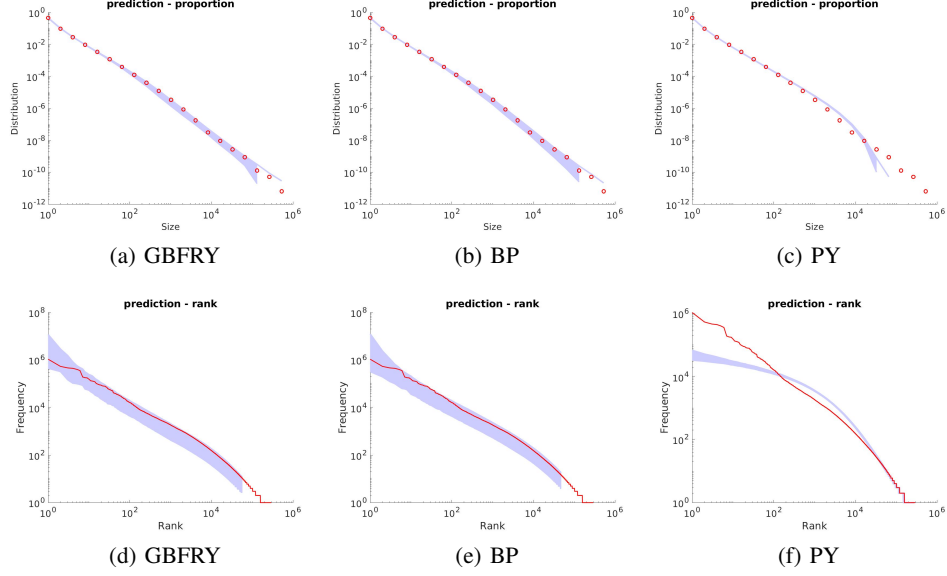


Figure 3. Results on the ANC dataset: 95% credible interval of the posterior predictive in blue, data in red. (Top) Proportion of occurrences of a given size. (Bottom) Ranked occurrences.

PY model (and similar the normalized GGP) fail to capture the power-law behavior for large frequencies. This behavior is better captured by the GBFRY and BP models. To illustrate quantitatively the comparison, we compute the average reweighted Kolmogorov-Smirnov divergence (Clauset et al., 2009) between the true data and the posterior predictive for each model, and report the results in Table 1. Finally, we report in Table 2 the 95% credible intervals of the parameters for each model and dataset. We can remark that to the exception of the NIPS dataset, we recover the Zipfian exponent $\tau = 1$ for large frequencies in text datasets.

6. Conclusion

In this paper we presented a novel class of random measures with double power-law behavior. We focused on the case of iid sampling from a normalized completely random measure. More generally, one could build on this class of models for other CRM-based constructions. In particular, it would be interesting to explore the asymptotic degree distribution when such models are used for random graph

models based on exchangeable point processes (Caron & Fox, 2017). Building hierarchical versions of such models as for the hierarchical Pitman-Yor process (Teh, 2006) would also be of interest. Finally, it would be useful to explore the connections between the models presented here and the two-stage urn process suggested by Gerlach & Altmann (2013) and investigate if other urn schemes could be derived that provably exhibit a double power-law behavior.

Acknowledgments Authors thank Valerio Perrone for providing NIPS dataset. JL and FC’s research leading to these results has received funding from European Research Council under the European Unions Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071 and from EPSRC under grant EP/P026753/1. FC acknowledges support from the Alan Turing Institute under EPSRC grant EP/N510129/1. JL acknowledges support from IITP grant funded by the Korea government(MSIT) (No.2017-0-01779, XAI) and Samsung Research Funding & Incubation Center under project number SRFC-IT1702-15.

References

- Bertoin, J., Fujita, T., Roynette, B., and Yor, M. On a particular class of self-decomposable random variables: the durations of Bessel excursions straddling independent exponential times. 2006.
- Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., and Wallach, D. S. Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1):4, 2015.
- Bingham, N. H., Goldie, C. M., and Teugels, J. L. *Regular variation*, volume 27. Cambridge university press, 1989.
- Brix, A. Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31(4):929–953, 1999.
- Broderick, T., Mackey, L., Paisley, J., and Jordan, M. I. Combinatorial clustering and the beta negative binomial process. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):290–306, 2015.
- Broderick, T., Wilson, A. C., and Jordan, M. I. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 24(4B):3181–3221, 2018.
- Cai, D., Campbell, T., and Broderick, T. Edge-exchangeable graphs and sparsity. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4249–4257. Curran Associates, Inc., 2016.
- Caron, F. Bayesian nonparametric models for bipartite graphs. In *Advances in neural information processing systems*, 2012.
- Caron, F. and Fox, E. B. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366, 2017.
- Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
- Crane, H. and Dempsey, W. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, 113(523):1311–1326, 2018.
- Csányi, G. and Szendrői, B. Structure of a large social network. *Physical Review E*, 69(3):036131, 2004.
- Devroye, L. and James, L. On simulation and properties of the stable law. *Statistical methods & applications*, 23(3):307–343, 2014.
- Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.
- Ferrer i Cancho, R. and Solé, R. V. Two regimes in the frequency of words and the origins of complex lexicons: Zipf’s law revisited. *Journal of Quantitative Linguistics*, 8(3):165–173, 2001.
- Font-Clos, F., Boleda, G., and Corral, A. A scaling law beyond Zipf’s law and its relation to Heaps’ law. *New Journal of Physics*, 15(9):093033, 2013.
- Gerlach, M. and Altmann, E. G. Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3(2):021006, 2013.
- Gnedin, A., Hansen, B., and Pitman, J. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability surveys*, 4:146–171, 2007.
- Goldwater, S., Johnson, M., and Griffiths, T. L. Interpolating between types and tokens by estimating power-law generators. In *Advances in neural information processing systems*, pp. 459–466, 2006.
- Hougaard, P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986.
- James, L. F., Lijoi, A., and Prünster, I. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.
- Kevei, P. and Mason, D. M. The limit distribution of ratios of jumps and sums of jumps of subordinators. *ALEA*, 11(2):631–642, 2014.
- Kingman, J. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.
- Lee, J., James, L. F., and Choi, S. Finite-dimensional BFRY priors and variational Bayesian inference for power law models. In *Advances in Neural Information Processing Systems*, pp. 3162–3170, 2016.
- Lee, J., Heaukulani, C., Ghahramani, Z., James, L. F., and Choi, S. Bayesian inference on random simple graphs with power law degree distributions. In *International Conference on Machine Learning*, pp. 2004–2013, 2017.
- Lijoi, A. and Prünster, I. Models beyond the Dirichlet process. *Bayesian nonparametrics*, 28(80):3, 2010.
- Lijoi, A., Mena, R. H., and Prünster, I. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.

- Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.
- Mochihashi, D., Yamada, T., and Ueda, N. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 100–108. Association for Computational Linguistics, 2009.
- Montemurro, M. A. Beyond the Zipf Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3-4):567–578, 2001.
- Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.
- Newman, M. E. J. Power laws, Pareto distributions and Zipf’s law. *Contemporary physics*, 46(5):323–351, 2005.
- Paleari, S., Redondi, R., and Malighetti, P. A comparative study of airport connectivity in China, Europe and US: which network provides the best service to passengers? *Transportation Research Part E: Logistics and Transportation Review*, 46(2):198–210, 2010.
- Pitman, J. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2): 145–158, 1995.
- Pitman, J. Poisson-Kingman partitions. *Lecture Notes-Monograph Series*, pp. 1–34, 2003.
- Pitman, J. *Combinatorial Stochastic Processes: Ecole d’Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.
- Pitman, J. and Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pp. 855–900, 1997.
- Regazzini, E., Lijoi, A., and Prünster, I. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2): 560–585, 2003.
- Sato, I. and Nakagawa, H. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 673–682. ACM, 2010.
- Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C., and Leskovec, J. Mobile call graphs: beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 596–604. ACM, 2008.
- Sudderth, E. B. and Jordan, M. I. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in neural information processing systems*, pp. 1585–1592, 2009.
- Teh, Y. W. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 985–992. Association for Computational Linguistics, 2006.
- Teh, Y. W. and Gorur, D. Indian buffet processes with power-law behavior. In *Advances in neural information processing systems*, pp. 1838–1846, 2009.
- Winkel, M. Electronic foreign-exchange markets and passage events of independent subordinators. *Journal of applied probability*, 42(1):138–152, 2005.
- Wood, F., Archambeau, C., Gasthaus, J., James, L., and Teh, Y. W. A stochastic memoizer for sequence data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1129–1136. ACM, 2009.
- Yang, J. and Leskovec, J. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 177–186. ACM, 2011.
- Zipf, G. The psycho-biology of language: an introduction to dynamic philology. 1935.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Beyond the Chinese Restaurant and Pitman-Yor processes: Statistical Models with double power-law behavior
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Proceedings of the International Conference on Machine Learning 2019

Student Confirmation

Student Name:	Fadhel Ayed		
Contribution to the Paper	For this paper, I worked alongside François Caron and Juho Lee. On the methodological part (how to construct doubly regularly varying CRMs), it is difficult to clearly separate the role that each of us had. I also contributed to proving the different results and conducted the experimentations for half of the models.		
Signature	Mohamed El Fadhel Ayed	Date	12/07/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor F Caron		
Supervisor comments			
Signature		Date	14/07/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

4

The Normal-Generalised Gamma-Pareto process

The Normal-Generalised Gamma-Pareto process: A novel pure-jump Lévy process with flexible tail and jump-activity properties

Fadhel Ayed
Department of Statistics, University of Oxford
and
Juho Lee
AI Graduate School, KAIST
and
François Caron
Department of Statistics, University of Oxford

June 15, 2020

Abstract

Pure-jump Lévy processes are popular classes of stochastic processes which have found many applications in finance, statistics or machine learning. In this paper, we propose a novel family of self-decomposable Lévy processes where one can control separately the tail behavior and the jump activity of the process, via two different parameters. Crucially, we show that one can sample exactly increments of this process, at any time scale; this allows the implementation of likelihood-free Markov chain Monte Carlo algorithms for (asymptotically) exact posterior inference. We use this novel process in Lévy-based stochastic volatility models to predict the returns of stock market data, and show that the proposed class of models leads to superior predictive performances compared to classical alternatives.

Keywords: Stochastic Volatility models, Power-law, Regular variation, Ornstein-Uhlenbeck, Bayesian inference, Pseudo-marginal Markov chain Monte Carlo

1 Introduction

Pure-jump Lévy processes are a flexible class of stochastic processes that have found a wide range of applications, including scalable Markov chain Monte Carlo (Şimşekli, 2017), tracking (Zhang and Paisley, 2018) or the analysis of phylogenetic traits (Landis et al., 2012; Landis and Schraiber, 2017). Finance is probably the main domain of application, as it is widely accepted that asset prices contain jumps, and such models have been used as building blocks of complex dynamic models of asset or option prices (Madan et al., 1998; Barndorff-Nielsen and Shephard, 2001; Carr et al., 2002; Cont and Tankov, 2004; Huang and Wu, 2004; Jing et al., 2012).

Let $(X_t)_{t \geq 0}$ be a real-valued pure-jump Lévy process. The process is said to have *heavy, power-law tails* if, for any $t, \Delta > 0$,

$$\Pr(|X_{t+\Delta} - X_t| > x) \stackrel{x \rightarrow \infty}{\sim} \Delta C_1 x^{-2\tau} \quad (1)$$

for some power-law exponent $\tau > 0$ and some constant $C_1 > 0$. That is, for large x , the survival function of the increments approximately behaves as a power function.

Many financial time series, such as historical asset returns, exhibit heavy-tails (Cont and Tankov, 2004, Section 7.3). Empirical evidence seems to indicate that the returns have nonetheless finite variance, hence corresponding to a power-law exponent $\tau > 1$. Starting from the early work of Mandelbrot (1963) with the stable distribution, various infinite-divisible distributions, closely related to Lévy processes, have been proposed to capture power-law tails. Examples include the student t (Blattberg and Gonedes, 1974) or Pareto (Champagnat et al., 2013) distributions; other models with (non power-law) semi-heavy tails such as the normal inverse Gaussian (Barndorff-Nielsen, 1997), generalized hyperbolic (Eberlein et al., 1998) and tempered stable distributions (Cont et al., 1997; Carr et al., 2002) have also been proposed; see (Cont and Tankov, 2004, Section 7.3) for a review.

Another quantity of interest of the Lévy process is the Blumenthal-Gettoor (BG) index $\beta \in [0, 2]$, also known as fractional order. It is defined by

$$\beta = \inf \left\{ r > 0 \mid \sum_{i \geq 1} |J_i|^r \mathbf{1}_{\theta_i \leq t} < \infty \right\} \quad (2)$$

where $(J_i, \theta_i)_{i \geq 1}$ is the set of jump values and jump time. The BG index measures the level of activity of the jumps: as the value of β increases, small jumps tend to become more and more frequent. It is also related to the smoothness properties of the time series (Cont and Tankov, 2004, Section 7.3) and therefore provides interpretable information on the process and its properties. A number of papers have proposed and analysed (model-free) estimators of this index (Aït-Sahalia and Jacod, 2009; Belomestny, 2010; Woerner, 2011; Belomestny and Panov, 2013). Some Lévy processes, such as the normal-tempered stable or tempered stable processes,

can capture the whole range $[0, 2)$ via a tuning parameter; other processes, such as the variance gamma ($\beta = 0$), normal-inverse Gaussian, student t or generalised hyperbolic ($\beta = 1$), have a fixed BG index.

For a pure-jump Lévy process, a typical way to obtain a given power-law exponent and BG index is to assume the regular variation of the tail intensity of the Lévy measure ν characterising the Lévy process, such that

$$\int_{|s| \geq x} \nu(ds) \stackrel{x \rightarrow \infty}{\sim} x^{-2\tau} C_1 \quad \text{and} \quad \int_{|s| \geq x} \nu(ds) \stackrel{x \rightarrow 0}{\sim} x^{-\beta} \ell(1/x)$$

for some slowly varying function ℓ , that is such that $\lim_{t \rightarrow \infty} \ell(ct)/\ell(t) = 1$ for all $c > 0$. While many Lévy measures have been proposed in the literature, no tractable model is able to capture both the whole range of power-law exponent $\tau > 0$ and BG index $\beta \in [0, 2)$. Normal-tempered stable process for example (which includes as special case the variance-gamma and normal-inverse Gaussian) capture the whole range of the β index but have light tails. The class of generalised hyperbolic processes can capture heavy tails, but has a fixed BG index equal to 1. The normal-stable process can capture both heavy tails and different BG indices, but the same parameter controls both properties, and the process has infinite variance.

In this paper, we introduce a novel four-parameter pure-jump Lévy process, called *normal generalised gamma-Pareto* (NGGP) process, with the following properties.

- The model can capture power-law, heavy tails with a single parameter $\tau > 0$; for $\tau > 1$, the process has finite variance;
- Another parameter $\sigma \in (-\infty, 1)$ controls the BG index and therefore the activity of the jumps, with $\beta = \max(0, 2\sigma)$. The process is finite-activity for $\sigma < 0$; it is infinite-activity for $\sigma \geq 0$; it is of bounded variation if $\sigma < 1/2$ and of unbounded variation if $\sigma \in [1/2, 1)$;
- The other two parameters respectively are inverse scale and time scale parameters;
- One can sample exactly from the distribution of the increments, at any time scale; this enables the use of likelihood-free Markov chain Monte Carlo methods for inference;
- The distribution of the increments of the Lévy process is self-decomposable.

The NGGP process is obtained via Brownian subordination, using the subordinator introduced by Ayed et al. (2019) for modeling power-law properties of text data. We derive a number of properties of the NGGP process and use the proposed model to predict the stock prices of some financial assets. We consider two Lévy based stochastic volatility models: an exponential Lévy model, and an Ornstein-Uhlenbeck based model Lévy-driven stochastic volatility model (Barndorff-Nielsen and Shephard, 2001). We show that, compared to other Lévy processes, the proposed model is both able to capture the heavy-tail and small-jump behaviours.

The article is organised as follows. In Section 2, we introduce the generalised gamma-Pareto subordinator, its properties, and the associated subordinated Brownian process. In Section 3 we describe two Lévy process based stochastic volatility models, and describe how to perform

asymptotically exact posterior inference under our Lévy process with both models. In Section 4 we present experimental results on the modelling of stock prices and show that our model provides a very good fit to the data and good predictive performances compared to classical alternatives.

Notations. We use the notation $a_n \stackrel{n \rightarrow \infty}{\sim} b_n$ for $\lim_{n \rightarrow \infty} a_n/b_n = 1$. For a random variable X , the notation $X \sim F$ indicates that X has distribution F . $\text{Gamma}(a, b)$ denotes the gamma distribution with shape parameter a and inverse scale parameter b . $\text{Poisson}(\lambda)$ denotes the standard Poisson distribution with rate λ .

2 The NGGP process

2.1 Generalised gamma subordinator

A generalised gamma (GG) subordinator (almost surely increasing Lévy process) $(Y_t)_{t \geq 0}$ has Lévy intensity (Hougaard, 1986; Aalen, 1992; Brix, 1999)

$$\rho_{\text{GG}}(w; \eta, \sigma, c) = \frac{\eta}{\Gamma(1 - \sigma)} w^{-1 - \sigma} e^{-cw}, \quad w > 0 \quad (3)$$

where $\eta > 0$ and $\sigma \in (-\infty, 1)$, $c > 0$ or $c = 0, \sigma \in (0, 1)$. The subordinator is finite-activity for $\sigma < 0$ and infinite-activity if $\sigma \in [0, 1)$. It admits as special cases the gamma process ($\sigma = 0$), inverse-Gaussian process ($\sigma = 1/2$) and stable process ($c = 0$). When $\sigma > 0$, the process belongs to the general family of tempered stable processes introduced by Rosiński (2007) and some authors referred to this process simply as a tempered stable process (Barndorff-Nielsen and Shephard, 2002; Liang and Li, 2015). Ignoring the drift term, Y_t has Laplace transform

$$\mathbb{E}[e^{-\vartheta Y_t}] = \exp(-t\psi_{\text{GG}}(\vartheta; \eta, \sigma, c)) \quad (4)$$

where the Laplace exponent is given by

$$\psi_{\text{GG}}(\vartheta; \eta, \sigma, c) = \begin{cases} \frac{\eta}{\sigma} [(\vartheta + c)^\sigma - c^\sigma] & \sigma \neq 0 \\ \eta \log(1 + \vartheta/c) & \sigma = 0. \end{cases} \quad (5)$$

Y_t is said to have the generalised gamma distribution with parameters $(\eta t, \sigma, c)$, and we write $Y_t \sim \text{GG}(\eta t, \sigma, c)$. For $\sigma = 0$, $Y_t \sim \text{Gamma}(\eta t, c)$, while for $\sigma < 0$, Y_t is a compound Poisson-gamma distribution with

$$Y_t \stackrel{d}{=} \sum_{k=1}^{K_t} Y_{t,k}$$

where $K_t \sim \text{Poisson}(\eta t \frac{c^\sigma}{-\sigma})$ and $Y_{t,k} \sim \text{Gamma}(-\sigma, c)$ for $k = 1, \dots, K_t$. For $\sigma > 0$, Y_t is an exponentially tilted stable random variable, for which exact samplers exist (Devroye, 2009; Hofert, 2011).

2.2 Generalised gamma-Pareto subordinator

2.2.1 Definition

Let $Z = (Z_t)_{t \geq 0}$ be a subordinator with no drift and Lévy intensity

$$\rho(w) = \frac{\eta}{c^\tau \Gamma(1-\sigma)} w^{-1-\tau} [\gamma(\tau - \sigma + 1, cw) + (cw)^{\tau-\sigma} e^{-cw}], \quad w > 0 \quad (6)$$

where $\eta > 0$, $c > 0$, $\sigma \in (-\infty, 1)$, $\tau > 0$ and $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$ is the lower incomplete gamma function. For $\tau > \sigma$, using the identity (41) in Appendix A, the Lévy intensity takes the simpler form

$$\rho(w) = \frac{\eta(\tau - \sigma)}{c^\tau \Gamma(1-\sigma)} w^{-1-\tau} \gamma(\tau - \sigma, cw) \quad (7)$$

which is the form in which Ayed et al. (2019) introduced the process, with a slightly different parameterisation. The Lévy intensity admits the following representation as a mixture of generalised gamma process

$$\rho(w) = \int_0^\infty u^{-1} \rho_{\text{GG}}\left(\frac{w}{u}; \frac{\eta(\tau - \sigma)}{\tau c^\sigma}, \sigma, c\right) f_U(u) du \quad (8)$$

where $f_U(u) = \tau u^{-1-\tau} \mathbf{1}_{u \geq 1}$ is the probability density function of a Pareto random variable Pareto($\tau, 1$) with support $[1, \infty)$ and power-law exponent $\tau > 0$, and ρ_{GG} is the Lévy intensity of a GG subordinator, defined in Equation (3). We will thereafter refer to the subordinator with intensity (6) as a *Generalised Gamma-Pareto* (GGP) process. For $x > 0$, let

$$\bar{\rho}(x) = \int_x^\infty \rho(w) dw \quad (9)$$

be the tail Lévy intensity. For $t > 0$, we denote F_{Z_t} the cumulative distribution function of the random variable Z_t , with Laplace transform

$$\mathbb{E}[e^{-\vartheta Z_t}] = e^{-t\psi(\vartheta)}$$

where ψ is the Laplace exponent which is given by, noting that $\gamma(\tau - \sigma + 1, cw) = w^{\tau-\sigma+1} \int_0^c u^{\tau-\sigma} e^{-wu} du$,

$$\psi(\vartheta) = \int_0^\infty (1 - e^{-w\vartheta}) \rho(w) dw = \frac{\eta}{c^\tau} \left[\frac{c^\tau}{\tau} - \int_0^c (u + \vartheta)^{\sigma-1} u^{\tau-\sigma} du + \frac{c^{\tau-\sigma}}{\sigma} ((\vartheta + c)^\sigma - c^\sigma) \right]. \quad (10)$$

Z_t is said to have GGP($t\eta, \sigma, \tau, c$) distribution.

2.2.2 Properties

We derive here a number of properties of the Lévy process Z and of the GGP distribution.

Positive stable process. The positive stable process with Lévy intensity $\frac{\eta}{\Gamma(1-\sigma)} w^{-1-\sigma}$ is obtained as a special case when $c = 1$, $\sigma = \tau \in (0, 1)$.

Scaled GG process. Let

$$Y_t = \sum_{i \geq 1} W_i \mathbb{1}_{\theta_i \leq t}$$

where $\{(W_i, \theta_i)\}_{i \geq 1}$ are the jump sizes and times of a GG subordinator. Then, for $\tau > \sigma$, the representation (8) implies that

$$Z_t \stackrel{d}{=} \sum_{i \geq 1} W_i U_i \mathbb{1}_{\theta_i \leq t}$$

where $U_i \sim \text{Pareto}(\tau, 1)$. The jump sizes of the GGP subordinator are obtained by scaling the jumps of a GG subordinator with independent Pareto random variables.

Moments and cumulants. We have $\mathbb{E}[Z_t^m] < \infty$ for $m < \tau$ and $\mathbb{E}[Z_t^m] = \infty$ otherwise. For $1 \leq m < \tau$, the m 's cumulant is given by

$$\kappa_m(Z_t) = t \int_0^\infty w^m \rho(w) dw = \frac{t\eta(\tau - \sigma)}{c^m(\tau - m)}.$$

In particular, for $\tau > 1$

$$\mathbb{E}[Z_t] = \frac{t\eta(\tau - \sigma)}{c(\tau - 1)}$$

and for $\tau > 2$,

$$\text{var}(Z_t) = \frac{t\eta(\tau - \sigma)}{c^2(\tau - 2)}.$$

Inverse scale parameter. If the intensity ρ is of the form (6) for some parameters (η, σ, τ, c) , then $\rho(w/c)/c$ is also of the form (6) with parameters $(\eta, \sigma, \tau, 1)$. c is therefore an inverse scale parameter, and if $Z_t \sim \text{GGP}(t\eta, \sigma, \tau, c)$ then $cZ_t \sim \text{GGP}(t\eta, \sigma, \tau, 1)$.

Activity of the jumps and BG index. The Lévy intensity (6) satisfies $\int_0^\infty \rho(w) dw = \infty$ if $\sigma \geq 0$ and the subordinator is therefore infinite-activity. If $\sigma < 0$, $\int_0^\infty \rho(w) dw < \infty$ and it is finite-activity. More precisely, as noted by Ayed et al. (2019), the tail Lévy intensity is regularly varying at 0

$$\bar{\rho}(x) \stackrel{x \rightarrow 0}{\sim} \ell(1/x)x^{-\alpha} \tag{11}$$

where $\alpha = \max(0, \sigma)$ is the BG index, and the slowly varying function ℓ is defined by

$$\ell(t) = \begin{cases} \frac{\eta}{c^\sigma \sigma \Gamma(1-\sigma)} & \sigma > 0 \\ \eta \log(t) & \sigma = 0 \\ \frac{\eta(\tau - \sigma)}{-\sigma \tau} & \sigma < 0. \end{cases} \tag{12}$$

It follows from the Abelian theorem (Gnedin et al., 2007, Proposition 17) that the Laplace exponent satisfies

$$\psi(\vartheta) \stackrel{\vartheta \rightarrow \infty}{\sim} \Gamma(1 - \alpha) \vartheta^\alpha \ell(\vartheta). \tag{13}$$

For $\sigma = \alpha \in (0, 1)$, the cumulative distribution function F_{Z_t} satisfies (Bingham et al., 1989, Theorem 8.2.2. p. 341)

$$-\log F_{Z_t}(z) \stackrel{z \rightarrow 0}{\sim} (1 - \alpha)\alpha^{\alpha/(1-\alpha)} \left(\frac{\eta t}{c^\alpha \alpha} \right)^{1/(1-\alpha)} z^{-\alpha/(1-\alpha)}. \quad (14)$$

Additionally, for small increments, we have, for all $\vartheta \geq 0$

$$\mathbb{E}[e^{-\vartheta \frac{cZ_t}{(\eta t/\alpha)^{1/\alpha}}}] \stackrel{t \rightarrow 0}{\rightarrow} e^{-\vartheta^\alpha}$$

hence $\frac{cZ_t}{(\eta t/\alpha)^{1/\alpha}}$ tends in distribution to a positive stable random variable with parameter $\alpha \in (0, 1)$ as $t \rightarrow 0$. For $\sigma < 0$, the cdf has a discontinuity at 0 with

$$F_{Z_t}(0) - F_{Z_t}(0_-) = \Pr(Z_t = 0) = e^{-t \frac{\eta(\tau-\sigma)}{-\sigma\tau}}.$$

Heavy tails and power-law behaviour. As noted by Ayed et al. (2019), the tail Lévy intensity is regularly varying at infinity, with power-law exponent τ . We have

$$\bar{\rho}(x) \stackrel{x \rightarrow \infty}{\sim} \frac{\eta\Gamma(\tau - \sigma + 1)}{\tau c^\tau \Gamma(1 - \sigma)} x^{-\tau}. \quad (15)$$

It follows from (Bingham et al., 1989, Theorem 8.2.1. page 341) that the survival function $1 - F_{Z_t}(z) = \Pr(Z_t > z)$ satisfies

$$1 - F_{Z_t}(z) \stackrel{z \rightarrow \infty}{\sim} \frac{\eta t \Gamma(\tau - \sigma + 1)}{\tau c^\tau \Gamma(1 - \sigma)} z^{-\tau} \quad (16)$$

and the increments have heavy, power-law tails with exponent $\tau > 0$.

Simulation of the increments. First note that if $\sigma < 0$, the subordinator is a compound Poisson process with jump rate $\frac{\eta(\tau-\sigma)}{-\sigma\tau}$ and jumps being GBFRY distributed (see Section B in the Appendix) with parameters $(-\sigma, \tau, c)$. We therefore have

$$Z_t \stackrel{d}{=} \sum_{j=1}^{K_t} G_{t,j} U_{t,j}$$

where $K_t \sim \text{Poisson}(t \frac{\eta(\tau-\sigma)}{-\sigma\tau})$, $G_{t,j} \sim \text{Gamma}(-\sigma, c)$ and $U_{t,j} \sim \text{Pareto}(\tau, 1)$, $j = 1, \dots, K_n$ are independent random variables. Consider now the case $\sigma \geq 0$. The Lévy measure admits the two-components mixture representation

$$\rho(w) = \frac{\eta c^{-\sigma}}{\Gamma(1 - \sigma)} w^{-1-\sigma} e^{-cw} + \frac{\eta}{c^\tau \Gamma(1 - \sigma)} w^{-1-\tau} \gamma(\tau + 1 - \sigma, cw). \quad (17)$$

The first component of the mixture representation (17) is the Lévy intensity of GG subordinator, while the second component is the intensity of a finite-activity GGP subordinator. We can therefore write

$$Z_t \stackrel{d}{=} Z_{t,1} + Z_{t,2} \quad (18)$$

where $Z_{t,1} \sim \text{GG}(\frac{\eta t}{c^\sigma}, \sigma, c)$ is an exponentially tilted stable random variable for which exact samplers exist (Devroye, 2009; Hofert, 2011), and

$$Z_{t,2} \stackrel{d}{=} \sum_{j=1}^{\tilde{K}_t} \tilde{G}_{t,j} \tilde{U}_{t,j}$$

where $\tilde{K}_t \sim \text{Poisson}(\frac{\eta t}{\tau})$, $\tilde{G}_{t,j} \sim \text{Gamma}(1 - \sigma, c)$ and $\tilde{U}_{t,j} \sim \text{Pareto}(\tau, 1)$.

Self-decomposability. The random variable Z_t is self-decomposable if $\sigma \geq 0$. That is, for any $a \in (0, 1)$, there is $Z_t^{(a)}$ independent of Z_t such that

$$Z_t \stackrel{d}{=} aZ_t + Z_t^{(a)}.$$

The self-decomposable random variable Z_1 admits the representation

$$Z_1 \stackrel{d}{=} \int_0^\infty e^{-s} d\tilde{Z}_s \quad (19)$$

where $(\tilde{Z}_s)_{s \geq 0}$ is termed the background driving Lévy process corresponding to the self-decomposable random variable Z_1 . $\tilde{Z}_t$ has Lévy intensity

$$\tilde{\rho}(w) = -\rho(w) - w\rho'(w) = \frac{\eta\sigma}{c^\sigma\Gamma(1-\sigma)} w^{-1-\sigma} e^{-cw} + \frac{\eta\tau}{c^\tau\Gamma(1-\sigma)} w^{-1-\tau} \gamma(\tau - \sigma + 1, cw). \quad (20)$$

Importantly, for $\sigma = 0$, the background Lévy process is a finite-activity GGP process with intensity

$$\tilde{\rho}(w) = \frac{\eta\tau}{c^\tau} w^{-1-\tau} \gamma(\tau + 1, cw). \quad (21)$$

Proof. Let $k(w) = w\rho(w)$. Consider first that $\tau > \sigma \geq 0$. From equation (7), we have $k(w) \propto w^{-\sigma} \int_0^c u^{\tau-\sigma-1} e^{-wu} du$ which is non-increasing. Consider that $0 < \tau \leq \sigma$. From Equation (6), k takes the form $k(w) \propto w^{-\tau} g(w)$ where

$$g(w) = \gamma(\tau - \sigma + 1, cw) + (cw)^{\tau-\sigma} e^{-cw}.$$

As $g'(w) = (\tau - \sigma)c(cw)^{\tau-\sigma-1} e^{-cw} \leq 0$, it follows that k is monotone decreasing. Hence for all $\sigma \geq 0$, k is monotone decreasing; using Proposition 15.3 p.485 in (Cont and Tankov, 2004), we conclude that the process is therefore self-decomposable. ■

Interpretability of the parameters. In summary, each of the four parameters governs a different property of the GGP process.

- $\eta > 0$ is a time-scaling parameter: if $(Z_t)_{t \geq 0}$ is a GGP process with parameters $(1, \sigma, \tau, c)$, then $(Z_{\eta t})_{t \geq 0}$ is a GGP process with parameters (η, σ, τ, c) ;
- $c > 0$ is an inverse-scale parameter: if (Z_t) is a GGP process with parameters $(\eta, \sigma, \tau, 1)$, then (Z_t/c) is a GGP process with parameters (η, σ, τ, c) ;

- $\sigma \in (-\infty, 1)$ tunes the activity of the jumps; the process is finite-activity if $\sigma < 0$ and infinite-activity otherwise, with corresponding BG index $\beta = \max(\sigma, 0)$;
- τ is the power-law exponent, controlling the tails of the distribution, with $\Pr(Z_t > z) \stackrel{z \rightarrow \infty}{\sim} Cz^{-\tau}$ for some constant C .

2.3 Normal GGP process

2.3.1 Definition

Let $Z = (Z_t)_{t \geq 0}$ be a GGP subordinator with no drift and Lévy intensity given by (6). Let $B = (B_t)_{t \geq 0}$ be a Brownian motion on $\mathbb{R}$, independent from Z . The normal generalised gamma Pareto (NGGP) Lévy process, taking values in $\mathbb{R}$, is defined via Brownian subordination by

$$X_t = B_{Z_t}.$$

For any $t > 0$, Let F_{X_t} denote the cumulative distribution function of the random variable X_t , with characteristic function (Cont and Tankov, 2004, Section 4.2)

$$\mathbb{E}[e^{i\lambda X_t}] = e^{t\Psi(\lambda)}$$

where the characteristic exponent is given by

$$\Psi(\vartheta) = -\psi(\vartheta^2/2) = \int_{\mathbb{R}} (e^{i\vartheta x} - 1)\nu(x)dx \quad (22)$$

where ψ is defined in Equation (10) and ν is a Lévy intensity on $\mathbb{R}$ defined by

$$\nu(x) = \int_0^\infty \frac{1}{\sqrt{2\pi w}} e^{-\frac{x^2}{2w}} \rho(w)dw.$$

For any $x > 0$, let

$$\bar{\nu}(x) = \int_{|s|>x} \nu(s)ds$$

denote the expected number of jumps of absolute value larger than x in an unit-length interval.

We also write $X_t \sim \text{NGGP}(\eta t, \sigma, \tau, c)$.

2.3.2 Properties

Most of the properties here follow from the properties of the subordinator. By construction, we have, for all t

$$X_t = \sqrt{Z_t}\epsilon_t \quad (23)$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$. $\sqrt{c}$ is therefore an inverse scale parameter, and if $X_t \sim \text{NGGP}(\eta t, \sigma, \tau, c)$ then $\sqrt{c}X \sim \text{NGGP}(\eta t, \sigma, \tau, 1)$.

Moments and cumulants. Let $1 \leq m < 2\tau$. For m odd, the m 'th raw moment and cumulant of X_t satisfy

$$\mathbb{E}[X_t^m] = \kappa_m(X_t) = 0. \quad (24)$$

For $2 \leq m < 2\tau$, m even,

$$\begin{aligned} \kappa_m(X_t) &= 2 \int_0^\infty \int_0^\infty \frac{x^m}{\sqrt{2\pi w}} e^{-\frac{x^2}{2w}} \rho(w) dw dx \\ &= \frac{2^{m/2}}{\sqrt{\pi}} \kappa_{m/2}(Z_t) \Gamma((m+1)/2) \\ &= \frac{t\eta(\tau-\sigma)2^{m/2}\Gamma((m+1)/2)}{c^{m/2}(\tau-m/2)\sqrt{\pi}} \end{aligned}$$

It follows, for $\tau > 1$, the Lévy process has finite variance with

$$\text{var}(X_t) = \mathbb{E}[Z_t] = \frac{t\eta(\tau-\sigma)}{c(\tau-1)} < \infty.$$

For $\tau > 2$, the excess kurtosis is finite and given by

$$\text{kurt}(X_t) = \frac{\kappa_4(X_t)}{\kappa_2(X_t)^2} = \frac{3\text{var}(Z_t)}{(\mathbb{E}[Z_t])^2} = \frac{3(\tau-1)^2}{t\eta(\tau-2)(\tau-\sigma)}.$$

Activity of the jumps and BG index. The Lévy process X is infinite-activity if $\sigma \geq 0$ and finite-activity otherwise. Using Proposition 1 in the Appendix, the regular variation of $\bar{\rho}$ at 0 in Equation (11) implies the regular variation of $\bar{\nu}$ at 0

$$\bar{\nu}(x) \stackrel{x \rightarrow 0}{\sim} \frac{2^{\alpha+1}\Gamma(\alpha+1/2)}{\sqrt{\pi}} \ell(1/x^2) x^{-2\alpha}. \quad (25)$$

The BG index of X is therefore equal to $2\alpha = 2\max(0, \sigma) \in [0, 2)$. When $\sigma = \alpha > 0$, combining (22) with (13), we obtain the small time limit

$$\mathbb{E} \left[\exp \left(i\vartheta \frac{Z_t \sqrt{2c}}{(\eta t)^{1/(2\alpha)}} \right) \right] \stackrel{t \rightarrow 0}{\rightarrow} e^{-|\vartheta|^{2\alpha}}$$

hence $\frac{Z_t \sqrt{2c}}{(\eta t)^{1/(2\alpha)}}$ tends in distribution to a symmetric stable distribution with parameter $2\alpha \in (0, 2)$ when t tends to 0.

Heavy tails and power-law behaviour. Using Proposition 1 in the Appendix, the regular variation of $\bar{\rho}$ at infinity in Equation (15) implies the regular variation of $\bar{\nu}$ at infinity

$$\bar{\nu}(x) \stackrel{x \rightarrow \infty}{\sim} C_1 x^{-2\tau} \quad (26)$$

where

$$C_1 = \frac{2^{\tau+1}\Gamma(\tau+1/2)}{\sqrt{\pi}} \frac{\eta\Gamma(\tau-\sigma+1)}{\tau c^\tau \Gamma(1-\sigma)}.$$

Additionally, we have

$$\Pr(|X_t| > x) \stackrel{x \rightarrow \infty}{\sim} C_1 t x^{-2\tau}. \quad (27)$$

The increments have therefore heavy tails with power-law exponent 2τ .

Simulation of the increments. As

$$X_t \stackrel{d}{=} \sqrt{Z_t} \epsilon_t$$

one can simulate increments exactly by sampling Z_t from (18) and $\epsilon_t \sim \mathcal{N}(0, 1)$.

Self-decomposability. The self-decomposability of X follows from the self-decomposability of the subordinator Z (Sato, 2001, Theorem 1).

Interpretability of the parameters. In summary, each of the four parameters governs a different property of the NGGP process.

- $\eta > 0$ is a time-scaling parameter: if $(X_t)_{t \geq 0}$ is a NGGP process with parameters $(1, \sigma, \tau, c)$, then $(X_{\eta t})_{t \geq 0}$ is a NGGP process with parameters (η, σ, τ, c) ;
- $\sqrt{c} > 0$ is an inverse-scale parameter: if (X_t) is a NGGP process with parameters $(\eta, \sigma, \tau, 1)$, then $(Z_t/\sqrt{c})$ is a NGGP process with parameters (η, σ, τ, c) ;
- $\sigma \in (-\infty, 1)$ tunes the activity of the jumps; the process is finite-activity if $\sigma < 0$ and infinite-activity otherwise, with corresponding BG index $\beta = \max(2\sigma, 0)$;
- 2τ is the power-law exponent, controlling the tails of the distribution, with $\Pr(X_t > z) \stackrel{z \rightarrow \infty}{\sim} t C_1 z^{-2\tau}$ for some constant C_1 .

2.4 Generalisation

One could consider more generally a Lévy intensity of the form

$$\rho(w) = \frac{\eta}{c^\sigma \Gamma(1-\sigma)} w^{-1-\sigma} \left(wc \int_0^1 u^{1-\sigma} h(u) e^{-ucw} du + h(1) e^{-cw} \right) \quad (28)$$

where $h : (0, 1] \rightarrow (0, \infty)$ is a differentiable function which satisfies

$$\int_0^1 h(u) du < \infty \quad \text{and} \quad h(u) \stackrel{u \rightarrow 0}{\sim} u^{\tau-1}. \quad (29)$$

For $\tau > \sigma$, we have $u^{1-\sigma} h(u) \rightarrow 0$ and (28) takes the alternative form

$$\rho(w) = \frac{\eta}{c^\sigma \Gamma(1-\sigma)} w^{-1-\sigma} \int_0^1 (u^{1-\sigma} h(u))' e^{-uwc} du. \quad (30)$$

The proposed subordinator (6) is obtained as a special case when $h(u) = u^{\tau-1}$.

The Lévy process is finite activity for $\sigma < 0$ and infinite-activity for $\sigma \geq 0$. Using Karamata's theorem for regularly varying functions, the tail Lévy intensity $\bar{\rho}$ of the Lévy intensity is regularly varying at 0 with BG index $\alpha = \max(0, \sigma)$ and at infinity with tail index τ . The Lévy intensity (28) takes the form of a sum of a compound Poisson intensity and a generalised gamma intensity. The compound Poisson intensity can be written as

$$\begin{aligned} \rho_1(w) &= \frac{\eta}{c^{\sigma-1} \Gamma(1-\sigma)} \int_0^1 h(u) u (uw)^{-\sigma} e^{-uw} du \\ &= \frac{\eta(1-\sigma) \int_0^1 h(u) du}{c^{\sigma-1}} \int_0^1 \frac{h(u)}{\int_0^1 h(v) dv} u \rho_{GG}(uw; \sigma-1, c) du \end{aligned}$$

which is a mixture of (finite-activity) generalised gamma processes. It follows that if Z is a subordinator with Lévy intensity (28), we therefore have

$$Z_t \stackrel{d}{=} Z_{t,1} + Z_{t,2}$$

where $Z_{t,1} \sim \text{GG}(\frac{t\eta h(1)}{c^\sigma}, \sigma, c)$ is an exponentially tilted stable random variable and

$$Z_{t,2} \stackrel{d}{=} \sum_{j=1}^{\tilde{K}_t} \tilde{G}_{t,j} \tilde{U}_{t,j}$$

where $\tilde{K}_t \sim \text{Poisson}(\eta \int_0^1 h(u) du)$, $\tilde{G}_{t,j} \sim \text{Gamma}(1 - \sigma, c)$ and $1/\tilde{U}_{t,j}$ have probability density function $\frac{h(u)}{\int_0^1 h(v) dv}$. Finally, if $u^{1-\sigma}h(u)$ is monotone increasing, with $\tau > \sigma$, then Equation (30) implies that the Lévy process is a tempered stable process if $\sigma > 0$, and a generalised gamma convolution if $\sigma = 0$; it is therefore self-decomposable.

2.5 Comparison to other models and discussion

Comparison to Ayed et al. The (normalised) GGP process was introduced by Ayed et al. (2019) as a prior for random probability measures with power-law properties, and applied to the modeling of word frequencies. Ayed et al. (2019) introduced the form (7) which is only valid for $\tau > \sigma$. The alternative form (6) we introduce here allows to deal with the case $0 < \tau \leq \sigma$ as well; in particular, one obtains the stable process as a particular case. Ayed et al. (2019) showed that the tail Lévy intensity of the GGP is regularly varying at 0 and infinity and deduced the asymptotic behaviour of large and small jumps. Here we derive a number of additional important properties of the process and of the distribution of the increments. We show that it is decomposable, and crucially, that one can sample exactly the increments at any time scale. Ayed et al. (2019) used the name GBFRY process for the process, due to its form similar to the form of the GBFRY distribution (see Section B in the Appendix); however, as it is customary to give the same name to the process and to the distribution of the increments, which are not GBFRY distributed, we prefer here to use the name generalised gamma-Pareto.

Tempered stable process and generalised gamma convolutions. If $\tau > \sigma > 0$, the subordinator falls in the general class of tempered stable processes, introduced by Rosiński (2007). Noting that $\gamma(\tau - \sigma, cw) = w^{\tau - \sigma + 1} \int_0^c u^{\tau - \sigma} e^{-wu} du$, the model (6) is indeed of the form $w^{-1 - \sigma} q(w)$ where the so-called tempering function q is given by

$$q(w) = \frac{\eta(\tau - \sigma)}{c^\tau \Gamma(1 - \sigma)} \int_0^c u^{\tau - \sigma - 1} e^{-wu} du. \quad (31)$$

By Bernstein's theorem, the function q is completely monotone.

For $\sigma = 0$, the subordinator belongs to the class of generalised gamma convolutions (Thorin, 1977; Bondesson, 1992; James et al., 2008), of the form $w^{-1} \int_0^\infty e^{wu} U(du)$ with Thorin measure $U(du) = \frac{\eta}{c^\tau} u^{\tau - 1} 1_{u \in (0, c)} du$.

Table 1: Comparison between different Lévy processes. VG: Variance Gamma; NIG: Normal inverse Gaussian; NGG: Normal generalised gamma; NS: Normal Stable; TS: Tempered Stable; St: Student t ; GH: Generalised hyperbolic

Model	Heavy tails	Finite 2nd moment	BG index	Tractable f_t for any t	Exact simulation from f_t for any t
VG	No	Yes	$\beta = 0$	Yes	Yes
NIG	No	Yes	$\beta = 1$	Yes	Yes
NGG	No	Yes	$\beta \in (0, 2)$	No	Yes
NS	Yes, $2\tau \in (0, 2)$	No	$\beta = 2\tau \in (0, 2)$	No	Yes
TS	No	Yes	$\beta \in [0, 2)$	No	Yes
St	Yes, $2\tau \in (0, \infty)$	Yes if $2\tau > 2$	$\beta = 1$	No	No
GH	Depends	Depends	$\beta = 1$	No	No
NGGP	Yes, $2\tau \in (0, \infty)$	Yes if $\tau > 1$	$\beta = 2 \max(0, \sigma) \in [0, 2)$	No	Yes

The subordinator, for any σ , also falls into the extended Thorin class described by Grigelionis (2007), see also the discussion in Section 1.8 in (James et al., 2008).

Comparison to other models. As mentioned in the introduction, a number of different Lévy processes have been proposed in the literature. While each process can capture some range of the different tail and jump behaviour, none of them is flexible enough to capture the whole range of tail and jump-activity indices. Variance gamma, normal inverse Gaussian, exponentially tilted stable and tempered stable process do not capture heavy tails; the normal stable process has infinite variance, and the same parameter tunes the activity of the jumps and the BG index; for generalised hyperbolic process, the BG index is fixed to 1.

A drawback of the proposed model is that, contrary to popular models such as the variance gamma or normal inverse Gaussian processes, the increments X_t do not have an analytical probability density function f_t . This is balanced however by the fact that one can sample exactly from the distribution of the increments, and one can therefore resort to likelihood-free methods for posterior inference, as described in the next section. Table 1 summarises the properties of the different models. Note that, as mentioned in (Cont and Tankov, 2004, Section 4.6), the generalised hyperbolic and student t are not closed under convolution, and so there is no analytic expression for f_t at any given time $t > 0$, which may be an issue if data are sampled irregularly.

Some interesting connections can be drawn with other classes of stochastic processes. If $\tau > \sigma$, due to the mixture form (8), the GGP distribution arises as the marginal distribution of a quantile clock process (see Theorem 3.1 by James and Zhang (2011)) with parameters (R, L) where R is a Pareto random variable and L a GG subordinator,

3 Lévy-driven stochastic volatility models

Let S_t denote the price of a financial asset, e.g. a market or a stock index, at time t . Denote $X_t = \log \left(\frac{S_t}{S_0} \right)$. Observations are obtained at fixed discrete times $t_0 = 0 < t_1 < t_2 < \dots < t_n$, and we write, for $k = 1, 2, \dots$

$$Y_k := \log \frac{S_{t_k}}{S_{t_{k-1}}} = X_{t_k} - X_{t_{k-1}} \quad (32)$$

the log-returns (or more shortly, called returns). Let $\Delta_k = t_k - t_{k-1}$ be the inter-arrival times between observations. We assume that

$$X_t = \mu_0 t + \mu_1 V_t^* + B_{V_t^*} \quad (33)$$

where μ_0 is the drift parameter, μ_1 is the risk premium, B_t is a Brownian motion, independent of the stochastic process V_t^* , which can be interpreted as the integrated stochastic volatility. For $k = 1, \dots, n$, let

$$\bar{V}_k = V_{t_k}^* - V_{t_{k-1}}^*$$

be the integrated stochastic volatility over the interval (t_{k-1}, t_k) . The observations $(Y_1, \dots, Y_n)$ are conditionally independent given $(\bar{V}_1, \dots, \bar{V}_n)$, with

$$Y_k \mid \bar{V}_k \sim \mathcal{N}(\mu_0 \Delta_k + \mu_1 \bar{V}_k, \bar{V}_k).$$

We consider two different stochastic processes for the integrated volatility process (V_t^*) : a Lévy process and a Ornstein-Uhlenbeck based model.

3.1 Exponentiated Lévy process

Assume that $(V_t^*)_{t \geq 0}$ is a subordinator with no drift with Lévy intensity ρ parameterised by a vector ϕ . The integrated volatilities $(\bar{V}_1, \dots, \bar{V}_n)$ are therefore conditionally independent, with

$$\bar{V}_k \mid \phi \sim F_{V_{\Delta_k}^*} \quad (34)$$

where $F_{V_t^*}$ denotes the distribution of V_t^* , with Laplace transform

$$\int_0^\infty e^{-\lambda x} dF_{V_t^*}(x) = e^{-t \int_0^\infty (1 - e^{-\lambda w}) \rho(w) dw}.$$

If V_t^* is taken to be the GGP model with intensity (6), then $B_{V_t^*}$ is a NGGP Lévy process.

3.2 Ornstein-Uhlenbeck based stochastic volatility model

We also consider a non-Gaussian Ornstein-Uhlenbeck based model (Barndorff-Nielsen and Shephard, 2001) with

$$V_t^* = \int_0^t V_t dt$$

where the instantaneous stochastic volatility process $(V_t)_{t \geq 0}$ is stationary and satisfies

$$V_t = V_0 e^{-\lambda t} + \int_0^t e^{\lambda(s-t)} dZ_s$$

for some $\lambda > 0$ and some background driving Lévy process Z_t with Lévy measure $\tilde{\rho}$. Additionally, for any $t > 0$ the random variable $V_t \sim F$ is infinite-divisible and self-decomposable with Laplace transform

$$\mathbb{E}[e^{-\vartheta V_t}] = e^{-\int_0^\infty (1 - e^{-\vartheta w}) \rho(w) dw}$$

where $\tilde{\rho}$ and ρ are related by the expression

$$\tilde{\rho}(w) = -\rho(w) - w\rho'(w).$$

To define the model, one can either define the mean measure $\tilde{\rho}$ of the subordinator Z_t , or choose the stationary (self-decomposable) distribution F of V_t , hence ρ . In practice, the second approach is often chosen; examples include the gamma (Roberts et al., 2004; Griffin and Steel, 2006; Frühwirth-Schnatter and Sögner, 2009), generalized inverse Gaussian (Gander and Stephens, 2007), and exponentially tilted stable distributions (Gander and Stephens, 2007; Andrieu et al., 2010) as marginals. The integrated stochastic volatilities over the interval (t_{k-1}, t_k) are obtained, for $k = 1, \dots, n$, by

$$\bar{V}_k = \int_{t_{k-1}}^{t_k} V_t dt = \lambda^{-1} (Z_{\lambda t_k} - V_{t_k} - (Z_{\lambda t_{k-1}} - V_{t_{k-1}})) \quad (35)$$

where $(\frac{V_{t_k}}{Z_{\lambda t_k}})$ follows a linear dynamic model with $Z_0 = 0$, $V_0 \sim F$, and for $k = 1, \dots, n$,

$$\begin{pmatrix} V_{t_k} \\ Z_{\lambda t_k} \end{pmatrix} = \begin{pmatrix} e^{-\lambda \Delta_k} V_{t_{k-1}} \\ Z_{\lambda t_{k-1}} \end{pmatrix} + \varepsilon_k, \quad \text{with } \varepsilon_k \stackrel{d}{=} \begin{pmatrix} e^{-\lambda \Delta_k} \int_0^{\Delta_k} e^{\lambda t} dZ_{\lambda t} \\ \int_0^{\Delta_k} dZ_{\lambda t} \end{pmatrix}. \quad (36)$$

Exact simulation of $(\bar{V}_1, \dots, \bar{V}_n)$ from the model defined by Equations (35-36) requires to be able to simulate from F and simulate the independent random variables $(\varepsilon_1, \dots, \varepsilon_n)$. We describe two models where exact simulation is possible.

3.2.1 Model with gamma marginal distribution

A classical choice (Barndorff-Nielsen and Shephard, 2001; Roberts et al., 2004; Griffin and Steel, 2006; Frühwirth-Schnatter and Sögner, 2009) is to take $F = \text{Gamma}(\eta, c)$ as marginal distribution for V_t . This corresponds to

$$\rho(w) = \eta w^{-1} e^{-cw}, \quad \tilde{\rho}(w) = \eta c e^{-cw}. \quad (37)$$

The background driving Lévy measure (Z_t) is therefore finite-activity, and one can sample exactly the state noise ε_k as follows.

1. Simulate $N_k \sim \text{Poisson}(\eta \lambda \Delta_k)$.
2. For $j = 1, \dots, N_k$, simulate $E_{kj} \sim \text{Exp}(c)$, $\theta_{kj} \sim U(0, \Delta_k)$.
3. Set $\varepsilon_k = \begin{pmatrix} e^{-\lambda \Delta_k} \sum_{j=1}^{N_k} e^{\lambda \theta_{kj}} E_{kj} \\ \sum_{j=1}^{N_k} E_{kj} \end{pmatrix}$.

3.2.2 Model with GGP marginal distribution

Let $\tau = \sigma > 0$ or $\tau > \sigma \geq 0$. As shown in Section 2.2, $\text{GGP}(\eta, \sigma, \tau, c)$ is self-decomposable. If V_t has marginal $F = \text{GGP}(\eta, \sigma, \tau, c)$ distribution, this corresponds to ρ be defined by Equation (6), and the Lévy intensity $\tilde{\rho}$ of the background driving Lévy intensity (Z_t) is given by Equation (20). For $\sigma > 0$, (Z_t) is infinite-activity, and one needs to resort to numerical methods to approximately sample (ε_k) . This could be done by using the representation of the process as a sum of GG process and a finite activity process as described in Section 2.2, and using a truncated series representation for simulating the GG process.

We focus here on the case $\sigma = 0$, where $\tilde{\rho}$ simplifies to

$$\tilde{\rho}(w) = \frac{\eta\tau}{c^\tau} w^{-1-\tau} \gamma(\tau+1, cw) \quad (38)$$

with $\int_0^\infty \tilde{\rho}(w)dw = \eta$, and the background Lévy process (Z_t) is therefore finite-activity. We can therefore simulate ε_k exactly as follows.

1. Simulate $N_k \sim \text{Poisson}(\eta\lambda\Delta_k)$
 2. For $j = 1, \dots, N$, simulate $E_{kj} \sim \text{Exp}(c)$, $U_{kj} \sim \text{Pareto}(\tau, 1)$, $\theta_{kj} \sim U(0, \Delta_k)$
 3. Set $\varepsilon_k = \left(\frac{e^{-\lambda\Delta_k} \sum_{j=1}^{N_k} e^{\lambda\theta_{kj}} E_{kj} U_{kj}}{\sum_{j=1}^{N_k} E_{kj} U_{kj}} \right)$
- which is similar to the model with gamma marginals, with $E_{kj}U_{kj}$ in place of E_{kj} .

3.3 Posterior Inference

Let ϕ denote the set of unknown parameters of both models. That is, ϕ includes the drift and risk premium parameters μ_0 and μ_1 , the parameters of the Lévy intensity and, for the Ornstein-Uhlenbeck based model, the discounting factor $\lambda > 0$. Let $\pi(\phi)$ be some prior density. We aim at approximating the posterior density $\pi(\phi \mid y_1, \dots, y_n)$. The marginal likelihood takes the form

$$p(y_1, \dots, y_n \mid \phi) = \int_{\mathbb{R}_+^n} \left[\prod_{k=1}^n p(y_k \mid \bar{v}_k, \mu_0, \mu_1) \right] d\mathbf{F}_n(\bar{v}_1, \dots, \bar{v}_n) \quad (39)$$

where

$$\mathbf{F}_n(\bar{v}_1, \dots, \bar{v}_n) = \Pr(\bar{V}_1 \leq \bar{v}_1, \dots, \bar{V}_n \leq \bar{v}_n \mid \phi)$$

denotes the joint cumulative distribution function of the integrated variances. In the exponentiated Lévy process, we have

$$\mathbf{F}_n(\bar{v}_1, \dots, \bar{v}_n) = \prod_{k=1}^n F_{V_{\Delta_k}}(\bar{v}_k).$$

If F_{V_t} does not admit a tractable probability density function, as it is the case for the proposed GGP model, neither $\mathbf{F}_n$ nor $p(y_1, \dots, y_n \mid \phi)$ are tractable, preventing the implementation of a Metropolis-Hastings Markov chain Monte Carlo algorithm. The same applies for the Ornstein-Uhlenbeck model.

We therefore propose to use a pseudo-marginal Markov chain Monte Carlo (MCMC) algorithm (Beaumont, 2003; Andrieu and Roberts, 2009), which only requires to simulate from $\mathbf{F}_n$. The pseudo-marginal algorithm replaces the untractable marginal likelihood (39) by an unbiased estimator, yet admitting the posterior distribution of interest as invariant distribution. Let q denote some proposal distribution for the parameters. At iteration i of the algorithm, we have

1. Sample $\phi^* \mid \phi^{(i-1)} \sim q(\cdot \mid \phi^{(i-1)})$
2. Compute an unbiased estimate $\hat{p}(y_1, \dots, y_n \mid \phi^*)$
3. With probability

$$\min \left(1, \frac{\hat{p}(y_1, \dots, y_n \mid \phi^*) \pi(\phi^*) q(\phi^{(i-1)} \mid \phi^*)}{\hat{p}(y_1, \dots, y_n \mid \phi^{(i-1)}) \pi(\phi^{(i-1)}) q(\phi^* \mid \phi^{(i-1)})} \right)$$

set $\phi^{(i)} = \phi^*$ and $\hat{p}(y_1, \dots, y_n \mid \phi^{(i)}) = \hat{p}(y_1, \dots, y_n \mid \phi^*)$.

Otherwise, set $\phi^{(i)} = \phi^{(i-1)}$ and $\hat{p}(y_1, \dots, y_n \mid \phi^{(i)}) = \hat{p}(y_1, \dots, y_n \mid \phi^{(i-1)})$.

In the exponential Lévy model, an unbiased estimator can be obtained via Monte Carlo approximation

$$\hat{p}(y_1, \dots, y_n \mid \phi) = \prod_{k=1}^n \frac{1}{n_p} \sum_{j=1}^{n_p} p(y_k \mid \bar{v}_k^{(j)}, \mu_0, \mu_1)$$

where $\bar{v}_k^{(j)} \sim F_{V_{\Delta_k}}$ for $k = 1, \dots, n$ and $j = 1, \dots, n_p$, with n_p the number of Monte Carlo samples (called particles thereafter).

In the Ornstein-Uhlenbeck model, the marginal likelihood can be approximated with a (bootstrap) sequential Monte Carlo algorithm, a standard inference technique for this class of models (Andrieu et al., 2010; Jasra et al., 2011; Chopin et al., 2013). The resulting algorithm is known in this case as a particle marginal Metropolis-Hastings algorithm (Andrieu et al., 2010).

4 Experiments

Priors. In all the experiments, the drift μ_0 and the premium μ_1 parameters are set to zero. For the GGP model, we assume that we are in the infinite-activity regime, with $\sigma \geq 0$, and with finite variance, hence $\tau > 1$. The priors are set as follows:

$$\eta \sim \text{Gamma}(0.1, 0.1), \quad c \sim \text{Gamma}(0.1, 0.1), \quad (\tau - 1) \sim \text{Gamma}(1, 1), \quad \sigma \sim \text{Unif}(0, 1).$$

The more informative prior for τ reflects the empirical evidence that, for many financial dataset, the power-law exponent (2τ for the GGP) is in the range $(2, 5)$ (Cont and Tankov, 2004, Section 7.1). For the Ornstein-Uhlenbeck model, we additionally set $\lambda \sim \text{Gamma}(0.1, 0.1)$.

Software. To fit both models, we use the Particles Library¹, which allows to perform posterior inference in state-space models using particle MCMC algorithms. The code and datasets can be

¹<https://github.com/nchopin/particles>

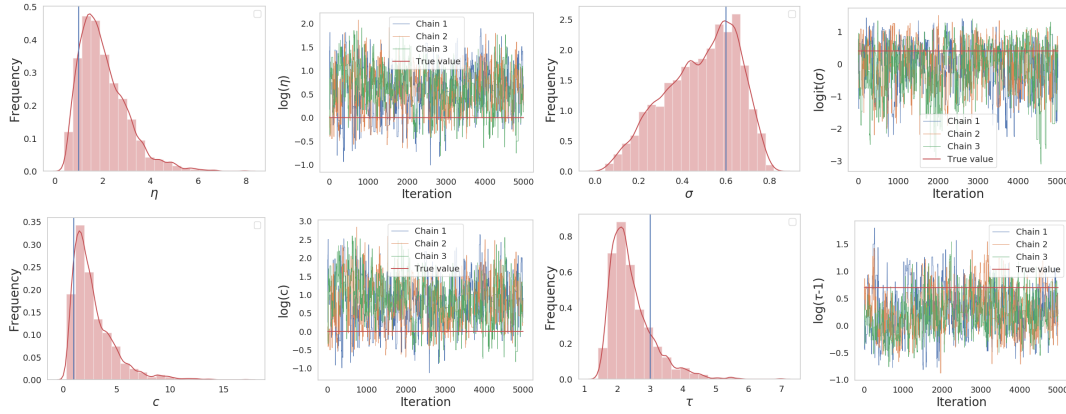


Figure 1: Histograms and trace plots of the posterior distributions of the parameters on the simulated data experiment. The trace plots correspond to the values of the transformed parameters. The blue line represents the value of the parameter used to simulate the data. (top-left) posterior of η , (top-right) posterior of σ , (bottom-left) posterior of c , (bottom-right) posterior of τ .

found on [github](https://github.com/FadhelA/ggp_dpl)².

4.1 Exponentiated Lévy model

4.1.1 Simulated datasets

We generate a synthetic dataset of $n = 5000$ unit-spaced observations from the NGGP model, with parameters $\eta = 1$, $\sigma = 0.6$, $\tau = 3$ and $c = 1$. We run three independent MCMC chains with $n_{mcmc} = 10\,000$ iterations each, of which 5 000 iterations are used for burn-in. The number of particles to compute the marginal likelihood estimates is set to $n_p = 4\,000$. In Figure 1 we report histograms and trace plots of the posterior samples for each of the four parameters. Trace plots suggest the convergence of the MCMC algorithm.

4.1.2 Real-world datasets.

Dataset. We consider a dataset composed of the time-series of the stock prices of six large technology companies: Apple, Amazon, Facebook, Google, Microsoft and Netflix. The data are sampled every minute from the 10th of July 2019 until the 22nd of January 2020, with approximately 50 000 time points. We subsample 1 500 observations as training data to estimate the parameters of each model, and use the rest of the observations as test data.

²https://github.com/FadhelA/ggp_dpl

Other models. We compare the fit of the NGGP model to classical Lévy processes on the first dataset. The models compared are the normal-GG (also known as normal-tempered stable or normal-exponentially tilted stable model), with parameters η , σ and c ; two special cases of this model, the variance-gamma ($\sigma = 0$) and normal-inverse-Gaussian ($\sigma = 0.5$); the generalised hyperbolic (GH) model, with four parameters, and the student model, a special case of the GH model with two parameters. We use vague Gamma(0.1, 0.1) priors on all parameters, except for the parameter σ in the GG model, where a uniform prior on $[0, 1]$ is used as for the GGP, and for the degrees of freedoms ν_0 in the student t model, which tunes the power-law tail, where $(\nu_0/2 - 1) \sim \text{Gamma}(1, 1)$ to reflect the prior assumptions on the tails (as for τ in the NGGP). Note that we can compare here to the GH and student models as the observations are equally spaced.

Results. We run 3 MCMC chains in parallel, with 5 000 iterations (2 500 burn-in) and 1 500 particles. The estimated parameters and 95% credible intervals for the parameters of the NGGP are reported in Table 2. The posterior mean for τ is around 2; this corresponds to a power-law exponent for X_t of around 4 which is in concordance with empirical observations (Cont and Tankov, 2004, Section 7.1). One exception is the Amazon stock, where τ is closer to 1, indicating a heavier tail. We first compare the models using the Kolmogorov-Smirnov (KS) statistics between the posterior predictive distribution and the empirical distribution of the test data. Results are reported in Table 3. The KS statistics is rather insensitive to the tail of the distribution, and the performances are similar for most models considered. To investigate the goodness-of-fit to the tails of the distribution, we compare the ranked empirical squared log-return to their posterior predictive distribution. Both GHD, VG and NIG, which have exponentially decaying tails, provide similar results, and we only report the results of the GH. Results are reported in Figure 2. We can see that the NGGP model successfully captures the behaviour of tails for the different datasets, while the GH fails to provide accurate posterior predictive for some datasets such as Facebook. The NS model, which has the same parameter to capture the jump-activity and the tail behaviour, underestimates the value of the tail exponent, and gives a poor fit. The student t model tends to provide poor credible intervals, possibly due to the lack of flexibility of this two-parameter model.

4.2 Ornstein-Uhlenbeck based model

We now consider the Ornstein-Uhlenbeck based stochastic volatility model with NGGP marginal with $\sigma = 0$. As discussed in section 3.2.2, in this case, the simulation of the state noise ε can be done exactly and exact posterior inference is possible. We compare the model with NGGP to the model with normal-gamma marginal (NG), described in Section 3, and demonstrate that NGGP better captures the heavy-tails with minimal computational overhead compared to NG.

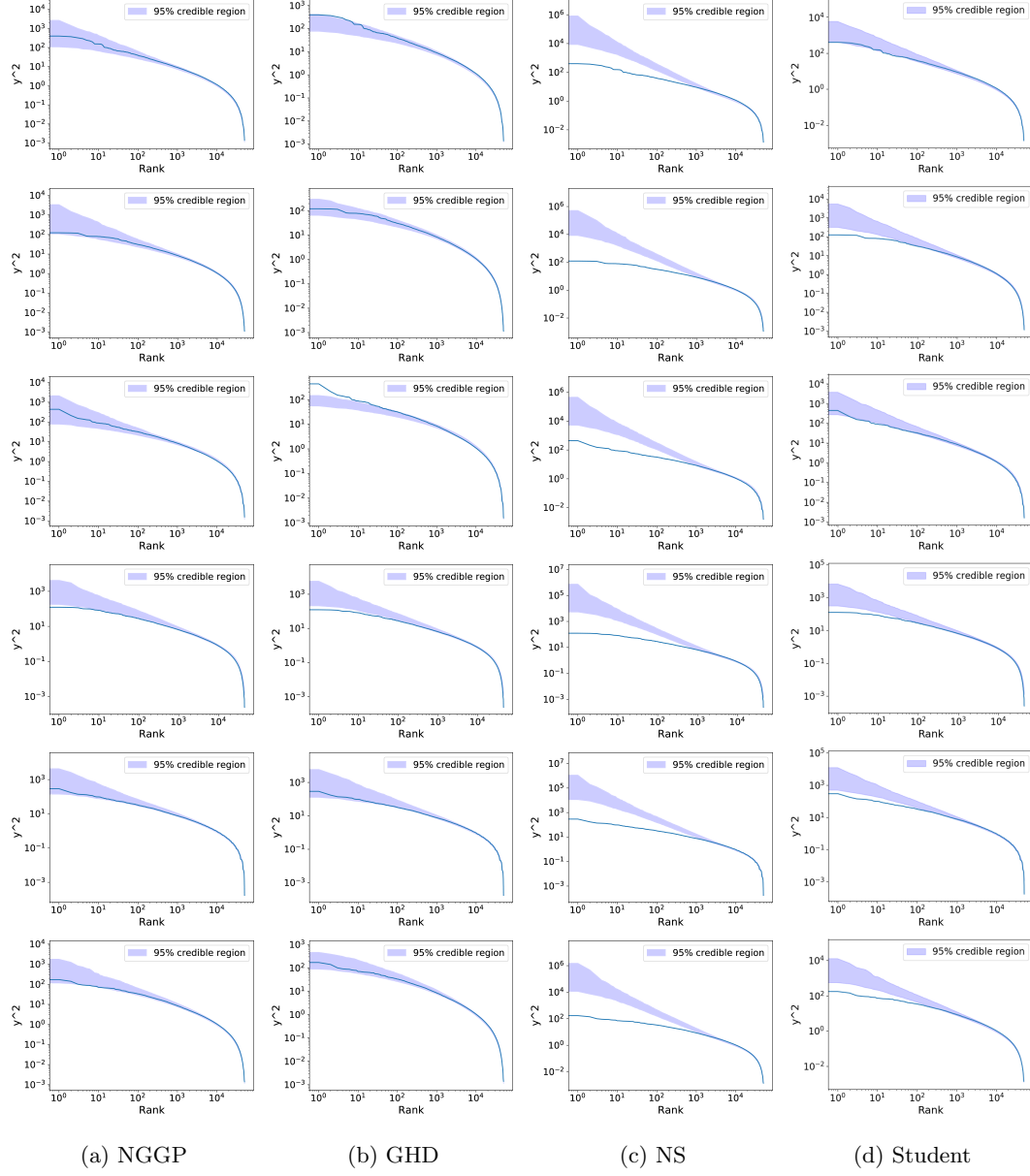


Figure 2: Ranked squared increments on the tech companies dataset. From top to bottom row: Apple, Amazon, Facebook, Google, Microsoft, Netflix. The line represents the ranked y^2 in the test dataset; the shaded area represent the 95% credible region. Results are given for the NGGP, GHD, NS and Student models in this order.

Table 2: Posterior mean and 95% credible interval for the four parameter of the NGGP model on the first dataset.

Data	η	σ	τ	c
Apple	0.51, (0.13, 1.21)	0.44, (0.21, 0.61)	1.82, (1.11, 3.51)	1.14, (0.13, 4.26)
Amazon	0.64, (0.19, 1.40)	0.39, (0.14, 0.58)	1.18, (1.15, 3.59)	1.46, (0.22, 4.61)
Facebook	0.85, (0.34, 1.46)	0.25, (0.04, 0.51)	1.97, (1.25, 3.90)	1.86, (0.38, 4.63)
Google	0.18, (0.02, 0.71)	0.64, (0.50, 0.73)	1.93, (1.08, 4.16)	0.34, (0.01, 1.98)
Microsoft	0.27, (0.05, 0.81)	0.55, (0.39, 0.66)	1.98, (1.11, 4.35)	0.56, (0.05, 2.51)
Netflix	0.21, (0.07, 0.45)	0.54, (0.44, 0.63)	2.55, (1.25, 5.14)	0.29, (0.06, 0.90)

Table 3: Kolmogorov-Smirnov distance between the empirical distribution of the test and the posterior predictive for different models on the first tech companies dataset (the smaller the better).

Data	NGGP	NGG	GHD	NIG	NS	VG	Student
Apple	0.0194	0.0196	0.0194	0.0194	0.0196	0.0218	0.0196
Amazon	0.0087	0.0087	0.0085	0.0085	0.0159	0.0145	0.0092
Facebook	0.0181	0.0182	0.0182	0.0183	0.0245	0.1413	0.0181
Google	0.0205	0.0209	0.0197	0.0193	0.0237	0.0848	0.0200
Microsoft	0.0285	0.0285	0.0287	0.0286	0.0286	0.1567	0.0289
Netflix	0.0079	0.0080	0.0080	0.0084	0.0098	0.0162	0.0079
Mean	0.0172	0.0173	0.0171	0.0171	0.0204	0.0726	0.0173

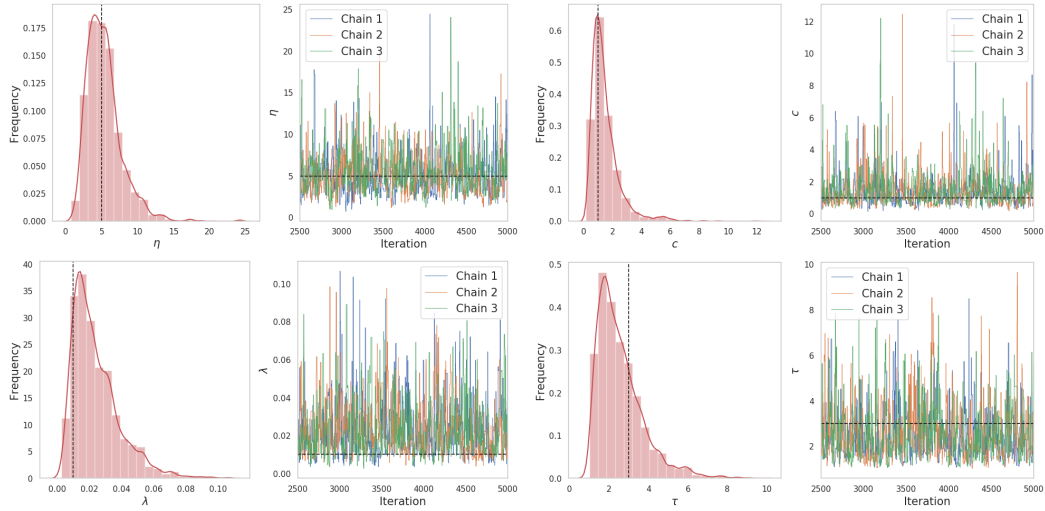


Figure 3: Posterior samples of the parameters on simulated data with $\tau = 3.0$.

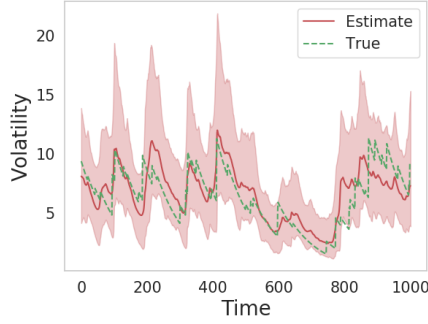


Figure 4: Posterior mean (solid red line) and 95% credible intervals (shaded area) of the integrated volatility. True volatility is in dashed green line.

4.2.1 Simulated dataset

We first demonstrate that our posterior sampler based on particle MCMC could successfully recover the true parameters on a simulated dataset. We simulate data from the Lévy-driven stochastic volatility model with NGGP marginal for $n = 1000$ time-steps and parameters $\eta = 5, c = 1, \tau = 3$ and $\lambda = 0.01$. We run three independent particle MCMC chains with 5000 iterations (2500 burn-in) and 1000 particles. Figure 3 shows that our sampler successfully recovers the parameters, especially the heavy-tail parameter τ . Trace plots suggest the convergence of the sampler. Figure 4 shows the posterior estimate and credible interval for the integrated volatility, together with the true value.

4.2.2 Real-world datasets

Dataset. The dataset is obtained from the Realized library ³. We collected 14 daily stock data from 05-11-2007 to 07-10-2011 (around the time of subprime mortgage crisis), and fitted the Lévy driven stochastic volatility models on daily log return values. The data is accompanied with the estimates of the integrated variances $\bar{v}_k$ for each day, obtained with an estimator based on higher-frequency data; we use these values as ground-truth of the integrated variance, and note it $\bar{v}_k^{\text{true}}$.

Evaluation metrics. For $k = 1, \dots, n$, let $\bar{V}_k^{(1)}, \dots, \bar{V}_k^{(n_s)}$ denote the posterior samples of the integrated variance over the k 's interval, where n_s is the number of MCMC iterations after burn-in. For $k = 1, \dots, n$ and any $\bar{v}_k \geq 0$, let $\hat{G}_k(\bar{v}_k) = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{1}_{\bar{V}_k^{(i)} \leq \bar{v}_k}$ be the Monte Carlo approximation of the posterior cumulative distribution function of the integrated variance $G_k(\bar{v}_k) = \Pr(\bar{V}_k \leq \bar{v}_k \mid y_1, \dots, y_n)$. For $k = 1, \dots, n$, denote

$$\zeta_k = 1 - \hat{G}_k(\bar{v}_k^{\text{true}}) = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{1}_{\bar{V}_k^{(i)} \geq \bar{v}_k^{\text{true}}}.$$

³<https://realized.oxford-man.ox.ac.uk>

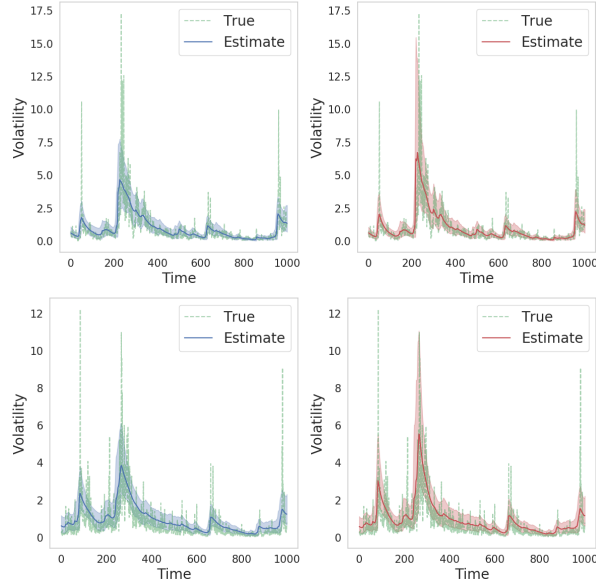


Figure 5: Posterior estimates (solid line) and 95% credible intervals of the integrated volatility under the NG (left) and NGGP (right) models for the AEX index (top row) and AORD index (bottom row). The true integrated volatility is represented by a green dashed line.

In order to assess the goodness-of-fit of the model, we calculate the KS statistics between the empirical distribution of $(\zeta_1, \dots, \zeta_n)$ and the distribution of a uniform random variable on $[0, 1]$. We also compare the fit of the model for different loss functions. Let L be a loss function. If $L(x, y) = (x - y)^2$ is the ℓ_2 loss, the Bayes estimator is the posterior mean. In case of $\ell_{1,\alpha}$ loss defined as

$$L(x, y) = \begin{cases} x - y & \text{if } x \geq y \\ \frac{1-\alpha}{\alpha} |x - y| & \text{if } x < y \end{cases},$$

the Bayes estimator is the α -quantile $G_k^{-1}(\alpha)$ of G_k . We assess the fit of each model by computing the average loss

$$\frac{1}{n} \sum_{k=1}^n L(\bar{v}_k^{\text{true}}, \hat{v}_k^L)$$

where $\hat{v}_k^L$ is the Bayes estimator under the loss L .

Results. For every stock data, we run three independent chains of particle MCMC with 5 000 iterations (2 500 burn-in) and 1 000 particles. The estimated parameters and credible intervals are given in Table 4. The comparisons between the fits of the two models in terms of KS statistics and empirical loss, for different loss functions, are reported in Table 5. The model with NGGP marginal outperformed the one with NG marginal for all used metrics. Especially, since NGGP better captures the heavy-tails, the performance gap becomes more significant for the metrics

Table 4: Posterior mean and 95% credible intervals of the parameters of the NG and NGGP marginals for the different indices.

	NG			NGGP			
	η	λ	c	η	λ	c	τ
AEX	1.43 (0.71, 2.50)	0.02 (0.01, 0.04)	1.41 (0.66, 2.56)	2.58 (1.17, 4.47)	0.03 (0.01, 0.05)	9.34 (3.00, 23.04)	1.49 (1.05, 2.37)
AORD	1.97 (0.92, 3.42)	0.02 (0.01, 0.03)	2.06 (0.88, 3.60)	3.62 (1.52, 6.87)	0.03 (0.01, 0.04)	11.26 (2.89, 29.19)	1.65 (1.09, 3.05)
DJI	1.28 (0.68, 2.05)	0.02 (0.01, 0.03)	1.26 (0.65, 2.10)	2.19 (1.10, 3.69)	0.03 (0.01, 0.04)	8.53 (2.30, 19.95)	1.38 (1.02, 2.25)
FTSE	1.34 (0.68, 2.22)	0.02 (0.01, 0.02)	1.46 (0.73, 2.37)	3.19 (1.35, 6.67)	0.02 (0.01, 0.04)	14.36 (3.55, 38.67)	1.36 (1.04, 2.08)
GSPTSE	1.23 (0.59, 2.13)	0.01 (0.01, 0.02)	1.28 (0.57, 2.31)	2.31 (0.92, 4.26)	0.02 (0.01, 0.03)	10.33 (2.42, 25.48)	1.42 (1.02, 2.29)
HSI	1.34 (0.58, 2.38)	0.01 (0.01, 0.02)	1.61 (0.73, 2.91)	2.51 (0.96, 4.60)	0.01 (0.01, 0.03)	10.00 (2.70, 23.93)	2.08 (1.07, 3.17)
IBEX	1.67 (0.88, 2.68)	0.03 (0.02, 0.06)	1.72 (0.86, 2.90)	2.92 (1.23, 5.91)	0.04 (0.02, 0.08)	9.17 (1.87, 26.00)	1.68 (1.04, 3.53)
IXIC	1.43 (0.77, 2.37)	0.02 (0.01, 0.03)	1.42 (0.72, 2.43)	2.06 (1.02, 3.84)	0.02 (0.01, 0.03)	6.15 (1.74, 15.39)	1.65 (1.04, 3.07)
KS11	1.55 (0.91, 2.42)	0.03 (0.02, 0.06)	1.69 (1.00, 2.57)	2.78 (1.20, 4.86)	0.06 (0.03, 0.09)	9.48 (2.12, 22.46)	1.69 (1.08, 3.54)
MXX	1.09 (0.43, 1.85)	0.01 (0.01, 0.02)	0.99 (0.38, 1.86)	2.04 (0.84, 3.99)	0.02 (0.01, 0.04)	7.31 (1.40, 19.96)	1.48 (1.04, 2.81)
N225	1.05 (0.86, 1.53)	0.03 (0.02, 0.04)	1.23 (0.78, 1.53)	1.47 (0.71, 2.37)	0.05 (0.03, 0.07)	3.67 (0.98, 8.77)	1.95 (1.27, 3.30)
RUT	1.41 (0.70, 2.35)	0.02 (0.01, 0.03)	1.33 (0.56, 2.31)	2.70 (1.33, 4.99)	0.03 (0.02, 0.05)	10.43 (3.19, 25.73)	1.39 (1.03, 2.31)
SPX	1.22 (0.67, 2.00)	0.02 (0.01, 0.03)	1.19 (0.59, 2.03)	2.07 (0.93, 3.52)	0.02 (0.01, 0.04)	8.91 (2.84, 19.89)	1.37 (1.02, 2.09)
SSMI	1.75 (0.95, 2.79)	0.03 (0.01, 0.05)	1.86 (0.96, 3.06)	2.86 (1.32, 5.25)	0.04 (0.02, 0.07)	9.18 (2.60, 24.71)	1.60 (1.06, 2.89)

emphasizing the heavy-tail regime ($\ell_{1,0.95}, \ell_{1,0.99}$). This is well highlighted in Figure 5, which gives the estimated volatility and credible intervals under both models for the AEX and AORD stock indices. NGGP in general better captures “spikes” in the log-return values while NG often fails to get credible interval with good coverage.

Acknowledgments. The authors thank Cian Naik, Lancelot James and Matthias Winkel for useful feedback on an earlier version of this article.

Table 5: Comparison of the fit of the NG and NGGP models under different metrics.

Data	KS($(\zeta_k), U(0, 1)$)		ℓ_2		$\ell_{1,0.5}$		$\ell_{1,0.95}$		$\ell_{1,0.99}$	
	NG	NGGP	NG	NGGP	NG	NGGP	NG	NGGP	NG	NGGP
AEX	0.237	0.200	0.920	0.950	0.398	0.396	0.127	0.113	0.074	0.053
AORD	0.531	0.511	0.688	0.680	0.465	0.453	0.117	0.101	0.065	0.047
DJI	0.371	0.341	1.859	1.677	0.476	0.456	0.162	0.138	0.107	0.077
FTSE	0.269	0.241	2.590	2.510	0.479	0.480	0.186	0.159	0.134	0.092
GSPTSE	0.450	0.432	13.993	13.656	0.615	0.612	0.277	0.259	0.226	0.198
HSI	0.351	0.335	1.081	1.056	0.426	0.411	0.133	0.126	0.086	0.072
IBEX	0.264	0.245	0.824	0.788	0.422	0.413	0.122	0.106	0.067	0.045
IXIC	0.433	0.421	0.849	0.898	0.412	0.418	0.102	0.092	0.055	0.042
KS11	0.237	0.177	1.740	1.207	0.405	0.358	0.151	0.095	0.096	0.039
MXX	0.580	0.553	1.030	1.142	0.523	0.518	0.095	0.087	0.047	0.036
N225	0.283	0.230	0.674	0.807	0.360	0.362	0.087	0.073	0.045	0.030
RUT	0.570	0.392	1.217	1.192	0.449	0.454	0.079	0.072	0.031	0.025
SPX	0.388	0.337	1.318	1.317	0.440	0.435	0.131	0.115	0.082	0.058
SSMI	0.276	0.259	1.420	1.292	0.438	0.437	0.160	0.135	0.098	0.062
Mean	0.374	0.344	2.157	2.084	0.451	0.443	0.138	0.119	0.087	0.063

A Useful identities

Recall that $\gamma(s, x)$ denotes the incomplete gamma function. We have, for $s, x > 0$

$$\gamma(s, x) = x^s \int_0^1 v^{s-1} e^{-vx} dv \quad (40)$$

$$s\gamma(s, x) = \gamma(s+1, x) + x^s e^{-x} \quad (41)$$

$$\gamma(s, x) \stackrel{x \rightarrow 0}{\sim} \frac{x^s}{s} \quad (42)$$

For any $\tau, \kappa, c > 0$, $\int_0^\infty x^{-1-\tau} \gamma(\kappa + \tau, cx) dx = \frac{c^\tau \Gamma(\kappa)}{\tau}$.

B Generalised BFRY distribution

A positive random variable X with generalised BFRY distribution has probability density function

$$\text{GBFRY}(x; \kappa, \tau, c) = \frac{\tau}{c^\tau \Gamma(\kappa)} x^{-1-\tau} \gamma(\kappa + \tau, cx) \quad (43)$$

for some parameters $\kappa, \tau, c > 0$. c is an inverse scale parameter; κ controls the behavior at 0 as, using Karamata's theorem (Bingham et al., 1989, Proposition 1.5.10) and Equation (42)

$$\Pr(X < x) \stackrel{x \rightarrow 0}{\sim} \frac{\tau c^\kappa}{\kappa(\kappa + \tau) \Gamma(\kappa)} x^\kappa$$

and τ is a power-law exponent controlling the tails of the distribution as, using Karamata's theorem

$$\Pr(X > x) \stackrel{x \rightarrow \infty}{\sim} \frac{\Gamma(\kappa + \tau)}{c^\tau \Gamma(\tau)} x^{-\tau}.$$

The moments are given by

$$\mathbb{E}[X^m] = \frac{\tau \Gamma(m + \kappa)}{c^m (\tau - m) \Gamma(\kappa)} \quad (44)$$

for $m < \tau$, and $\mathbb{E}[X^m] = \infty$ otherwise. A GBFRY random variable admits the following representation

$$X \stackrel{d}{=} Y/Z$$

where $Y \sim \text{Gamma}(\kappa, c)$ random variable and $Z \sim \text{Beta}(\tau, 1)$ where $\text{Gamma}(\kappa, c)$ denotes the gamma distribution with shape parameter $\kappa > 0$ and inverse scale parameter $c > 0$, and $\text{Beta}(a, b)$ denotes the beta distribution with parameters $a, b > 0$. Note that $Z^{-1} \sim \text{Pareto}(\tau, 1)$.

The generalised BFRY was introduced by Ayed et al. (2019) as a generalisation to the BFRY distribution (Pitman and Yor, 1997; Winkel, 2005; Bertoin et al., 2006) whose density is $\kappa/\Gamma(\kappa)x^{-1-\kappa}(1 - e^{-x})$; it is obtained as a special case when $c = 1$ and $\kappa = 1 - \tau \in (0, 1)$. The term BFRY was coined after Bertoin, Fujita, Roynette and Yor by Devroye and James (2014).

C Tauberian-Abelian theorem for subordinated Brownian

Proposition 1 *Let ϱ be a Lévy measure on $(0, \infty)$ and let $\bar{\rho}(x) = \int_x^\infty \varrho(dw)$ be its tail Lévy intensity. For $x > 0$, let*

$$\bar{\nu}(x) = \int_{|s|>x} \nu(s) ds \quad (45)$$

where, for $s \in \mathbb{R}$,

$$\nu(s) = \int_0^\infty (2\pi w)^{-1/2} e^{-\frac{s^2}{2w}} \varrho(dw).$$

If $\bar{\rho}(x)$ is a regularly-varying function at 0 with

$$\bar{\rho}(x) \stackrel{x \rightarrow 0}{\sim} x^{-\alpha} \ell_1(1/x) \quad (46)$$

where $\alpha \in [0, 1]$ and ℓ_1 is a slowly varying function, then $\bar{\nu}(x)$ is also regularly varying at 0, with

$$\bar{\nu}(x) \stackrel{x \rightarrow 0}{\sim} \frac{2^{\alpha+1} \Gamma(\alpha + 1/2)}{\sqrt{\pi}} x^{-2\alpha} \ell_1(1/x^2). \quad (47)$$

If $\bar{\rho}(x)$ is a regularly-varying function at infinity with

$$\bar{\rho}(x) \stackrel{x \rightarrow \infty}{\sim} x^{-\tau} \ell_2(x) \quad (48)$$

where $\tau \geq 0$ and ℓ_2 is a slowly varying function, then $\bar{\nu}(x)$ is also regularly varying at infinity, with

$$\bar{\nu}(x) \stackrel{x \rightarrow \infty}{\sim} \frac{2^{\tau+1} \Gamma(\tau + 1/2)}{\sqrt{\pi}} x^{-2\tau} \ell_2(x^2) \quad (49)$$

If additionally $x \rightarrow x^{1/2} \bar{\rho}(2x)$ is ultimately monotone, then (47) also implies (46); if $x \rightarrow x^{-1/2} \bar{\rho}(1/(2x))$ is ultimately monotone, then (49) also implies (48).

Proof. We have

$$\bar{\nu}(x) = \sqrt{\frac{2}{\pi}} \int_{\mu_0+x}^{\infty} \int_0^{\infty} w^{-1/2} e^{-\frac{(s-\mu_0)^2}{2w}} \varrho(dw) ds.$$

With the change of variable $u = \frac{(s-\mu_0)^2}{2wx^2}$, we obtain

$$\bar{\nu}(x) = \frac{2x}{\sqrt{\pi}} \int_0^{\infty} \int_{1/(2u)}^{\infty} u^{-1/2} e^{-ux^2} \varrho(dw) du = \frac{2x}{\sqrt{\pi}} \int_0^{\infty} u^{-1/2} e^{-ux^2} \bar{\rho}(1/(2u)) du$$

We have

$$\begin{aligned} u^{-1/2} \bar{\rho}(1/(2u)) &\stackrel{u \rightarrow 0}{\sim} 2^{\tau} u^{\tau-1/2} \ell_2(1/u) \\ u^{-1/2} \bar{\rho}(1/(2u)) &\stackrel{u \rightarrow \infty}{\sim} 2^{\alpha} u^{\alpha-1/2} \ell_1(u). \end{aligned}$$

It follows from Karamata Abelian theorem (Feller, 1971, Chapter XIII, Section 5) that

$$\begin{aligned} \bar{\nu}(x) &\stackrel{x \rightarrow \infty}{\sim} \frac{2^{\tau+1} \Gamma(\tau+1/2)}{\sqrt{\pi}} x^{-2\tau} \ell_2(x^2) \\ \bar{\nu}(x) &\stackrel{x \rightarrow 0}{\sim} \frac{2^{\alpha+1} \Gamma(\alpha+1/2)}{\sqrt{\pi}} x^{-2\alpha} \ell_1(1/x^2). \end{aligned}$$

The reverse Tauberian result holds under monotonicity conditions near 0 or infinity of the function $u \rightarrow u^{-1/2} \bar{\rho}(1/(2u))$, see (Bingham et al., 1989, Theorem 1.7.2 page 39). ■

References

- Aalen, O. (1992). Modelling heterogeneity in survival analysis by the compound poisson distribution. *The Annals of Applied Probability*, 951–972.
- Aït-Sahalia, Y. and J. Jacod (2009). Estimating the degree of activity of jumps in high frequency data. *The Annals of Statistics* 37(5A), 2202–2244.
- Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(3), 269–342.
- Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* 37(2), 697–725.
- Ayed, F., J. Lee, and F. Caron (2019). Beyond the Chinese restaurant and Pitman-Yor processes: Statistical models with double power-law behavior. In *International Conference on Machine Learning*, pp. 395–404.
- Barndorff-Nielsen, O. and N. Shephard (2002). Normal modified stable processes. *Theory Probab. Math. Statist* 65, 1–20.
- Barndorff-Nielsen, O. E. (1997). Processes of normal inverse Gaussian type. *Finance and stochastics* 2(1), 41–68.

- Barndorff-Nielsen, O. E. and N. Shephard (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(2), 167–241.
- Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* 164(3), 1139–1160.
- Belomestny, D. (2010). Spectral estimation of the fractional order of a Lévy process. *The Annals of Statistics* 38(1), 317–351.
- Belomestny, D. and V. Panov (2013). Estimation of the activity of jumps in time-changed Lévy models. *Electronic Journal of Statistics* 7, 2970–3003.
- Bertoin, J., T. Fujita, B. Roynette, and M. Yor (2006). On a particular class of self-decomposable random variables : the durations of Bessel excursions straddling independent exponential times. *Probability and Mathematical Statistics* 26(2), 315–366.
- Bingham, N. H., C. M. Goldie, and J. L. Teugels (1989). *Regular variation*, Volume 27. Cambridge university press.
- Blattberg, R. C. and N. J. Gonedes (1974). A comparison of the stable and student distributions as statistical models for stock prices. *The journal of business* 47(2), 244–280.
- Bondesson, L. (1992). *Generalized gamma convolutions and related classes of distributions and densities*, Volume 76. Lecture notes in Statistics. Springer Verlag.
- Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability* 31(4), 929–953.
- Carr, P., H. Geman, D. B. Madan, and M. Yor (2002). The fine structure of asset returns: An empirical investigation. *The Journal of Business* 75(2), 305–332.
- Champagnat, N., M. Deaconu, A. Lejay, N. Navet, and S. Boukherouaa (2013). An empirical analysis of heavy-tails behavior of financial data: The case for power laws.
- Chopin, N., P. E. Jacob, and O. Papaspiliopoulos (2013). SMC²: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(3), 397–426.
- Cont, R., M. Potters, and J. P. Bouchaud (1997). Scaling in stock market data: stable laws and beyond. In *Scale invariance and beyond*, pp. 75–85. Springer.
- Cont, R. and P. Tankov (2004). *Financial modelling with jump processes*, Volume 2. CRC press.

- Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 19(4), 18.
- Devroye, L. and L. F. James (2014). On simulation and properties of the stable law. *Statistical Methods and Applications* 23(3), 307–343.
- Eberlein, E., U. Keller, and K. Prause (1998). New insights into smile, mispricing, and value at risk: The hyperbolic model. *The Journal of Business* 71(3), 371–405.
- Feller, W. (1971). *An Introduction to Probability theory and its application Vol II*. John Wiley and Sons.
- Frühwirth-Schnatter, S. and L. Sögner (2009). Bayesian estimation of stochastic volatility models based on OU processes with marginal gamma law. *Annals of the Institute of Statistical Mathematics* 61(1), 159–179.
- Gander, M. and D. A. Stephens (2007). Stochastic volatility modelling in continuous time with general marginal distributions: Inference, prediction and model selection. *Journal of Statistical Planning and Inference* 137(10), 3068–3081.
- Gnedin, A., B. Hansen, and J. Pitman (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability surveys* 4, 146–171.
- Griffin, J. E. and M. F. J. Steel (2006). Inference with non-Gaussian Ornstein–Uhlenbeck processes for stochastic volatility. *Journal of Econometrics* 134(2), 605–644.
- Grigelionis, B. (2007). Extended thorin classes and stochastic integrals. *Lithuanian Mathematical Journal* 47(4), 406–411.
- Hofert, M. (2011). Sampling exponentially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 22(1), 3.
- Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 73(2), 387–396.
- Huang, J.-Z. and L. Wu (2004). Specification analysis of option pricing models based on time-changed Lévy processes. *The Journal of Finance* 59(3), 1405–1439.
- James, L. F., B. Roynette, and M. Yor (2008). Generalized gamma convolutions, Dirichlet means, Thorin measures, with explicit examples. *Probability Surveys* 5, 346–415.
- James, L. F. and Z. Zhang (2011). Quantile clocks. *Ann. Appl. Probab.* 21(5), 1627–1662.

- Jasra, A., D. A. Stephens, A. Doucet, and T. Tsagaris (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics* 38(1), 1–22.
- Jing, B.-Y., X.-B. Kong, and Z. Liu (2012). Modeling high-frequency financial data by pure jump processes. *The Annals of Statistics* 40(2), 759–784.
- Landis, M. J. and J. G. Schraiber (2017). Pulsed evolution shaped modern vertebrate body sizes. *Proceedings of the National Academy of Sciences* 114(50), 13224–13229.
- Landis, M. J., J. G. Schraiber, and M. Liang (2012). Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. *Systematic biology* 62(2), 193–204.
- Liang, C. and S. Li (2015). Option pricing and hedging in incomplete market driven by normal tempered stable process with stochastic volatility. *Journal of mathematical Analysis and Applications* 423(1), 701–719.
- Madan, D. B., P. Carr, and E. Chang (1998). The variance gamma process and option pricing. *Review of Finance* 2(1), 79–105.
- Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business* 36(4), 394–419.
- Pitman, J. and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 855–900.
- Roberts, G. O., O. Papaspiliopoulos, and P. Dellaportas (2004). Bayesian inference for non-Gaussian Ornstein–Uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(2), 369–393.
- Rosiński, J. (2007). Tempering stable processes. *Stochastic processes and their applications* 117(6), 677–707.
- Sato, K. (2001). Subordination and self-decomposability. *Statistics & probability letters* 54(3), 317–324.
- Šimšekli, U. (2017). Fractional Langevin Monte Carlo: Exploring Lévy driven stochastic differential equations for Markov chain Monte Carlo. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3200–3209. JMLR. org.
- Thorin, O. (1977). On the infinite divisibility of the lognormal distribution. *Scandinavian Actuarial Journal* 1977(3), 121–148.
- Winkel, M. (2005). Electronic foreign-exchange markets and passage events of independent subordinators. *Journal of applied probability* 42(1), 138–152.

- Woerner, J. (2011). Analyzing the fine structure of continuous time stochastic processes. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pp. 473–492. Springer.
- Zhang, A. and J. Paisley (2018). Deep Bayesian nonparametric tracking. In *International Conference on Machine Learning*, pp. 5828–5836.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	The Normal-Generalised Gamma-Pareto process: A novel pure-jump Lévy process with flexible tail and jump-activity properties
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Submitted to the Journal of Business and Economic Statistics and is under review

Student Confirmation

Student Name:	Fadhel Ayed		
Contribution to the Paper	I contributed in deriving the theoretical properties of the proposed model. I was also in charge of the conducting the experimentations for the exponentiated Lévy model.		
Signature	Mohamed El Fadhel Ayed	Date	12/07/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor F. Coron		
Supervisor comments			
Signature		Date	14/07/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

Part II

Estimation of rare events and missing mass in species sampling and feature allocation models

Abstract

This part is composed of three independent papers; The first paper has been submitted to the journal *Annales de l'Institut Henri Poincaré, Series B* and is under review. The second one has been published in the *Electronic Journal of Statistics*. The third one has been submitted to the *Journal of Machine Learning Research*, and is now under revision.

Table of Contents

5	On consistent and rate optimal estimation of the missing mass in species sampling	99
6	A Good-Turing estimator for feature allocation models	127
7	Consistent estimation of rare masses in feature allocation models	159

5

On consistent and rate optimal estimation
of the missing mass in species sampling

On consistent and rate optimal estimation of the missing mass

Fadhel Ayed¹, Marco Battiston², Federico Camerlenghi³ and Stefano Favaro⁴

¹*Department of Statistics, University of Oxford
OX1 3LB, Oxford, United Kingdom.
e-mail: fadhel.ayed@gmail.com*

²*Department of Mathematics and Statistics, University of Lancaster
LA1 4YF, Lancaster, United Kingdom.
e-mail: m.battiston@lancaster.ac.uk*

³*Department of Economics, Management and Statistics, University of Milano-Bicocca
20126, Milano, Italy.
e-mail: federico.camerlenghi@unimib.it*

⁴*Department of Economics and Statistics, University of Torino
10134, Torino, Italy.
e-mail: stefano.favaro@unito.it*

Abstract: Given n samples from a population of individuals belonging to different types with unknown proportions, how do we estimate the probability of discovering a new type at the $(n + 1)$ -th draw? This is a classical problem in statistics, commonly referred to as the missing mass estimation problem. Recent results have shown: i) the impossibility of estimating the missing mass without imposing further assumptions on type's proportions; ii) the consistency of the Good-Turing estimator of the missing mass under the assumption that the tail of type's proportions decays to zero as a regularly varying function with parameter $\alpha \in (0, 1)$; iii) the rate of convergence $n^{-\alpha/2}$ for the Good-Turing estimator under the class of $\alpha \in (0, 1)$ regularly varying P . In this paper we introduce an alternative, and remarkably shorter, proof of the impossibility of a distribution-free estimation of the missing mass. Beside being of independent interest, our alternative proof suggests a natural approach to strengthen, and expand, the recent results on the rate of convergence of the Good-Turing estimator under $\alpha \in (0, 1)$ regularly varying type's proportions. In particular, we show that the convergence rate $n^{-\alpha/2}$ is the best rate that *any* estimator can achieve, up to a slowly varying function. Furthermore, we prove that a lower bound to the minimax estimation risk must scale at least as $n^{-\alpha/2}$, which leads to conjecture that the Good-Turing estimator is a rate optimal minimax estimator under regularly varying type proportions.

MSC 2010 subject classifications: 62G05, 62C20.

Keywords and phrases: Good-Turing estimator, Minimax rate, Missing mass, Optimal rate of convergence, Regular variation, Two-parameter Poisson-Dirichlet.

⁴Also affiliated to Collegio Carlo Alberto

1. Introduction

Given n samples from a population of individuals belonging to different types with unknown proportions, how do we estimate the probability of discovering a new type at the $(n + 1)$ -th draw? This is a classical problem in statistics, commonly referred to as the missing mass estimation problem. It first appeared in ecology (e.g., Fisher et al. [9] and Good [15]), and its importance has grown in recent years driven by applications in biological and physical sciences (e.g., Kroes et al. [20], Gao et al. [11] and Ionita-Laza et al. [17]), machine learning and theoretical computer science (e.g., Motwani and Vassilvitskii [25] and Bubeck et al. [5]), and information theory (e.g., Orlitsky et al. [29] and Ben-Hamou et al. [3]). To introduce the missing mass estimation problem, let $P = \sum_{j \geq 1} p_j \delta_{\theta_j}$ be an unknown discrete distribution, where $(\theta_j)_{j \geq 1}$ are atoms on a measurable space and $(p_j)_{j \geq 1}$ are the corresponding probability masses, i.e. $p_j \in [0, 1]$ such that $\sum_{j \geq 1} p_j = 1$. If $\mathbf{X}_n = (X_1, \dots, X_n)$ is a collection of independent and identically distributed random variables from P , then we define the missing mass as

$$M_n(P, \mathbf{X}_n) = \sum_{j \geq 1} p_j \mathbb{1}(\theta_j \notin \mathbf{X}_n), \quad (1.1)$$

where $\mathbb{1}(\cdot)$ is the indicator function. The Good-Turing estimator (Good [15]) is arguably the most popular nonparametric estimator of $M_n(P, \mathbf{X}_n)$. It has been the subject of numerous studies. These include, e.g., asymptotic normality and large deviations (Zhang and Zhang [38], Gao [10] and Grabchak and Zhang [16]), admissibility and concentration properties (McAllester and Ortiz, [21], Ohannessian and Dahleh [27] and Ben-Hamou et al. [2]), consistency and convergence rates (McAllester Schapire [22], Wagner et al. [37] and Mossel and Ohannessian [24]), optimality and minimax properties (Orlitsky et al. [28] and Rajaraman et al. [35]).

Let $\hat{M}_n(\mathbf{X}_n)$ denote an estimator of $M_n(P, \mathbf{X}_n)$. Motivated by the recent works of Ohannessian and Dahleh [27], Mossel and Ohannessian [24], Ben-Hamou et al. [2] and Grabchak and Zhang [16], in this paper we consider the problem of consistent estimation of the missing mass under the multiplicative loss function

$$L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) = \left| \frac{\hat{M}_n(\mathbf{X}_n)}{M_n(P, \mathbf{X}_n)} - 1 \right|. \quad (1.2)$$

As discussed in Ohannessian and Dahleh [27], the loss function (1.2) is adequate for estimating small value parameters, since it allows to achieve more informative results. Such a multiplicative loss has been already used in statistics, for instance in the estimation of small value probabilities using importance sampling (Chatterjee and Diaconis [6]) and in the estimation of tail probabilities in extreme value theory (Beirlant and Devroye [1]). Under the loss function (1.2), Ohannessian and Dahleh [27] showed that: i) the Good-Turing estimator may be inconsistent; ii) the Good-Turing estimator is strongly consistent if the tail of P decays to zero as a regularly varying function with parameter $\alpha \in (0, 1)$

(e.g., Bingham et al. [4]). In particular, Grabchak and Zhang [16] showed that the convergence rate of the Good-Turing estimator is $n^{-\alpha/2}$, up to a slowly varying function. Mossel and Ohannessian [24] strengthened the inconsistency result of Ohannessian and Dahleh [27], showing the impossibility of estimating $M_n(P, \mathbf{X}_n)$ in a completely distribution-free fashion, that is without imposing further structural assumptions on P . In this paper we strengthen, and expand, the result of Grabchak and Zhang [16] on the rate of convergence of the Good-Turing estimator.

We start by introducing an alternative, and remarkably shorter, proof of the impossibility result of Mossel and Ohannessian [24]. Our proof relies on simple Bayesian nonparametric ideas, and in particular on the use the Ferguson-Dirichlet process prior (Ferguson [8]) for the unknown distribution P . The prior assumption on P allows us to avoid the winding constructive arguments of Mossel and Ohannessian [24], and prove the impossibility of estimating $M_n(P, \mathbf{X}_n)$ by exploiting well-known properties of the posterior distribution of $M_n(P, \mathbf{X}_n)$. Beside being of independent interest, our alternative proof suggests a natural approach to study rates of convergence of the Good-Turing estimator under the class of $\alpha \in (0, 1)$ regularly varying P . In particular, we make use of the two-parameter Poisson-Dirichlet process prior (Pitman and Yor [34]) for the unknown distribution P , which is known to generate (almost surely) discrete distributions whose tails decay to zero as a regularly varying function with parameter $\alpha \in (0, 1)$. See, e.g., Gneden et al. [13] and references therein. Under this prior assumption on P , we exploit properties of the posterior distribution of $M_n(P, \mathbf{X}_n)$ to prove that $n^{-\alpha/2}$ is the best rate of convergence that any estimator of the missing mass $M_n(P, \mathbf{X}_n)$ can achieve, up to a slowly varying function. Our result then strengthen the result of Grabchak and Zhang [16], showing the optimality, up to a slowly varying function, of the rate of convergence of the Good-Turing estimator. Finally, still relying on the two-parameter Poisson-Dirichlet process prior for the unknown P , we first investigate minimax rates for the Good-Turing estimator under the class of $\alpha \in (0, 1)$ regularly varying P . In particular, we show that the minimax rate for estimating $M_n(P, \mathbf{X}_n)$ must be at least $n^{-\alpha/2}$, and we conjecture that the Good-Turing estimator is an asymptotically optimal minimax estimator under the class of $\alpha \in (0, 1)$ regularly varying P .

The paper is structured as follows. In Section 2 we present the alternative proof of the result of Mossel and Ohannessian [24]. In Section 2 we state and prove our main results on the rate of convergence and on the minimax rate of the Good-Turing estimator under $\alpha \in (0, 1)$ regularly varying P . Section 4 contains a discussion of our results, and some open problems on the missing mass estimation and its generalizations. Auxiliary results and technical lemmas are deferred to Appendix A. The following notation is adopted throughout the paper: $[0, 1]$ is the unit interval, and $\mathcal{B}([0, 1])$ its Borel σ -algebra; $\mathcal{P}$ is the space of discrete distributions on $[0, 1]$, endowed with the smallest σ -algebra making $P \mapsto P(A)$ measurable for every $A \in \mathcal{B}([0, 1])$; P^n is the n -fold product of P on $[0, 1]$, and $\mathbb{E}_P$ the expectation with respect P ; for easiness of notation, we also use $\mathbb{E}_P$ to denote the expectation with respect to P^n ; ℓ is a slowly varying

function, i.e. a function satisfying $\ell(xc)/\ell(x) \rightarrow 1$ as $x \rightarrow \infty$ for every $c > 0$; C denotes a generic strictly positive constant; given a sequence of probabilities $(p_j)_{j \geq 1}$, $(p_{[j]})_{j \geq 1}$ is the corresponding ordered sequence, i.e. $p_{[1]} \geq p_{[2]} \geq \dots$; given two functions f and g , $f \sim g$ stands for $\lim f/g = 1$, $f = \mathcal{O}(g)$ for $\limsup |f|/|g| < C$, $f = o(g)$ for $\lim f/g = 0$; $\mathcal{B}(a, b)$ is the Beta integral of parameters a and b .

2. Impossibility of estimating $M_n(P, \mathbf{X}_n)$

Let $\mathbf{X}_n = (X_1, \dots, X_n)$ be a collection of independent and identically distributed random variables from an unknown discrete distribution P . The actual values taken by the observations, X_i 's, is not relevant for the missing mass estimation problem and, without loss of generality, they can be assumed to be values in the set $[0, 1]$. Therefore, $P(\cdot) = \sum_j p_j \delta_{\theta_j}(\cdot)$ is supposed to be a discrete distribution on the sample space $[0, 1]$, given a sequence of atoms $\theta_j \in [0, 1]$ and masses $p_j < 1$ such that $\sum_{j \geq 1} p_j = 1$. Both atoms and masses of the distribution P are assumed to be unknown. Given the sample $\mathbf{X}_n$, we are interested in estimating the missing $M_n(P, \mathbf{X}_n)$ defined in (1.1). The function $M_n(P, \mathbf{X}_n)$ is a jointly measurable function of P and $\mathbf{X}_n$, as proved in Proposition A.1. Given an estimator $\hat{M}_n(\mathbf{X}_n) : [0, 1]^n \rightarrow [0, 1]$ of $M_n(P, \mathbf{X}_n)$, we will measure its statistical performance by using the multiplicative loss function defined in (1.2). As we discussed in the introduction, this loss function is suitable to study theoretical properties of parameters or functionals taking small values, and it has already been used in previous works on missing mass estimation, e.g., Ohannessian and Dahleh [27], Mossel and Ohannessian [24], Ben-Hamou et al. [2] and Grabchak and Zhang [16].

A sequence of estimators $\hat{M}_n(\mathbf{X}_n)$ is said to be consistent for $M_n(P, \mathbf{X}_n)$, under parameter space $\mathcal{P}$ and the loss function L , if the loss incurred by the estimator converges in probability to zero under all points in the parameter space. Formally, $\hat{M}_n(\mathbf{X}_n)$ is consistent for $M_n(P, \mathbf{X}_n)$ if for all $P \in \mathcal{P}$ and for all $\epsilon > 0$,

$$P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) > \epsilon) \rightarrow 0 \quad (2.1)$$

as $n \rightarrow \infty$. Furthermore, $\hat{M}_n(\mathbf{X}_n)$ is strongly consistent if (2.1) is replaced by almost sure convergence. Under this setting Mossel and Ohannessian [24] proved Theorem 2.1 below. Here we present an alternative, and remarkably shorter, proof of Theorem 2.1. Our proof relies on Bayesian nonparametric ideas, and in particular on the use the Ferguson-Dirichlet process prior for the unknown distribution P .

Theorem 2.1. *Let $\mathcal{P}$ be the set of all discrete distributions on $[0, 1]$ and L be the loss function defined as (1.2). Then, there do not exist any consistent estimators for the missing mass $M_n(P, \mathbf{X}_n)$, i.e. there are no estimators $\hat{M}_n(\mathbf{X}_n)$ satisfying (2.1).*

Proof. We are going to show that for every estimator $\hat{M}_n(\mathbf{X}_n)$, there exists $\epsilon > 0$

such that

$$\sup_{P \in \mathcal{P}} \limsup_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) > \epsilon) > 0 \quad (2.2)$$

and, therefore, there exists $P \in \mathcal{P}$ such that $L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n))$ does not converge to zero in probability. First, note that for $\epsilon < 1/2$, Lemma (A.1) implies that

$$P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) > \epsilon) \geq P^n(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > 2\epsilon), \quad (2.3)$$

so it is sufficient to show that there exists $0 < \epsilon < 1$ such that for every estimator $\hat{M}_n(\mathbf{X}_n)$,

$$\sup_{P \in \mathcal{P}} \limsup_n P^n(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) > 0. \quad (2.4)$$

We prove that (2.4) holds for all $0 < \epsilon < 1/4$ (and therefore (2.2) holds for any $0 < \epsilon < 1/8$). Let $\epsilon \in (0, 1/4)$. Let DP_γ denote the Ferguson-Dirichlet process on $\mathcal{P}$ (Ferguson [8]), with base measure γ on $[0, 1]$. We choose γ uniform, i.e. $\gamma(d\theta) = \mathbb{1}(0 < d\theta < 1)$. Now, we can lower bound the supremum in (2.4) by an average over $\mathcal{P}$ with respect to DP_γ and then swap the integration by Fubini theorem, therefore

$$\begin{aligned} & \sup_{P \in \mathcal{P}} \limsup_n P^n(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) \\ & \geq \int_{\mathcal{P}} \limsup_n P^n(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) \text{DP}_\gamma(dP) \\ & \geq \limsup_n \int_{\mathcal{P}} \int_{[0,1]^n} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) P^n(d\mathbf{X}_n) \text{DP}_\gamma(dP) \\ & = \limsup_n \int_{[0,1]^n} \int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) \text{DP}_{\gamma + \sum_{i=1}^n \delta_{X_i}}(dP) P_{\text{DP}_\gamma}^n(d\mathbf{X}_n) \\ & \geq \limsup_n \int_{[0,1]^n} \inf_{x \geq 0} \int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), x) > \epsilon) \text{DP}_{\gamma + \sum_{i=1}^n \delta_{X_i}}(dP) d\mathbf{X}_n \end{aligned}$$

where the first inequality follows since we can lower bound the supremum by an average, the second from reverse Fatou's lemma, the equality comes by swapping the marginal of P and conditional of $\mathbf{X}_n$ given P with the marginal of $\mathbf{X}_n$, denoted $P_{\text{DP}_\gamma}^n$, and the conditional of P given $\mathbf{X}_n$, the last inequality follows since we are considering the infimum over all possible values of $\hat{M}_n(\mathbf{X}_n)$. Also recall that, when P is distributed as DP_γ , then the marginal of $\mathbf{X}_n$, $P_{\text{DP}_\gamma}^n$, is a Generalized Polya urn, while the conditional of P given $\mathbf{X}_n$ is $\text{DP}_{\gamma + \sum_{i=1}^n \delta_{X_i}}$ (see Theorem 4.6 and subsection 4.1.4 of Ghosal and Van der Vaart [12]). From Proposition A.2, $M_n(P, \mathbf{X}_n)$ under the posterior distribution $\text{DP}_{\gamma + \sum_{i=1}^n \delta_{X_i}}$ is distributed according to a Beta random variable $\text{Beta}(1, n)$. Therefore, we can write

$$\int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), x) > \epsilon) \text{DP}_{\gamma + \sum_{i=1}^n \delta_{X_i}}(dP) = \mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) \quad (2.5)$$

where $Z \sim \text{Beta}(1, n)$. We are now going to lower bound the probability of the event on the right hand side of (2.5). First let us consider $x \in (0, \frac{1}{1+\epsilon}]$ and $n \geq 2$, that is

$$\begin{aligned}
 \mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) &= \mathbb{P}(Z > (1 + \epsilon)x) + \mathbb{P}(Z < (1 - \epsilon)x) \\
 &= 1 + (1 - (1 + \epsilon)x)^n - (1 - (1 - \epsilon)x)^n \\
 &= 1 - 2x\epsilon \sum_{k=0}^{n-1} (1 - (1 + \epsilon)x)^{n-1-k} (1 - (1 - \epsilon)x)^k \quad (2.6) \\
 &\geq 1 - 2x\epsilon \sum_{k=0}^{n-1} (1 - (1 - \epsilon)x)^{n-1} \\
 &= 1 - 2x\epsilon n (1 - (1 - \epsilon)x)^{n-1} \\
 &\geq 1 - 2 \frac{\epsilon}{(1 - \epsilon)} n (1 - \epsilon)x (1 - (1 - \epsilon)x)^{n-1} \\
 &\geq 1 - 2 \frac{\epsilon}{(1 - \epsilon)} (1 - 1/n)^{n-1} \quad (2.7) \\
 &\geq 1 - \frac{2\epsilon}{(1 - \epsilon)}
 \end{aligned}$$

where we have used $a^n - b^n = (a - b) \sum_{k=0}^{n-1} a^{n-1-k} b^k$ in (2.6) and that the maximum of the function $x \mapsto x(1 - x)^{n-1}$ is achieved in $1/n$ in (2.7). Now let $x > \frac{1}{1+\epsilon}$, noticing that $\frac{2\epsilon}{(1+\epsilon)} < 1$, it comes that

$$\begin{aligned}
 \mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) &= \mathbb{P}(Z < (1 - \epsilon)x) \geq \mathbb{P}\left(Z < \frac{1 - \epsilon}{1 + \epsilon}\right) = 1 - 2^n \frac{\epsilon^n}{(1 + \epsilon)^n} \\
 &\geq 1 - \frac{2\epsilon}{(1 + \epsilon)} \geq 1 - \frac{2\epsilon}{(1 - \epsilon)}.
 \end{aligned}$$

Therefore, $\mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) \geq 1 - \frac{2\epsilon}{(1 - \epsilon)}$ for all $x \geq 0$ and $n \geq 2$. Plugging this estimate in place of (2.5), we obtain

$$\sup_{P \in \mathcal{P}} \limsup_n P^n(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) \geq 1 - \frac{2\epsilon}{(1 - \epsilon)}$$

and the right hand side is strictly positive for all $0 < \epsilon < 1/4$. $\square$

Mossel Ohannessian [24] proved Theorem 2.1 by exploiting a coupling of two generalized (dithered) geometric distributions. We presented an alternative proof of Theorem 2.1. While the proof of Mossel and Ohannessian [24] has the merit to be constructive, our approach has the merit to be direct by exploiting properties of the posterior distribution of $M_n(P, \mathbf{X}_n)$ under a Ferguson-Dirichlet prior for P . Beside being of independent interest, our alternative proof suggests a natural approach to study rates of convergence of the Good-Turing estimator

under the class of $\alpha \in (0, 1)$ regularly varying P . Indeed similar Bayesian arguments will be crucial in the next section to study the rate of convergence and the minimax rate of the Good-Turing estimator under the class of $\alpha \in (0, 1)$ regularly varying P .

Theorem 2.1 shows that for any asymptotic result to hold uniformly over a set of possible distributions, the parametric space $\mathcal{P}$ must be restricted to a suitable subclass. In particular, the proof of Theorem 2.1 shows that there are no consistent estimators for the class of distributions sampled from a Ferguson-Dirichlet process. This suggests that some conditions must be imposed on the tail decay of the elements of the parameter space. From Kingman [19] (Equation 65), we have that, if P is sampled from a Ferguson-Dirichlet process, its sequence of ordered masses behaves like $\log p_{[j]} \sim -jC$, as $j \rightarrow +\infty$. Therefore, the tail of P has approximately exponential form, resembling a geometric distribution and satisfying $p_{[j]} = o(j^{-\frac{1}{\alpha}})$ for every $\alpha \in (0, 1)$. Indeed, a geometric distribution was first used in Ohannessian and Dahleh [27] as an example to prove that the Good-Turing estimator may be inconsistent. Theorem 2.1 shows that, under this very light regime, any estimator of the missing mass, not just the Good-Turing, fails to be consistent under multiplicative loss. This motivates us to consider the class of P s having heavy enough tails. This will be the subject of the next section.

3. Consistent and minimax rate optimal estimation of $M_n(P, \mathbf{X}_n)$

We start by recalling the Good-Turing estimator (Good [15]) of the missing mass $M_n(P, \mathbf{X}_n)$, and then we study its convergence rate and minimax risk for regularly varying P . The definition of the Good-Turing estimator makes use of the proportion of unique values in the sample to estimate the missing mass. Let $Y_{n,j}(\mathbf{X}_n)$ be the number of times the value θ_j is observed in the sample $\mathbf{X}_n$, namely

$$Y_{n,j}(\mathbf{X}_n) = \sum_{i=1}^n \mathbb{1}(X_i = \theta_j).$$

Furthermore, let $K_{n,r}(\mathbf{X}_n)$ and $K_n(\mathbf{X}_n)$ be the number of values observed $1 \leq r \leq n$ times and the total number of distinct values, respectively, observed in $\mathbf{X}_n$, i.e.,

$$K_{n,r}(\mathbf{X}_n) = \sum_{j=1}^{\infty} \mathbb{1}(Y_{n,j} = r) \quad K_n(\mathbf{X}_n) = \sum_{r=1}^n K_{n,r}(\mathbf{X}_n).$$

The Good-Turing estimator of $M_n(P, \mathbf{X}_n)$ is defined in terms of the statistic $K_{n,1}(\mathbf{X}_n)$, that is

$$\hat{G}T(\mathbf{X}_n) = \frac{K_{n,1}(\mathbf{X}_n)}{n}. \quad (3.1)$$

Ohannessian and Dahleh [27] first showed the inconsistency of $\hat{G}T(\mathbf{X}_n)$ under the choice of P being a geometric distribution. In the same paper, it is shown

that under the assumption that the tail of P decays to zero as a regularly varying function with parameter $\alpha \in (0, 1)$, the Good-Turing estimator is strongly consistent. This latter result was generalized to the range $\alpha \in (0, 1]$ in Ben-Hamou et al. [2].

The assumption of regularly varying distribution P provides with a generalization of the power law tail decay, adding some more flexibility by the introduction of the slowly varying function ℓ . Power-law distributions are observed in the empirical distributions of many quantities in different applied areas, and their study have attracted a lot of interest in recent years. For extensive discussions of power laws in empirical data and their properties, the reader is referred to Mitzenmacher [23], Goldwater et al. [14], Newman [26], Clauset et al. [7] and Sornette [36]. Restricting the parameter space to probability distributions having regularly varying tail is not a mere technical assumption and, on the contrary, it represents a natural subset of the parameter space to consider, which we expect to contain the true data generating distribution for many different applications. To move into the concrete setting of regular variation (e.g., Bingham et al. [4]), for every $P \in \mathcal{P}$ we define a counting measure on $[0, 1]$ as $\nu_P(dx) = \sum_j \delta_{p_j}(dx)$, with corresponding tail function defined as $\vec{\nu}_P(x) = \nu([x, +\infty))$ for all $x > 0$. Then a distribution $P \in \mathcal{P}$ is said to be regularly varying with parameter $\alpha \in (0, 1)$ if

$$\vec{\nu}_P(x) \stackrel{x \downarrow 0}{\sim} x^{-\alpha} \ell(1/x), \quad (3.2)$$

where ℓ is a slowly varying function. From Lemma 22 and Proposition 23 of Gneden et al. [13], (3.2) is equivalent to the more explicit condition in term of ordered masses of P

$$p_{[j]} \stackrel{j \uparrow \infty}{\sim} j^{-1/\alpha} \ell_*(j), \quad (3.3)$$

where ℓ_* is a slowly varying function depending on ℓ . We denote by $\mathcal{P}_{RV_\alpha} \subseteq \mathcal{P}$ the set of all regularly varying distribution on $[0, 1]$ with parameter α . From (3.3) it is clear that such a class includes distributions having power-law tail decay, which correspond to the particular case of ℓ_* being a constant, which is equivalent to ℓ being constant. We denote the class of distributions having power law tail decay by $\mathcal{P}_{PL_\alpha} \subseteq \mathcal{P}_{RV_\alpha}$. In the following results, we will restrict our attention to the estimation problem under restricted parameter spaces $\mathcal{P}_{RV_\alpha}$ and $\mathcal{P}_{PL_\alpha}$.

Ohannessian and Dahleh [27] proved that the Good-Turing estimator is consistent under all distributions P in the class $\mathcal{P}_{RV_\alpha}$. Grabchak and Zhang [16] then proved that the convergence rate of the Good-Turing estimator is $n^{-\alpha/2}$, up to a slowly varying function. Hereafter we strengthen, and expand, the result of Grabchak and Zhang [16]. Recall that, under the loss function (1.2), a sequence $(r_n)_{n \in \mathbb{N}}$ is a convergence rate of an estimator $\hat{M}_n(\mathbf{X}_n)$ for the distribution $P \in \mathcal{P}$ if

$$\lim_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) > T_n r_n) = 0$$

for all sequences $T_n \rightarrow \infty$. For the sake of clarity, in the next proposition we state the result of Grabchak and Zhang [16] on the rate of convergence of

the Good-Turing estimator. An alternative proof of the next proposition follows from Proposition 3.2 below along with a straightforward application of Markov's inequality.

Proposition 3.1. *Let $\hat{GT}(\mathbf{X}_n)$ be the Good-Turing estimator, defined in (3.1). Then, for every $P \in \mathcal{P}_{RV_\alpha}$ and for all $T_n \rightarrow \infty$,*

$$\lim_n P^n(L(\hat{GT}(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) > T_n n^{-\alpha/2} \ell^{-1/2}(n)) = 0, \quad (3.4)$$

where ℓ in (3.4) is the slowly varying function specific to P appearing in (3.2). Therefore, up to slowly varying functions, $n^{-\alpha/2}$ is a convergence rate for the Good-Turing estimator within the class $\mathcal{P}_{RV_\alpha}$.

The next theorem strengthen Proposition 3.1 on the rate of convergence of the Good-Turing estimator. Indeed it shows that the convergence rate achieved by the Good-Turing estimator is actually almost the best convergence rate any estimator of $M_n(\mathbf{X}_n, P)$ can achieve. For any other estimator of the missing mass, it is possible to find a point $P \in \mathcal{P}_{PL_\alpha}$ for which the rate of convergence is not faster than $n^{-\alpha/2}$.

Theorem 3.1. *For any estimator $\hat{M}_n(\mathbf{X}_n)$, there exists $P \in \mathcal{P}_{PL_\alpha} \subset \mathcal{P}_{RV_\alpha}$ such that for every $T_n \rightarrow 0$*

$$\liminf_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) < T_n n^{-\alpha/2}) = 0. \quad (3.5)$$

Therefore the convergence rate of $\hat{M}_n(\mathbf{X}_n)$ cannot be faster than $n^{-\alpha/2}$.

Proof. Let $(T_n)_n$ any non-negative sequence converging to 0. We will show that for any estimator $\hat{M}_n(\mathbf{X}_n)$,

$$\inf_{P \in \mathcal{P}_{RV_\alpha}} \liminf_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) < T_n n^{-\alpha/2}) = 0. \quad (3.6)$$

Let us denote by SP_α the law of a stable process on $[0, 1]$ of parameter α . This a subordinator with Levy intensity, $\nu(d\omega) = \frac{\alpha}{\Gamma(1-\alpha)} \omega^{-1-\alpha} d\omega$. See Kingman [19] and Pitman [33] for details and additional references. Because of $\nu([x, \infty)) = \frac{x^{-\alpha}}{\Gamma(1-\alpha)}$, the stable process samples probability measures belonging to $\mathcal{P}_{PL_\alpha}$. Now we can upper bound the infimum in (3.6) by an average with respect to SP_α ,

$$\begin{aligned} & \inf_{P \in \mathcal{P}_{RV_\alpha}} \liminf_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) < T_n n^{-\alpha/2}) \\ & \leq \int_{\mathcal{P}} \liminf_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) < T_n n^{-\alpha/2}) SP_\alpha(dP) \\ & \leq \liminf_n \int_{\mathcal{P}} \int_{[0,1]^n} \mathbb{1}(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) < T_n n^{-\alpha/2}) P^n(d\mathbf{X}_n) SP_\alpha(dP) \end{aligned}$$

where the last equality follows by applying Fatou's Lemma.

Take n large enough so that $T_n n^{-\alpha/2} < 1/2$. Let us denote by $P_{SP_\alpha}^n$ the marginal law of the observations under an α -stable process, when P is integrated out, i.e. the probability measure on $[0, 1]^n$ defined as $P_{SP_\alpha}^n(A) = \int_{\mathcal{P}} P^n(A) SP_\alpha(dP)$ for all $A \in \mathcal{B}([0, 1]^n)$. We swap the integration of the marginal of P and the conditional of $\mathbf{X}_n$ given P with the marginal of $\mathbf{X}_n$ and the conditional of P given $\mathbf{X}_n$ and then apply Lemma A.1 to obtain

$$\begin{aligned} & \int_{\mathcal{P}} \int_{[0,1]^n} \mathbb{1}(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) < T_n n^{-\alpha/2}) P^n(d\mathbf{X}_n) SP_\alpha(dP) \\ &= \int_{[0,1]^n} \int_{\mathcal{P}} \mathbb{1}(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) < T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) P_{SP_\alpha}^n(d\mathbf{X}_n) \\ &\leq \int_{[0,1]^n} \int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) P_{SP_\alpha}^n(d\mathbf{X}_n) \end{aligned}$$

where $SP_\alpha|\mathbf{X}_n$ denotes the posterior of P given the sample. Therefore, taking $s > 1$, we can upper bound the quantity appearing on the l.h.s. of (3.6) by

$$\begin{aligned} & \limsup_n \int_{[0,1]^n} \mathbb{1}(K_n(\mathbf{X}_n) \in (n^\alpha/s, sn^\alpha)) \\ & \times \int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) P_{SP_\alpha}^n(d\mathbf{X}_n) \quad (3.7) \\ & + \limsup_n P_{SP_\alpha}^n(K_n(\mathbf{X}_n) \notin (n^\alpha/s, sn^\alpha)) \end{aligned}$$

We will upper-bound the two terms of the sum in (3.7) independently. Let us focus on the first term of (3.7). Let n large enough so that $3 < \frac{\alpha n^\alpha}{s} < \alpha s n^\alpha < n - 3$ and $T_n n^{-\alpha/2} < 1/4$. From Proposition A.2, under the posterior $SP_\alpha|\mathbf{X}_n$, $M_n(P, \mathbf{X}_n)$ is distributed according to a Beta random variable $\text{Beta}(\alpha K_n(\mathbf{X}_n), n - \alpha K_n(\mathbf{X}_n))$. Let us denote $a(\mathbf{X}_n) = \alpha K_n(\mathbf{X}_n)$, $b(\mathbf{X}_n) = n - \alpha K_n(\mathbf{X}_n)$, and for easiness of notation we will simply write a and b in the following calculations. Also let $F_{a,b}$ be the cumulative distribution function of the beta random variable $\text{Beta}(a, b)$. From Proposition A.2 in the Appendix, we have that

$$\begin{aligned} & \int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) \\ &= F_{a,b}((1 + 2T_n n^{-\alpha/2})\hat{M}_n(\mathbf{X}_n)) - F_{a,b}((1 - 2T_n n^{-\alpha/2})\hat{M}_n(\mathbf{X}_n)) \\ &\leq \sup_{x \in [0,1]} \left(F_{a,b}((1 + 2T_n n^{-\alpha/2})x) - F_{a,b}((1 - 2T_n n^{-\alpha/2})x) \right). \end{aligned}$$

Consider the function $\psi : \mathbb{R}_+ \rightarrow [0, 1]$ defined by

$$\psi(x) = F_{a,b}((1 + 2T_n n^{-\alpha/2})x) - F_{a,b}((1 - 2T_n n^{-\alpha/2})x).$$

Notice that $\psi \in \mathcal{C}^2$ and that $\psi(0) = \lim_{x \rightarrow \infty} \psi(x) = 0$. Therefore, ψ reaches its maximum in $x^*(a, b) \in \mathbb{R}_+$ (denoted x^* for easiness of notation) satisfying

$$\begin{aligned} \psi'(x^*) &= (1 + 2T_n n^{-\alpha/2}) f_{a,b}((1 + 2T_n n^{-\alpha/2})x^*) \\ &\quad - (1 - 2T_n n^{-\alpha/2}) f_{a,b}((1 - 2T_n n^{-\alpha/2})x^*) = 0, \end{aligned}$$

where $f_{a,b}$ denotes the density function of the Beta(a, b) distribution. On the event $K_n(\mathbf{X}_n) \in (n^\alpha/s, sn^\alpha)$, we have $a, b > 3$, and so $f_{a,b}$ is bell-shaped with second inflexion point,

$$\begin{aligned} \kappa(a, b) &= \frac{a-1}{a+b-2} + \frac{\sqrt{\frac{(a-1)(b-1)}{a+b-3}}}{a+b-2} \\ &\leq \frac{\alpha sn^\alpha}{n-2} + \frac{\sqrt{\alpha sn^\alpha}}{n-2} \\ &\leq \frac{2\alpha sn^\alpha}{n-2}. \end{aligned}$$

Therefore, $f'_{a,b}$ is non decreasing on the interval $[\kappa(a, b), \infty)$ and as a consequence, ψ'' is non negative on $[\frac{\kappa(a,b)}{(1-2T_n n^{-\alpha/2})}, \infty)$, from which we can deduce that ψ' is non decreasing on the same interval. Now, since $\lim_{x \rightarrow \infty} \psi'(x) = 0$, it follows that $\psi'(x) \leq 0$ on $[\frac{\kappa(a,b)}{(1-2T_n n^{-\alpha/2})}, \infty)$. Therefore,

$$x^*(a, b) \leq \frac{\kappa(a, b)}{(1 - 2T_n n^{-\alpha/2})} \leq \frac{2\alpha sn^\alpha}{(n-2)(1 - 2T_n n^{-\alpha/2})} \leq \frac{4\alpha sn^\alpha}{n-2}.$$

We can now upper-bound $\sup_{x \geq 0} \psi(x)$ as follows:

$$\sup_{x \geq 0} \psi(x) = \psi(x^*) \leq 4T_n n^{-\alpha/2} x^* \sup_{x \geq 0} f_{a,b}(x) \leq 16T_n n^{-\alpha/2} \frac{\alpha sn^\alpha}{n-2} \sup_{x \geq 0} f_{a,b}(x)$$

From Lemma A.2 in Appendix A, it follows that, on the event $K_n \in (n^\alpha/s, sn^\alpha)$, for n large enough,

$$\begin{aligned} \sup_{x \geq 0} \psi(x) &\leq 128T_n n^{-\alpha/2} \frac{\alpha sn^\alpha}{n-2} (a+b)^{3/2} a^{-1/2} b^{-1/2} \\ &\leq 128T_n n^{-\alpha/2} \frac{\alpha sn^\alpha}{n-2} n^{3/2} (\alpha n^\alpha/s)^{-1/2} (n - s\alpha n^\alpha)^{-1/2} = T_n g(\alpha, s, n) \end{aligned}$$

and notice that $\limsup_{n \rightarrow +\infty} T_n g(\alpha, s, n) = 0$.

From all previous computations, we can deduce that, on the event $K_n(\mathbf{X}_n) \in (1/sn^\alpha, sn^\alpha)$, we can find $n_0(\alpha, s)$ (n_0 does not depend on the value of K_n) such that for all $n \geq n_0(\alpha, s)$, the inequality

$$\int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha | \mathbf{X}_n(dP) \leq T_n g(\alpha, s, n)$$

holds, leading to

$$\begin{aligned} & \limsup_n \int_{[0,1]^n} \mathbb{1}(K_n(\mathbf{X}_n) \in (1/sn^\alpha, sn^\alpha)) \\ & \quad \times \int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha | \mathbf{X}_n(dP) P_{SP_\alpha}^n(d\mathbf{X}_n) \\ & \leq \limsup_n T_n g(\alpha, s, n) = 0. \end{aligned}$$

Therefore the first term in (3.7) is equal to zero. Let us consider the second term,

$$\limsup_n P_{SP_\alpha}^n(K_n(\mathbf{X}_n) \notin (n^\alpha/s, sn^\alpha)).$$

From Theorem 3.8 of Pitman [33], under the α -stable process, $\frac{K_n(\mathbf{X}_n)}{n^\alpha} \rightarrow S_\alpha$ almost surely, where S_α is a random variable on $\mathbb{R}_+$ distributed according to a Stable distribution of parameter α . Therefore,

$$\limsup_n P_{SP_\alpha}^n(K_n(\mathbf{X}_n) \notin (n^\alpha/s, sn^\alpha)) = \mathbb{P}(S_\alpha \notin (1/s, s)). \quad (3.8)$$

Taking $s \rightarrow \infty$, (3.8) converges to zero, and then so does (3.7). $\square$

Proposition 3.1 and Theorem 3.1 together show that the Good-Turing estimator achieves the best convergence rate up to a slowly varying function. In particular, if the distribution P has a power-law decay, i.e. $P \in \mathcal{P}_{PL_\alpha}$, the two rates match and the Good-Turing estimator achieves the best rate possible. In particular, because $\hat{GT}(\mathbf{X}_n)$ does not depend on α , it follows that the Good-Turing estimator is actually *rate adaptive* for the class of power law distributions, $\mathcal{P}_{PL} = \cup_{0 < \alpha < 1} \mathcal{P}_{PL_\alpha}$.

The next theorem considers the asymptotic minimax estimation risk for the missing mass under the loss function (1.2) and with parameter space $\mathcal{P}_{PL_\alpha}$. In particular Theorem 3.2 provides with a lower bound for the estimation risk of this statistical problem, showing that the minimax rate is not smaller than $n^{-\alpha/2}$.

Theorem 3.2. *Let $\mathcal{P}_{PL_\alpha}$ be the class of discrete distributions on $[0, 1]$ with power law tail function and let L denote the multiplicative loss function (1.2). Then, there exists a positive constant $C > 0$ such that*

$$\liminf_n n^{\alpha/2} \inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P \in \mathcal{P}_{PL_\alpha}} \mathbb{E}_P \left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) \right) > C$$

where the infimum is taken over all possible estimators $\hat{M}_n(\mathbf{X}_n)$.

Proof. In the following, we make use of the generic notation C to refer to constants that can only depend on α (its value can change from a line to the other). As in the proof of Theorem 3.1, let SP_α denote the law of a stable process of parameter α and $P_{SP_\alpha}^n$ the marginal law of the observations under this prior.

We can lower bound the minimax risk by the Bayesian risk with prior SP_α . Indeed,

$$\begin{aligned}
 & \inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P \in \mathcal{P}_{PL_\alpha}} \mathbb{E}_P \left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) \right) \\
 & \geq \inf_{\hat{M}_n(\mathbf{X}_n)} \int_{\mathcal{P}} \mathbb{E}_P \left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) \right) SP_\alpha(dP) \\
 & = \inf_{\hat{M}_n(\mathbf{X}_n)} \int_{[0,1]^n} \int_{\mathcal{P}} L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) SP_\alpha|_{\mathbf{X}_n}(dP) P_{SP_\alpha}^n(d\mathbf{X}_n) \\
 & \geq \int_{[0,1]^n} \inf_{\hat{M}_n(\mathbf{X}_n)} \int_{\mathcal{P}} L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) SP_\alpha|_{\mathbf{X}_n}(dP) P_{SP_\alpha}^n(d\mathbf{X}_n) \quad (3.9)
 \end{aligned}$$

From Proposition A.3 in Appendix A, the posterior distribution of missing mass $M_n(P, \mathbf{X}_n)$ under SP_α is distributed according to $\text{Beta}(\alpha K_n(\mathbf{X}_n), n - \alpha K_n(\mathbf{X}_n))$. Let $a(\mathbf{X}_n) = \alpha K_n(\mathbf{X}_n)$ and $b(\mathbf{X}_n) = n - \alpha K_n(\mathbf{X}_n)$, and for easiness of notation we will simply write a and b in the following calculations. The inner integral in (3.9) equals

$$\begin{aligned}
 & \int_{\mathcal{P}} L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) SP_\alpha|_{\mathbf{X}_n}(dP) \\
 & = \int_0^1 \frac{|\hat{M}_n(\mathbf{X}_n) - x|}{x} \frac{x^{a-1}(1-x)^{b-1}}{\mathcal{B}(a, b)} dx \\
 & = \frac{\mathcal{B}(a-1, b)}{\mathcal{B}(a, b)} \int_0^1 |\hat{M}_n(\mathbf{X}_n) - x| \frac{x^{a-2}(1-x)^{b-1}}{\mathcal{B}(a-1, b)} dx \\
 & = \frac{\mathcal{B}(a-1, b)}{\mathcal{B}(a, b)} \mathbb{E}_{M'} (|M' - \hat{M}_n(\mathbf{X}_n)|)
 \end{aligned}$$

where M' is a random variable distributed according to $\text{Be}(a-1, b)$. Plugging this quantity inside (3.9) we find

$$\begin{aligned}
 & \inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P \in \mathcal{P}_{PL_\alpha}} \mathbb{E}_P \left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) \right) \\
 & \geq \int_{[0,1]^n} \frac{\mathcal{B}(a-1, b)}{\mathcal{B}(a, b)} \inf_{\hat{M}_n(\mathbf{X}_n)} \mathbb{E}_{M'} (|M' - \hat{M}_n(\mathbf{X}_n)|) P_{SP_\alpha}^n(d\mathbf{X}_n) \\
 & = \int_{[0,1]^n} \frac{\mathcal{B}(a-1, b)}{\mathcal{B}(a, b)} \mathbb{E}_{M'} (|M' - \text{med}(M')|) P_{SP_\alpha}^n(d\mathbf{X}_n) \quad (3.10)
 \end{aligned}$$

where $\text{med}(M')$ denotes the median of M' . Now, let us denote by $f_{a,b}$ and $m_{a,b}$ the density function and the median of a Beta random variable of parameters

a and b . We can rewrite the inner expectation in (3.10) as

$$\begin{aligned}
 \mathbb{E}_{M'}(|M' - m_{a-1,b}|) &= \int_0^1 |x - m_{a-1,b}| f_{a-1,b}(x) dx \\
 &= \int_0^{m_{a-1,b}} |x - m_{a-1,b}| f_{a-1,b}(x) dx + \int_{m_{a-1,b}}^1 |x - m_{a-1,b}| f_{a-1,b}(x) dx \\
 &= - \int_0^{m_{a-1,b}} (x - m_{a-1,b}) f_{a-1,b}(x) dx + \int_{m_{a-1,b}}^1 (x - m_{a-1,b}) f_{a-1,b}(x) dx \\
 &= \int_{m_{a-1,b}}^1 x f_{a-1,b}(x) dx - \frac{1}{2} m_{a-1,b} - \int_0^{m_{a-1,b}} x f_{a-1,b}(x) dx + \frac{1}{2} m_{a-1,b} \\
 &= \int_{m_{a-1,b}}^1 x f_{a-1,b}(x) dx - \int_0^{m_{a-1,b}} x f_{a-1,b}(x) dx
 \end{aligned}$$

Therefore, the term inside the integral in (3.10) is

$$\begin{aligned}
 &\frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \mathbb{E}_{M'}(|M' - m_{a-1,b}|) \\
 &= \frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \int_{m_{a-1,b}}^1 x f_{a-1,b}(x) dx - \frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \int_0^{m_{a-1,b}} x f_{a-1,b}(x) dx \\
 &= \int_{m_{a-1,b}}^1 f_{a,b}(x) dx - \int_0^{m_{a-1,b}} f_{a,b}(x) dx \\
 &= \int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx + \int_{m_{a,b}}^1 f_{a,b}(x) dx - \int_0^{m_{a,b}} f_{a,b}(x) dx - \int_{m_{a,b}}^{m_{a-1,b}} f_{a,b}(x) dx \\
 &= \int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx + \frac{1}{2} - \frac{1}{2} - \int_{m_{a,b}}^{m_{a-1,b}} f_{a,b}(x) dx \\
 &= \int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx + \int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx = 2 \int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx
 \end{aligned}$$

and so we have,

$$\begin{aligned}
 &\inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P \in \mathcal{P}_{PL_\alpha}} \mathbb{E}_P \left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) \right) \\
 &\geq 2 \int_{[0,1]^n} \left(\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx \right) P_{S_{P_\alpha}}^n(d\mathbf{X}_n) \quad (3.11)
 \end{aligned}$$

where we recall that $a = \alpha K_n(\mathbf{X}_n)$ and $b = n - \alpha K_n(\mathbf{X}_n)$.

From Theorem 3.8 of Pitman [33], when P is distributed according to the α -stable process, $\frac{K_n(\mathbf{X}_n)}{n^\alpha} \xrightarrow{d} S_\alpha$, where S_α is a random variable on $[0, \infty)$ with Stable distribution of parameter α . Therefore, there exist n_0 and two positive bounded values w_α and W_α such that for all $n > n_0$, $\mathbb{P}(\frac{K_n(\mathbf{X}_n)}{n^\alpha} \in [w_\alpha, W_\alpha]) \geq \frac{1}{2}$ and $8 < 2\alpha w_\alpha n^\alpha \leq 2\alpha W_\alpha n^\alpha < n - 1$.

We are now ready to lower bound (3.11). We will make use of some technical lemmas regarding the density and median of the Beta distribution, whose statements and proofs are in Appendix A. (3.11) can be lower bounded by the following quantity

$$2\mathbb{P}(w_\alpha n^\alpha \leq K_n(\mathbf{X}_n) \leq W_\alpha n^\alpha) \times \mathbb{E}_{P_{SP_\alpha}^n} \left[\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx \mid w_\alpha n^\alpha \leq K_n(\mathbf{X}_n) \leq W_\alpha n^\alpha \right] \quad (3.12)$$

Given our choice of n_0 , w_α and W_α , for $n > n_0$, we have that $\mathbb{P}(w_\alpha n^\alpha \leq K_n \leq W_\alpha n^\alpha) \geq \frac{1}{2}$. Recall now that $a = \alpha K_n(\mathbf{X}_n)$ and $b = n - \alpha K_n(\mathbf{X}_n)$. Hence, noticing that on the event we are conditioning on, $3 < a < b$ and $a < b/2$, by applying Lemma A.3, we can lower bound (3.12) by

$$2\frac{1}{2}\mathbb{E}_{P_{SP_\alpha}^n} \left[\frac{C}{\sqrt{a}} \mid w_\alpha n^\alpha \leq K_n(\mathbf{X}_n) \leq W_\alpha n^\alpha \right],$$

for some strictly positive constant C . Ultimately this leads to

$$\begin{aligned} & \inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P \in \mathcal{P}_{RV_\alpha}} \mathbb{E}_P \left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) \right) \\ & \geq C \mathbb{E}_{P_{SP_\alpha}^n} \left[a^{-1/2} \mid w_\alpha n^\alpha \leq K_n(\mathbf{X}_n) \leq W_\alpha n^\alpha \right] \\ & = C \alpha^{-1/2} \mathbb{E}_{P_{SP_\alpha}^n} \left[K_n^{-1/2} \mid w_\alpha n^\alpha \leq K_n(\mathbf{X}_n) \leq W_\alpha n^\alpha \right] \\ & \geq C (w_\alpha n^\alpha)^{-1/2} = C n^{-\alpha/2}, \end{aligned}$$

which provides the lower bound rate for the minimax risk,

$$\liminf_n n^{\alpha/2} \inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P \in \mathcal{P}_{PL_\alpha}} \mathbb{E}_P \left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) \right) > C$$

□

The lower bound of Theorem 3.2 can be used to derive the minimax rate, by matching it with appropriate upper bounds of specific estimators of the missing mass. This lower bound trivially still holds for any parametric set larger than $\mathcal{P}_{PL_\alpha}$ and, therefore, the theorem also provides with a lower bound of the estimation risk under the larger parameter space $\mathcal{P}_{RV_\alpha}$. In the next Proposition, we show that for a fixed distribution $P \in \mathcal{P}_{RV_\alpha}$, the Good-Turing estimator achieves the best possible rate of Theorem 3.2 up to a slowly varying term.

Proposition 3.2. *Let $\hat{GT}(\mathbf{X}_n)$ be the Good-Turing estimator and let $P \in \mathcal{P}_{RV_\alpha}$. Then, there exists a finite constant C such that for every n*

$$\mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), M_n(P, \mathbf{X}_n))) \leq C n^{-\alpha/2} \ell^{-1/2}(n), \quad (3.13)$$

where ℓ is the slowly varying function specific to P appearing in (3.2).

Proof. Let $P \in \mathcal{P}_{RV_\alpha}$ and ℓ defined as in (3.2). In the following, we use the generic notation C and C' to refer to constants that can only depend on P (their values can change from a line to the other). Here we study the convergence rate under the assumption of regular variation of the Good-Turing estimator, $\hat{GT}(\mathbf{X}_n) = \frac{K_{n,1}(\mathbf{X}_n)}{n}$, proving that

$$\mathbb{E}_P \left(L(\hat{GT}(\mathbf{X}_n), M_n(P)) \right) = O(n^{-\alpha/2} \ell(n)^{-1/2}).$$

Let us first notice that for $a \geq 0$, $b, c > 0$

$$L(a, c) = \left| \frac{a}{b} \left(\frac{b}{c} - 1 \right) + \frac{a}{b} - 1 \right| \leq L(a, b) + \frac{a}{b} L(b, c).$$

Therefore, we can upper bound the loss of $\hat{GT}(\mathbf{X}_n) = \frac{K_{n,1}(\mathbf{X}_n)}{n}$ by

$$\begin{aligned} & L(\hat{GT}(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) \\ & \leq L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n))) + \frac{\hat{GT}(\mathbf{X}_n)}{\mathbb{E}_P(\hat{GT}(\mathbf{X}_n))} L(\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)), \mathbb{E}_P(M_n(P, \mathbf{X}_n))) \\ & \quad + \frac{\hat{GT}(\mathbf{X}_n)}{\mathbb{E}_P(M_n(P, \mathbf{X}_n))} L(\mathbb{E}_P(M_n(P, \mathbf{X}_n)), M_n(P, \mathbf{X}_n)), \end{aligned}$$

and consequently its risk by

$$\begin{aligned} & \mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), M_n(P, \mathbf{X}_n))) \leq \mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n)))) \\ & \quad + L(\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)), \mathbb{E}_P(M_n(P, \mathbf{X}_n))) \\ & \quad + \mathbb{E}_P \left(\frac{\hat{GT}(\mathbf{X}_n)}{\mathbb{E}_P(M_n(P, \mathbf{X}_n))} L(\mathbb{E}_P(M_n(P, \mathbf{X}_n)), M_n(P, \mathbf{X}_n)) \right). \end{aligned} \quad (3.14)$$

We will now separately upper bound the three components of the r.h.s. of this inequality. Let us first focus on the quantity $L(\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)), \mathbb{E}_P(M_n(P, \mathbf{X}_n)))$. From the work of Karlin [18] (see also Theorem 4.2 of Ben-Hamou et al. [2]), we know that

$$\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)) \sim \alpha \Gamma(1 - \alpha) n^{\alpha-1} \ell(n),$$

as $n \rightarrow \infty$ and since $0 \leq \mathbb{E}_P(\hat{GT}(\mathbf{X}_n)) - \mathbb{E}_P(M_n(P, \mathbf{X}_n)) \leq \frac{1}{n}$, we deduce that

$$L(\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)), \mathbb{E}_P(M_n(P, \mathbf{X}_n))) \leq C n^{-\alpha} \ell(n)^{-1}. \quad (3.15)$$

Let us now consider the first term in the r.h.s. of (3.14),

$$\mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n)))) = \mathbb{E}_P \left(\left| \frac{K_{n,1}(\mathbf{X}_n)}{\mathbb{E}_P(K_{n,1}(\mathbf{X}_n))} - 1 \right| \right). \quad (3.16)$$

As a result of Ben-Hamou et al. [2] Proposition 3.5 (see also the proof of corollary 5.3 of the same paper), for every $\epsilon > 0$, we have

$$P^n(L(K_{n,1}(\mathbf{X}_n), \mathbb{E}_P(K_{n,1}(\mathbf{X}_n))) \geq \epsilon) \leq 4e^{-\epsilon^2 A_n^2},$$

where

$$A_n = \frac{\mathbb{E}(K_{n,1}(\mathbf{X}_n))}{\sqrt{8(\mathbb{E}(K_{n,1}(\mathbf{X}_n)) \vee 2\mathbb{E}(K_{n,2}(\mathbf{X}_n))) + 4/3}}.$$

Hence, we can now bound (3.16) as follows

$$\begin{aligned} \mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n)))) &= \mathbb{E}_P(L(K_{n,1}(\mathbf{X}_n), \mathbb{E}_P(K_{n,1}(\mathbf{X}_n)))) \\ &= \int_0^\infty P^n(L(K_{n,1}(\mathbf{X}_n), \mathbb{E}_P(K_{n,1}(\mathbf{X}_n))) \geq \epsilon) d\epsilon \\ &\leq 4 \int_0^\infty e^{-\epsilon^2 A_n^2} d\epsilon = \frac{4}{A_n} \int_0^\infty e^{-y^2} dy = C A_n^{-1}, \end{aligned}$$

where we have used the change of variables $y = \epsilon A_n$. Therefore, from the asymptotic behaviors of $\mathbb{E}_P(K_{n,1}(\mathbf{X}_n))$ and $\mathbb{E}_P(K_{n,2}(\mathbf{X}_n))$ that are provided in the work of Karlin [18] (see also Theorem 4.2 of Ben-Hamou et al. [2]), we conclude that

$$\mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n)))) \leq C n^{-\alpha/2} \ell(n)^{-1/2}. \quad (3.17)$$

Finally, let us look at the third term in (3.14),

$$\mathbb{E}_P \left(\frac{\hat{GT}(\mathbf{X}_n)}{\mathbb{E}_P M_n(P, \mathbf{X}_n)} L(\mathbb{E}_P(M_n(P, \mathbf{X}_n)), M_n(P, \mathbf{X}_n)) \right). \quad (3.18)$$

Notice that (3.18) is equal to

$$\mathbb{E}_P \left(\frac{\hat{GT}(\mathbf{X}_n)}{M_n(P, \mathbf{X}_n)} L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n))) \right), \quad (3.19)$$

and from the Cauchy-Schwarz inequality, we can upper bound (3.19) by

$$\sqrt{\mathbb{E}_P \left(\frac{\hat{GT}(\mathbf{X}_n)^2}{M_n(P, \mathbf{X}_n)^2} \right)} \sqrt{\mathbb{E}_P (L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n)))^2)}. \quad (3.20)$$

We will first compute the asymptotic behavior of the second term in (3.20) and then show that the first term is asymptotically bounded. By applying Theorem 3.9 of Ben-Hamou et al. [2] and the asymptotic regimes of Karlin [18], we obtain that for every $\epsilon > 0$,

$$P^n(L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n))) \geq \epsilon) \leq 2e^{-\epsilon^2 B_n},$$

where

$$B_n \leq C n^\alpha \ell(n)$$

(see for example the proof of Corollary 5.3 of Ben-Hamou et al. [2]). Therefore, for all $\epsilon > 0$

$$P^n(L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n)))^2 \geq \epsilon) \leq 2e^{-\epsilon B_n}$$

and, following the same reasoning used before, we obtain

$$\mathbb{E}_P \left(L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n)))^2 \right) \leq CB_n^{-1}$$

which leads to

$$\sqrt{\mathbb{E}_P \left(L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n)))^2 \right)} \leq Cn^{-\alpha/2} \ell(n)^{-1/2}. \quad (3.21)$$

It remains only to prove that the first term in (3.20) is bounded. First note that

$$\frac{\hat{GT}(\mathbf{X}_n)}{M_n(P, \mathbf{X}_n)} \leq \frac{K_n(\mathbf{X}_n)}{n \sum_{j > K_n(\mathbf{X}_n)} p[j]}$$

For $t \geq 1$, let us define the function f by $f(t) = \frac{t}{\sum_{j > t} p[j]}$. Noticing that $f(t) \geq 1 > 0$ we can write

$$\frac{\hat{GT}(\mathbf{X}_n)}{M_n(P, \mathbf{X}_n)} \leq \frac{f(K_n(\mathbf{X}_n))}{f(\mathbb{E}(K_n(\mathbf{X}_n)))} \frac{f(\mathbb{E}(K_n(\mathbf{X}_n)))}{n}. \quad (3.22)$$

Denoting by $\ell_\alpha^{\frac{1}{\alpha}\#}$ the de Bruijn conjugate of $\ell_\alpha^{\frac{1}{\alpha}}$ (see subsection 1.5.7 of Bingham et al. [4] for a definition), Proposition 23 of Gneden et al. [13] implies the following

$$f(t) \sim Ct^{\frac{1}{\alpha}} \ell_\alpha^{\frac{1}{\alpha}\#}(t^{\frac{1}{\alpha}}),$$

which in turns implies that f^2 is regularly varying with index $2/\alpha$. Since f is non decreasing, it is bounded on any set of the form $[1, T]$. Therefore, we can apply Potter's Theorem (Theorem 1.5.6, Bingham et al. [4]) in order to obtain that

$$\frac{f(K_n(\mathbf{X}_n))^2}{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))^2} \leq C \left[\left(\frac{K_n(\mathbf{X}_n)}{\mathbb{E}_P(K_n(\mathbf{X}_n))} \right)^{\frac{1}{\alpha}} + \left(\frac{K_n(\mathbf{X}_n)}{\mathbb{E}_P(K_n(\mathbf{X}_n))} \right)^{\frac{3}{\alpha}} \right] + C'.$$

Following the same lines of reasoning that we used before, we can show that for all $\eta > 1$

$$\lim_{n \rightarrow +\infty} \mathbb{E}_P [L(K_n(\mathbf{X}_n), \mathbb{E}_P(K_n(\mathbf{X}_n)))^\eta] = 0, \quad (3.23)$$

and, thanks to the elementary inequality $x/|x-1| \leq 2$ for $x \geq 2$, it follows that, for all $\eta > 1$,

$$\left(\frac{K_n(\mathbf{X}_n)}{\mathbb{E}_P(K_n(\mathbf{X}_n))} \right)^\eta \leq 2^\eta + 2^\eta L(K_n(\mathbf{X}_n), \mathbb{E}_P(K_n(\mathbf{X}_n)))^\eta.$$

As a consequence of this last inequality, along with (3.23), for all $\eta > 1$ we obtain

$$\mathbb{E}_P \left(\left(\frac{K_n(\mathbf{X}_n)}{\mathbb{E}_P(K_n(\mathbf{X}_n))} \right)^\eta \right) \leq C$$

from which we get that

$$\mathbb{E}_P \left(\left(\frac{f(K_n(\mathbf{X}_n))}{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))} \right)^\eta \right) \leq C. \quad (3.24)$$

Besides, since $\mathbb{E}_P(K_n(\mathbf{X}_n))^{\frac{1}{\alpha}} \sim n\ell^{\frac{1}{\alpha}}(n)$, which diverges to infinity, the uniform convergence theorem for slowly varying functions (Theorem 1.2.1, Bingham et al. [4]) gives that

$$\ell^{\frac{1}{\alpha}\#}(\mathbb{E}_P(K_n(\mathbf{X}_n))^{\frac{1}{\alpha}}) \sim \ell^{\frac{1}{\alpha}\#}(n\ell^{\frac{1}{\alpha}}(n)).$$

As a consequence of this and of the asymptotic properties of f , we obtain that

$$\frac{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))}{n} \sim \ell^{\frac{1}{\alpha}}(n)\ell^{\frac{1}{\alpha}\#}(n\ell^{\frac{1}{\alpha}}(n)),$$

which, from the definition of the de Bruijn conjugate, in turn gives

$$\frac{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))}{n} \sim 1,$$

and then

$$\frac{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))^2}{n^2} \leq C. \quad (3.25)$$

From (3.22), (3.24) and (3.25) together, we finally obtain

$$\mathbb{E}_P \left(\frac{\hat{G}T(\mathbf{X}_n)^2}{M_n(P, \mathbf{X}_n)^2} \right) \leq C,$$

which together with (3.15), (3.17) and (3.21) concludes the proof. $\square$

Extending Proposition 3.2 to hold uniformly over the class $\mathcal{P}_{RV_\alpha}$ is an open problem and probably requires a careful control over the size of $\mathcal{P}_{RV_\alpha}$. Indeed, the classes of distributions we are considering are defined through the asymptotic properties of their elements, while to obtain minimax results we need a control for each $n \in \mathbb{N}$. Even though Proposition 3.2 does not directly provide with the minimax rate of the Good-Turing estimator, it still provides with a sanity check for its asymptotic risk. Specifically, Proposition 3.2 implies that for every $P \in \mathcal{P}_{PL_\alpha}$,

$$\limsup_n n^{\alpha/2} \mathbb{E}_P(L(\hat{G}T(\mathbf{X}_n), M_n(P, \mathbf{X}_n))) < +\infty$$

Moreover, from a minor change at the beginning of the proof of Theorem 3.2, we can also prove that for every estimator $\hat{M}_n(\mathbf{X}_n)$ and every sequence $(T_n)_n$ diverging to infinity, we can find an element $P \in \mathcal{P}_{PL_\alpha}$ such that one has $\limsup_n T_n n^{\alpha/2} \mathbb{E}_P(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n))) = +\infty$. This leads us to conjecture that the Good-Turing estimator is a rate optimal minimax estimator under the loss function (1.2).

4. Discussion

In this paper we have considered the problem of consistent estimation of the missing mass under a suitable multiplicative loss function. We have presented an alternative, and simpler, proof of the result by Mossel and Ohannessian [24] on the impossibility of a distribution-free estimation of the missing mass. Our results relies on novel arguments from Bayesian nonparametric statistics, which are then exploited to study convergence rates and minimax rates of the Good-Turing estimator under the class of $\alpha \in (0, 1)$ regularly varying P . In Proposition 3.1 and Theorem 3.1 it has been shown that, within the class $\mathcal{P}_{PL_\alpha}$, the Good-Turing estimator achieves the best convergence rate possible, while for the class, $\mathcal{P}_{RV_\alpha}$, this rate is the best up to a slowly varying function. An open problem is to understand whether this additional slowly varying term is intrinsic to the problem or our results can actually be improved to make the rate of the Good-Turing estimator matches the best possible rate also within the class of regularly varying distributions. Under the restricted parametric spaces, in Theorem 3.2 we have provided a lower bound for the asymptotic risk. This bound can be used to compare estimators from a minimax point of view, by finding suitable upper bounds matching the lower bound rate. In particular, in Proposition 3.2 we have shown that the asymptotic rate of the risk of the Good-Turing estimator matches the lower bound rate, up to a slowly varying function. However, the rate of Proposition 3.2 is a pointwise result, for a fixed $P \in \mathcal{P}_{RV_\alpha}$. An open problem is to extend Proposition 3.2 to the uniform case, when considering the supremum of the risk over all $P \in \mathcal{P}_{RV_\alpha}$. This extension probably requires a careful analysis and control of the size of this parameter space $\mathcal{P}_{RV_\alpha}$. Work on this is ongoing.

Appendix A: Appendix

Lemma A.1. *For $\epsilon < 1/2$ and $a, b \geq 0$, $L(a, b) \leq \epsilon$ implies $L(b, a) \leq 2\epsilon$.*

Proof. Let a, b be positive real numbers, $\epsilon < 1/2$ and suppose $L(a, b) = |\frac{a}{b} - 1| \leq \epsilon$. Straightforwardly, have that

$$-b\epsilon \leq a - b \leq b\epsilon \quad (\text{A.1})$$

From the lower bound of (A.1), $a \geq (1 - \epsilon)b \geq b/2$ and, therefore, $\frac{1}{a} \leq \frac{2}{b}$. Multiplying (A.1) by this last inequality, we have $L(b, a) \leq 2\epsilon$. $\square$

Proposition A.1. *The missing mass $M_n(P, \mathbf{X}_n)$ is a jointly measurable mapping.*

Proof. Recall that $\mathcal{P}$ is endowed with the smallest σ -algebra making the mappings $P \mapsto P(A)$ measurable for every $A \in \mathcal{B}([0, 1])$. This is also the Borel σ -algebra generated by the weak convergence topology, which can be induced by the *bounded Lipschitz metric* (see Appendix A of Ghosal and Van der Vaart

[12]), defined as

$$d_{BL}(P, Q) = \sup_{\|f\|_{C^1} \leq 1} \left| \int f dP - \int f dQ \right|$$

where the supremum is over all real functions satisfying $|f(x) - f(y)| \leq |x - y|$ for any $x, y \in [0, 1]$. $[0, 1]^n$ is endowed with the Euclidean topology, which can be induced by the ℓ_∞ norm.

Let us consider $O_n(P, \mathbf{X}_n) = 1 - M_n(P, \mathbf{X}_n)$ and let us define for any $\eta > 0$,

$$f_{\eta, X_i} : x \mapsto \max(0, \eta - |X_i - x|).$$

Also let $f_{\eta, \mathbf{X}_n} = \max_i f_{\eta, X_i}$, which is 1-Lipschitz function, since all f_{η, X_i} are 1-Lipschitz. Now, let $O_{\eta, n}$ be defined as follows

$$O_{\eta, n}(P, \mathbf{X}_n) = \frac{1}{\eta} \sum_{j \geq 1} p_j f_{\eta, \mathbf{X}_n}(\theta_j) = \int \frac{1}{\eta} f_{\eta, \mathbf{X}_n} dP.$$

Let $(P, \mathbf{X}_n) \in \mathcal{P} \times [0, 1]^n$. We have that

$$\lim_{\eta \rightarrow 0} O_{\eta, n}(P, U_{1:n}) = O_n(P, U_{1:n}). \quad (\text{A.2})$$

Indeed, notice that for any $x \in [0, 1]$,

$$\frac{f_{\eta, \mathbf{X}_n}(x)}{\eta} > 0 \Leftrightarrow \exists i \in \{1, \dots, n\}, |X_i - x| < \eta.$$

Hence, if $x \notin \mathbf{X}_n$, $\lim_{\eta \rightarrow 0} \frac{f_{\eta, \mathbf{X}_n}(x)}{\eta} = 0$. Besides, $\frac{f_{\eta, \mathbf{X}_n}(X_i)}{\eta} = 1$ for any η and i . Finally, since $\frac{f_{\eta, \mathbf{X}_n}(X_i)}{\eta} \leq 1$, dominated convergence theorem gives (A.2).

Let $\epsilon > 0$, and take $\mathbf{X}_n, \mathbf{Y}_n \in [0, 1]^n$ such that $\|\mathbf{X}_n - \mathbf{Y}_n\|_\infty \leq \eta\epsilon/2$. Take $P, Q \in \mathcal{P}$ such that $d_{BL}(P, Q) \leq \eta\epsilon/2$. Now, for any $x \in [0, 1]$,

$$|f_{\eta, \mathbf{X}_n}(x) - f_{\eta, \mathbf{Y}_n}(x)| \leq \eta\epsilon/2.$$

Indeed, suppose for instance $f_{\eta, \mathbf{X}_n}(x) \geq f_{\eta, \mathbf{Y}_n}(x)$. Suppose $f_{\eta, \mathbf{X}_n}(x) > 0$. Now consider X_i the closest point to x . We have that

$$f_{\eta, \mathbf{X}_n}(x) = \eta - |X_i - x| \leq \eta - |Y_i - x| + \eta\epsilon/2 = f_{\eta, \mathbf{Y}_n}(x) + \eta\epsilon/2.$$

Finally, let us compute the distance between the two images,

$$\begin{aligned} \eta |O_{\eta, n}(P, \mathbf{X}_n) - O_{\eta, n}(Q, \mathbf{Y}_n)| &= \left| \int f_{\eta, \mathbf{X}_n} dP - \int f_{\eta, \mathbf{Y}_n} dQ \right| \\ &\leq \left| \int f_{\eta, \mathbf{X}_n} dP - \int f_{\eta, \mathbf{X}_n} dQ \right| + \left| \int f_{\eta, \mathbf{X}_n} dQ - \int f_{\eta, \mathbf{Y}_n} dQ \right| \\ &\leq \left| \int f_{\eta, \mathbf{X}_n} dP - \int f_{\eta, \mathbf{X}_n} dQ \right| + \|f_{\eta, \mathbf{X}_n} - f_{\eta, \mathbf{Y}_n}\|_\infty \leq \eta\epsilon \end{aligned}$$

which gives

$$|O_{\eta,n}(P, \mathbf{X}_n) - O_{\eta,n}(Q, \mathbf{Y}_n)| \leq \epsilon.$$

Therefore $O_{\eta,n}$ is continuous and hence measurable. And finally M_n is measurable as limit and sum of measurable functions. $\square$

Proposition A.2. *Let $\mathbf{X}_n = (X_1, \dots, X_n)$ be a sample such that $X_i|P \stackrel{iid}{\sim} P$ for all $1 \leq i \leq n$. Then,*

1. *if $P \sim DP(\gamma)$, where $\gamma(d\theta) = \mathbb{1}(0 < d\theta < 1)$, then $M_n(\mathbf{X}_n, P)|\mathbf{X}_n \sim \text{Beta}(1, n)$*
2. *if $P \sim SP_\alpha$, then $M_n(\mathbf{X}_n, P)|\mathbf{X}_n \sim \text{Beta}(\alpha K_n(\mathbf{X}_n), n - \alpha K_n(\mathbf{X}_n))$*

Proof. We are going to derive the posterior distribution of $M_n(\mathbf{X}_n, P)$ when P is distributed according to the law of a two-parameter Poisson-Dirichlet process (Pitman and Yor [34]), $P \sim \text{PY}(\eta, \alpha)$, with $\alpha < 1$ and $\eta > -\alpha$. Point i) in the statement is the particular case $\text{PY}(1, 0)$, while point ii) corresponds to $\text{PY}(0, \alpha)$.

From Corollary 20 of Pitman [32], the posterior distribution of P given $\mathbf{X}_n$ under the two-parameter Poisson-Dirichlet process satisfies the following distributional equality,

$$P|\mathbf{X}_n \stackrel{d}{=} \sum_{i=1}^{K_n} w_i \delta_{X_i^*} + w_0 \tilde{P} \quad (\text{A.3})$$

where $(X_1^*, \dots, X_{K_n}^*)$ are K_n the distinct values in the sample $\mathbf{X}_n$ and having multiplicities $(n_1, \dots, n_{K_n})$, $w = (w_0, w_1, \dots, w_{K_n})$ is a random vector distributed according to a Dirichlet distribution $\text{Dir}(\eta + K_n\alpha, n_1 - \alpha, \dots, n_{K_n} - \alpha)$ and $\tilde{P} \sim \text{PY}(\alpha, \eta + K_n\alpha)$ independent of w . Therefore,

$$M_n(\mathbf{X}_n, P) = P(\{\mathbf{X}_n\}^c)|\mathbf{X}_n \stackrel{d}{=} \sum_{i=1}^{K_n} w_i \delta_{X_i^*}(\{\mathbf{X}_n\}^c) + w_0 \tilde{P}(\{\mathbf{X}_n\}^c) \quad (\text{A.4})$$

The point masses in (A.4) are all equal to zero, while $\tilde{P}(\{\mathbf{X}_n\}^c) = 1$ since the base measure of $\tilde{P}$ is diffuse. Therefore, $M_n(\mathbf{X}_n, P) \stackrel{d}{=} w_0$ and w_0 is distributed according to $\text{Beta}(\eta + K_n\alpha, n - \alpha K_n)$ from the aggregation property of the Dirichlet distribution. $\square$

Lemma A.2. *Let $f_{a,b}$ denote the density of the $\text{Beta}(a, b)$ distribution. Then, there exists $n_0 \in \mathbb{N}$ such that for all $b > a > n_0$, we have*

$$\sup_{x \in [0,1]} f_{a,b}(x) < 8(a+b)^{3/2} a^{-1/2} b^{-1/2} \quad (\text{A.5})$$

Proof. Take first $a, b > 2$. The mode of the beta distribution is $\frac{a-1}{a+b-2}$, therefore

$$\sup_{x \in [0,1]} f_{a,b}(x) = f_{a,b}\left(\frac{a-1}{a+b-2}\right) \leq \frac{1}{B(a,b)} \frac{a^{a-1} b^{b-1}}{(a+b-2)^{a+b-2}} \quad (\text{A.6})$$

From the Stirling's formula, there exists n_0 such that for $a, b > n_0$,

$$\frac{1}{\mathcal{B}(a, b)} \leq \frac{(a+b)^{a+b-1/2}}{a^{a-1/2}b^{b-1/2}}.$$

Plugging this quantity into (A.6), we have

$$\sup_{x \in [0,1]} f_{a,b}(x) \leq a^{-1/2}b^{-1/2}(a+b)^{3/2} \frac{1}{(1 - \frac{2}{a+b})^{a+b}}.$$

Finally, since for n_0 large enough $\frac{1}{(1 - \frac{2}{a+b})^{a+b}} \leq 8$, we have (A.5). $\square$

Lemma A.3. *There exists a constant C such that for any $a, b > 3$ such that $a < b/2$,*

$$\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx \geq \frac{C}{\sqrt{a}} \quad (\text{A.7})$$

Proof. In the following, we use the generic notation C to refer to universal constants. The value of C can change from a line to another. To simplify the notations, let $n = a + b$ and denote $I_x(a, b)$ the normalized incomplete Beta function defined as

$$I_x(a, b) = \int_0^x f_{a,b}(x) dx.$$

It is well-known that

$$I_x(a+1, b) = I_x(a, b) - \frac{x^a(1-x)^b}{a\mathcal{B}(a, b)},$$

(such a result can be obtain using integration by part). Successively applying this result for $x = m_{a-1,b}$, and using the definition of the median we can deduce that

$$\begin{aligned} I_{m_{a-1,b}}(a, b) &= I_{m_{a-1,b}}(a-1, b) - \frac{m_{a-1,b}^{a-1}(1-m_{a-1,b})^b}{(a-1)\mathcal{B}(a-1, b)} \\ &= 1/2 - \frac{m_{a-1,b}^{a-1}(1-m_{a-1,b})^b}{(a-1)\mathcal{B}(a-1, b)} \\ &= I_{m_{a,b}}(a, b) - \frac{m_{a-1,b}^{a-1}(1-m_{a-1,b})^b}{(a-1)\mathcal{B}(a-1, b)}, \end{aligned}$$

which in turn leads to

$$\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x) dx = I_{m_{a,b}}(a, b) - I_{m_{a-1,b}}(a, b) = \frac{m_{a-1,b}^{a-1}(1-m_{a-1,b})^b}{(a-1)\mathcal{B}(a-1, b)}. \quad (\text{A.8})$$

From the Stirling formula applied to the Beta function, we know that for all $a, b > 1$,

$$\mathcal{B}(a, b) \leq C \frac{a^{a-1/2}b^{b-1/2}}{(a+b)^{a+b-1/2}} = C \frac{a^{a-1/2}b^{b-1/2}}{n^{n-1/2}},$$

which leads to

$$\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx \geq C \frac{(n-1)^{n-3/2}}{(a-1)^{a-1/2}b^{b-1/2}} m_{a-1,b}^{a-1} (1 - m_{a-1,b})^b \quad (\text{A.9})$$

when plugging in (A.8).

Now, since $a - 1 < b$, the mode-median-mean inequality (see Payton et al. [31]) gives that

$$\frac{a-2}{n-3} \leq m_{a-1,b} \leq \frac{a-1}{n-1},$$

from which we deduce

$$m_{a-1,b}^{a-1} \geq \frac{(a-2)^{a-1}}{(n-3)^{a-1}} \geq \frac{(a-2)^{a-1}}{(n-1)^{a-1}}$$

and

$$(1 - m_{a-1,b})^b \geq \frac{b^b}{(n-1)^b}.$$

Now, together with (A.9), the previous two inequalities yield

$$\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx \geq C \frac{(n-1)^{n-3/2}}{(a-1)^{a-1/2}b^{b-1/2}} \frac{(a-2)^{a-1}b^b}{(n-1)^{n-1}} \geq C \sqrt{\frac{b}{an}}$$

where we used the fact that for $x > 2$, $x^x \geq (x-1)^x \geq Cx^x$. We then conclude by noticing that $a < b/2$ implies that $b/n < 2/3$. $\square$

References

- [1] Beirlant, J. and Devroye, L. (1999). On the impossibility of estimating densities in the extreme tail. *Statist. Probab. Lett.* **43**, 57–64.
- [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.
- [3] Ben-Hamou, A., Boucheron, S. and Gassiat, E. (2018). Pattern coding meets censoring: (almost) adaptive coding on countable alphabets. *Preprint: arXiv:1608.08367*
- [4] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation*. Cambridge University Press.
- [5] Bubeck, S., Ernst, D. and Garivier, A. (2013). Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality. *J. Mach. Learn. Res.* **14**, 601–623.
- [6] Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *Ann. Appl. Probab.* **28**, 1099–1135.
- [7] Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). Power-law Distributions in Empirical Data. *SIAM Rev.* **51**, 661–703.

- [8] Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.* **1**, 209–230.
- [9] Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58.
- [10] Gao, F. (2013). Moderate deviations for a nonparametric estimator of sample coverage. *Ann. Statist.* **41**, 641–669.
- [11] Gao, Z., Tseng, C.H., Pei, Z. and Blaser, M.J. (2007). Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl. Acad. Sci. USA* **104**, 2927–2932.
- [12] Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.
- [13] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problems with infinitely many boxes: general asymptotics and power law. *Probab. Surv.* **4**, 146–171.
- [14] Goldwater, S., Griffiths, T. and Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In *Proceedings of the Conference in Advances in Neural Information Processing Systems*.
- [15] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264 (1953)
- [16] Grabchak, M. and Zhang, Z. (2017). Asymptotic properties of Turing’s formula in relative error. *Mach. Learn.* **106**, 1771–1785.
- [17] Ionita-Laza, I., Lange, C. and Laird, N.M. (2009). Estimating the number of unseen variants in the human genome. *Proc. Natl. Acad. Sci. USA* **106**, 5008–5013.
- [18] Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17**, 373–401.
- [19] Kingman, J.F.K. (1975). Random Discrete Distributions. *J. Roy. Statist. Soc. Ser. B* **37**, 1–22.
- [20] Kroes, I., Lepp, P.W. and Relman, D.A. (1999). Bacterial diversity within the human subgingival crevice. *Proc. Natl. Acad. Sci. USA* **96**, 14547–14552.
- [21] McAllester, D. and Ortiz, L. (2003). Concentration inequalities for the missing mass and for histogram rule error. *J. Mach. Learn. Res.* **4**, 895–911.
- [22] McAllester, D. and Schapire, R.E. (2000). On the convergence rate of Good-Turing estimators. In *Proceedings of the Conference on Computational Learning Theory*.
- [23] Mitzenmacher, M. (2004) A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* **1**, 226–251.
- [24] Mossel, E. and Ohannessian, M.I. (2019) On the impossibility of learning the missing mass. *Entropy* **21**, 28.
- [25] Motwani, S. and Vassilvitskii, S. (2006) Distinct value estimators in power law distributions. In *Proceedings of the Workshop on Analytic Algorithms and Combinatorics*.
- [26] Newman, M.E.J. (2003). The Structure and Function of Complex Networks. *SIAM Rev.* **45**, 167–256.

- [27] Ohannessian, M.I. and Dahleh, M.A. (2012). Rare probability estimation under regularly varying heavy tails. *J. Mach. Learn. Res.* **23**, 1–24.
- [28] Orlitsky, A., Santhanam, N.P. and Zhang, J. (2003). Always Good-Turing: asymptotically optimal probability estimation. *Science* **302**, 427–431.
- [29] Orlitsky, A., Santhanam, N.P. and Zhang, J. (2004). Universal compression of memoryless sources over unknown alphabets. *IEEE Trans. Inf. Theory* **50**, 1469–1481.
- [30] Orlitsky, A., Suresh, A.T. and Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113**, 13283–13288.
- [31] Payton, M.E., Young, L.J. and Young, J.H. (1989). Bounds for the difference between median and mean of beta and negative binomial distributions. *Metrika* **36**, 347–354.
- [32] Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory*, Institute of Mathematical Statistics.
- [33] Pitman, J. (2006) Combinatorial Stochastic Processes. *Ecole d’Eté de Probabilités de Saint-Flour XXXII*, Lecture notes in mathematics, Springer.
- [34] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.
- [35] Rajaraman, N., Thangaraj, A. and Suresh, A.T. (2017). Minimax Risk for Missing Mass Estimation. In *Proceedings of the IEEE International Symposium on Information Theory*.
- [36] Sornette, D. (2006). *Critical Phenomena in Natural Sciences*. Springer.
- [37] Wagner, B., Viswanath and P., Kulkarni, S.R. (2017). Strong consistency of the Good-Turing estimator. In *Proceedings of the IEEE International Symposium on Information Theory*.
- [38] Zhang, C.H. and Zhang, Z. (2009). Asymptotic normality of a nonparametric estimator of sample coverage. *Ann. Statist.* **37**, 2582–2595.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	On consistent and rate optimal estimation of the missing mass
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Submitted to the journal Annales de l'Institut Henri Poincarés, Series B and is under review

Student Confirmation

Student Name:	Fadhel Ayed		
Contribution to the Paper	I am the main author of this paper.		
Signature		Date	12/07/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor F. Caron		
Supervisor comments			
Signature		Date	14/07/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

6

A Good-Turing estimator for feature allocation models

A Good-Turing estimator for feature allocation models

Fadhel Ayed¹, Marco Battiston²,
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).

MSC 2010 subject classifications: 62G05, 62C20.

Keywords and phrases: Bernoulli product model, feature allocation model, Good-Turing estimator, minimax rate optimality, missing mass, non-asymptotic uncertainty quantification, nonparametric empirical Bayes, SNP data.

Received December 2018.

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 $(F_j)_{j \geq 1}$. 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 $Y_i = (Y_{i,j})_{j \geq 1}$ of independent Bernoulli random variables with unknown success probabilities $(p_j)_{j \geq 1}$, 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 $\mathbf{Y}_n = (Y_1, \dots, Y_n)$ denote a collection of n random samples collected under the Bernoulli product model with unknown feature probabilities $(p_j)_{j \geq 1}$. 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.

$$\begin{aligned} M_n(\mathbf{Y}_n, (p_j)_{j \geq 1}) &= \mathbb{E} \left(\sum_{j \geq 1} \mathbb{I}\{X_{n,j} = 0, Y_{n+1,j} = 1\} \mid \mathbf{Y}_n \right) \\ &= \sum_{j \geq 1} p_j \mathbb{I}\{X_{n,j} = 0\}, \end{aligned} \quad (1.1)$$

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 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 $\mathbf{Y}_n = (Y_1, \dots, Y_n)$ 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 $i = 1, \dots, n$, 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 $a_n/b_n \rightarrow 1$ as $n \rightarrow +\infty$. 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) ,

$$\begin{aligned} \mathbb{E}(M_n) &= \sum_{j \geq 1} p_j \mathbb{P}(X_{n,j} = 0) = \sum_{j \geq 1} p_j (1 - p_j)^n \\ &= \frac{1}{\binom{n+1}{1}} \sum_{j \geq 1} \binom{n+1}{1} p_j (1 - p_j)^{(n+1)-1} = \frac{1}{n+1} \mathbb{E}(K_{n+1,1}) \simeq \frac{1}{n} \mathbb{E}(K_{n,1}) \end{aligned}$$

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., $(Y_1, \dots, Y_n)$ 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 $K_{n,1} \in \{0, 1, \dots, n\}$. 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., $M_{n-1} = \mathbb{E}[K_n - K_{n-1}(j) | \{Y_1, \dots, Y_n\} \setminus Y_j]$. 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 $\nu(ds, df) = \rho(s)ds\mathbb{I}(0, 1)(f)df$ 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

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

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 $\mathcal{P}_W = \{(p_j)_{j \geq 1} : p_j \in (0, 1) \text{ and } \sum_{j \geq 1} p_j \leq W\}$ 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 $\mathcal{P}_W$, the L_2 -risk of the estimator $\hat{T}_n$ is defined as follows

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

and the minimax risk is

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

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 $\mathcal{P}_W$. 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 $\mathcal{P}_W$. Then,

i) for any $n \geq 1$

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

ii) for any $n \geq 2$

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

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 $a_n \leq C b_n$ 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 $\sup_{(p_j)_{j \geq 1} \in \mathcal{P}_W} R_n(\hat{M}_n; (p_j)_{j \geq 1}) \lesssim W/n$, 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 $R_n^* \gtrsim W/n$. 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 $\mathcal{P}_W$ of admissible probabilities' masses up to a constant factor.

In Theorem 2.1 we considered the class $\mathcal{P}_W$ of feature probabilities having total mass bounded by W . We highlight the crucial role of 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 $(p_j^{(0)})_{j \geq 1}, \dots, (p_j^{(W-1)})_{j \geq 1} \in \mathcal{P}_1$, i.e. such that $\sum_j p_j^{(r)} \leq 1$ for all $r \in \{0, \dots, W-1\}$. Now, let us construct another vector of frequencies $P = (p_j)_{j \geq 1} \in \mathcal{P}_W$ 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 $(p_j^{(0)})_{j \geq 1}, \dots, (p_j^{(W-1)})_{j \geq 1}$, such that in the first W entries of P we put the first elements $(p_1^{(0)}, \dots, p_1^{(W-1)})$, followed by the second ones $(p_2^{(0)}, \dots, p_2^{(W-1)})$, the third ones $(p_3^{(0)}, \dots, p_3^{(W-1)})$ 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 $Y_i^{(r)} = (Y_{i,j}^{(r)})_{j \geq 1}$, $r \in \{0, \dots, W-1\}$, 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 $\mathbf{Y}^{(r)} = (Y_1^{(r)}, \dots, Y_n^{(r)})$. 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 $\delta \in (0, 1)$

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

and

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

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

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

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

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

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 $W \rightarrow 0$. 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 $(\tilde{Y}_1, \dots, \tilde{Y}_{\tilde{n}})$ 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 $\mathbb{E}_{\tilde{n}}(\frac{1}{\tilde{n}} \mathbb{I}\{\tilde{n} > 0\})$. 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

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

However, we also need to estimate $W = \sum_j p_j$ in order to make use of $\tilde{M}_{\tilde{n}}^{\text{GT}}$ 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 $\tilde{M}_{\tilde{n}}^{\text{GT}}$ for $\tilde{M}_{\tilde{n}}$, we obtain the following estimator of M_n ,

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

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 is 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 larger 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 presented 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 $\mathbf{Y}_n = (Y_1, \dots, Y_n)$, 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 $\nu(ds, df) = \rho(s)ds\mathbb{I}(0,1)(f)df$, 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 $F_1^*, \dots, F_{K_n}^*$ be the K_n distinct features observed in it, the posterior estimate of M_n , under a square loss function, equals

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

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 $\nu_n(ds, df) = (1-s)^n \rho(s)ds\mathbb{I}(0,1)(f)df$. Hence the Bayesian estimator in (A.1) boils down to

$$\begin{aligned} \tilde{M}_n(\vartheta, \alpha, \beta) &= \mathbb{E} \left[\tilde{\mu}_n(\{F_1^*, \dots, F_{K_n}^*\}^c) + \sum_{\ell=1}^{K_n} J_\ell \delta_{F_\ell^*}(\{F_1^*, \dots, F_{K_n}^*\}^c) \right] \\ &= \mathbb{E}[\tilde{\mu}_n((0, 1))] = -\frac{d}{du} \mathbb{E}[e^{-u\tilde{\mu}_n((0,1))}] \Big|_{u=0} = \vartheta \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\beta + n + 1)} \end{aligned}$$

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

$$\bar{\rho}(s) := \int_s^1 \rho(s)ds = \frac{\vartheta}{\Gamma(1-\alpha)} \int_s^1 s^{-1-\alpha}(1-s)^{\beta+\alpha-1}ds$$

fulfills

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

Hence we just proved that $\bar{\rho}(x) \simeq x^{-\alpha}\vartheta/(\Gamma(1-\alpha)\alpha)$ as $n \rightarrow +\infty$, and therefore an application of [12, Corollary 21] gives the asymptotic relation $\frac{K_{n,1}}{n^\alpha} \rightarrow \vartheta$. Moreover, from Stirling's approximation, $\frac{\Gamma(\alpha+\beta+n)}{\Gamma(n+1+\beta)} \simeq n^{\alpha-1}$. Therefore, $\hat{\vartheta}_n = \frac{K_{n,1}\Gamma(\beta+n+1)}{\Gamma(\beta+\alpha+n)n} \rightarrow \vartheta$ 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 $\tilde{M}_n(\hat{\vartheta}_n, \alpha, \beta) = \frac{K_{n,1}}{n} = \hat{M}_n$. $\square$

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 $\mathcal{P}_W$.

In order to bound the L_2 -risk of the estimator $\hat{M}_n$ from above for any $(p_j)_{j \geq 1} \in \mathcal{P}_W$, 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

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

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

$$\begin{aligned}\mathrm{Var}(\hat{M}_n - M_n) &\leq \sum_{j \geq 1} \mathbb{E} \left(\frac{1}{n} \mathbb{I}\{X_{n,j} = 1\} - p_j \mathbb{I}\{X_{n,j} = 0\} \right)^2 \\ &= \sum_{j \geq 1} \left(\left(\frac{1}{n} \right)^2 \mathbb{P}(X_{n,j} = 1) + p_j^2 \mathbb{P}(X_{n,j} = 0) \right) \\ &= \left(\frac{1}{n} \right)^2 \sum_{j \geq 1} \mathbb{P}(X_{n,j} = 1) + \sum_{j \geq 1} p_j^2 \mathbb{P}(X_{n,j} = 0) \\ &= \frac{1}{n^2} \sum_{j \geq 1} \binom{n}{1} p_j (1 - p_j)^{n-1} + \sum_{j \geq 1} p_j^2 (1 - p_j)^n \\ &= \frac{1}{n} \sum_{j \geq 1} p_j (1 - p_j)^{n-1} + \frac{1}{\binom{n+1}{1}} \sum_{j \geq 1} \binom{n+1}{1} p_j^2 (1 - p_j)^n \\ &= \frac{1}{n} \mathbb{E}(M_{n-1}) + \frac{1}{n+1} \mathbb{E}(M_{n+1,1}) \leq \frac{2n+1}{n(n+1)} W.\end{aligned}$$

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 $\mathcal{P}_W$, 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 $\inf_{\hat{T}_n \leq W} \sup_{(p_j)_{j \geq 1} \in \mathcal{P}_W} \mathbb{E}[(M_n - \hat{T}_n)^2]$. 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 $\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_m, 0, 0, \dots)$ where m is a Poisson random variable with mean $\lambda = nW$ and $\mathbf{p}_j$ are independent beta random variables with parameters $a = 1$ and $b = 2n - 1$. We denote by $\mathbf{P}_F := \sum_{j=1}^m \mathbf{p}_j$ the total mass of the random sequence $(\mathbf{p}_j)_{j \geq 1}$; we further define $X := (X_{n,j})_{j \geq 1}$ and it is worth noticing that, conditionally on $\mathbf{P}$, the $X_{n,j}$'s are independent, having a Binomial distribution with parameters $(n, \mathbf{p}_j)$. 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

$$\begin{aligned}
& \inf_{\hat{T}_n \leq W} \sup_{(p_j)_j \in \mathcal{P}_W} \mathbb{E}[(M_n - \hat{T}_n)^2] \\
& \geq \inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathbf{P}} \left\{ \mathbb{I}\{\mathbf{P} \in \mathcal{P}_W\} \mathbb{E}_{X|\mathbf{P}}[(M_n - \hat{T}_n)^2 | \mathbf{P}] \right\} \\
& \geq \inf_{\hat{T}_n \leq W} \left\{ \mathbb{E}_{\mathbf{P}} \mathbb{E}_{X|\mathbf{P}}[(M_n - \hat{T}_n)^2 | \mathbf{P}] \right. \\
& \quad \left. - \mathbb{E}_{\mathbf{P}}(1 - \mathbb{I}\{\mathbf{P} \in \mathcal{P}_W\}) \mathbb{E}_{X|\mathbf{P}}[(M_n - \hat{T}_n)^2 | \mathbf{P}] \right\} \\
& \geq \inf_{\hat{T}_n \leq W} \left\{ \mathbb{E}_{\mathbf{P}} \mathbb{E}_{X|\mathbf{P}}[(M_n - \hat{T}_n)^2 | \mathbf{P}] - \mathbb{E}_{\mathbf{P}}(1 - \mathbb{I}\{\mathbf{P} \in \mathcal{P}_W\}) \mathbf{P}_F^2 \right\} \\
& \geq \inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathbf{P}} \mathbb{E}_{X|\mathbf{P}}[(M_n - \hat{T}_n)^2 | \mathbf{P}] - \mathbb{E}(\mathbf{P}_F^2 \mathbb{I}\{\mathbf{P}_F > W\}) \quad (\text{A.2})
\end{aligned}$$

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

$$\begin{aligned}
& \mathbb{E}(\mathbf{P}_F^2 \mathbb{I}\{\mathbf{P}_F > W\}) \\
& = \int s^2 \mathbb{I}\{s > W\} \mathbb{P}_{\mathbf{P}_F}(ds) = 2 \int \int t \mathbb{I}\{t < s\} dt \mathbb{I}\{s > W\} \mathbb{P}_{\mathbf{P}_F}(ds) \\
& = 2 \int t \int \mathbb{I}\{t < s\} \mathbb{I}\{s > W\} \mathbb{P}_{\mathbf{P}_F}(ds) dt = 2 \int t \mathbb{P}(\mathbf{P}_F > \max(t, W)) dt
\end{aligned}$$

which leads to the more convenient form

$$\mathbb{E}(\mathbf{P}_F^2 \mathbb{I}\{\mathbf{P}_F > W\}) = W^2 \mathbb{P}(\mathbf{P}_F > W) + 2 \int_W^\infty t \mathbb{P}(\mathbf{P}_F > t) dt. \quad (\text{A.3})$$

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

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

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

$$\begin{aligned}
& \mathbb{E}(\mathbf{P}_F - \mathbb{E} \mathbf{P}_F)^3 = \mu_3(\mathbf{P}_F) \\
& = \mathbb{E}[\mu_3(\mathbf{P}_F | m)] + \mu_3(\mathbb{E}[\mathbf{P}_F | m]) + 3 \text{Cov}(\mathbb{E}[\mathbf{P}_F | m], \text{Var}[\mathbf{P}_F | m]) \\
& = \mathbb{E}[\mu_3(\mathbf{P}_F | m)] + \mu_3\left(\frac{m}{2n}\right) + 3 \text{Cov}\left(\frac{m}{2n}, \frac{mab}{(a+b)^2(a+b+1)}\right) \\
& = \mathbb{E}[\mu_3(\mathbf{P}_F | m)] + \frac{1}{8n^3} \mu_3(m) + 3 \frac{2n-1}{8n^3(2n+1)} \text{Var}(m)
\end{aligned}$$

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 $\mathbb{E}[\mu_3(\mathbf{P}_F|m)] + \frac{W}{2n^2}$. Conditioning on m , the total mass $\mathbf{P}_F$ is a sum of independent and identically distributed random variables so $\mu_3(\mathbf{P}_F|m) = m\mu_3(p_1)$. Since p_1 is Beta distributed with parameters a , and b , its third centralized moment is given by

$$\begin{aligned}\mu_3(p_1) &= [\text{Var}(p_1)]^{3/2} \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}} = \frac{2ab(b-a)}{(a+b)^3(a+b+1)(a+b+2)} \\ &= \frac{4(2n-1)(n-1)}{(2n)^3(2n+1)2(n+1)} = \frac{(2n-1)(n-1)}{4n^3(2n+1)(n+1)}.\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(\mathbf{P}_F > t) &= \mathbb{P}(\mathbf{P}_F - \mathbb{E}\mathbf{P}_F > t - W/2) = \mathbb{P}((\mathbf{P}_F - \mathbb{E}\mathbf{P}_F)^3 > (t - W/2)^3) \\ &\leq \frac{\mu_3(\mathbf{P}_F)}{(t - W/2)^3} \leq \frac{W}{n^2(t - W/2)^3}\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}(\mathbf{P}_F^2 \mathbb{I}\{\mathbf{P}_F > W\}) &= W^2 \mathbb{P}(\mathbf{P}_F > W) + 2 \int_W^\infty t \mathbb{P}(\mathbf{P}_F > t) dt \\ &\leq W^2 \frac{8W}{n^2 W^3} + \frac{2W}{n^2} \int_W^\infty \frac{t}{(t - W/2)^3} dt \\ &= \frac{8}{n^2} + \frac{2W}{n^2} \left[\int_W^\infty \frac{1}{(t - W/2)^2} dt + \int_W^\infty \frac{W/2}{(t - W/2)^3} dt \right] \\ &= \frac{8}{n^2} + \frac{2W}{n^2} \left[\frac{2}{W} + \frac{1}{W} \right] = \frac{14}{n^2}\end{aligned}$$

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

$$\begin{aligned}&\inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathbf{P}} \mathbb{E}_{X|\mathbf{P}}[(M_n - \hat{T}_n)^2 | \mathbf{P}] \\ &= \inf_{\hat{T}_n \leq W} \mathbb{E}_m \mathbb{E}_{\mathbf{P}|m} \left\{ \mathbb{E}_{X|\mathbf{P}}[(M_n - \hat{T}_n)^2 | \mathbf{P}] \mid m \right\} \\ &\geq \mathbb{E}_m \inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathbf{P}|m} \left\{ \mathbb{E}_{X|\mathbf{P}}[(M_n - \hat{T}_n)^2 | \mathbf{P}] \mid m \right\}.\end{aligned}\tag{A.4}$$

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 $\mathbf{P}_F$.

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

$$\begin{aligned}
&= \inf_{\hat{T}_n \leq W} \mathbb{E}_{X|m} \mathbb{E}_{\mathbf{P}|X,m} [(M_n - \hat{T}_n)^2 | X, m] \\
&\geq \mathbb{E}_{X|m} \inf_{\hat{T}_n} \mathbb{E}_{\mathbf{P}|X,m} [(M_n - \hat{T}_n)^2 | X, m] \\
&\geq \mathbb{E}_{X|m} \text{Var}_{\mathbf{P}|X,m}(M_n | X, m) = \mathbb{E}_{X|m} \text{Var}_{\mathbf{P}|X,m} \left(\sum_{j=1}^m p_j \mathbb{I}\{X_{n,j} = 0\} \middle| X, m \right) \\
&= \mathbb{E}_{X|m} \left[\sum_{j=1}^m \mathbb{I}\{X_{n,j} = 0\} \frac{a(n+b)}{(a+n+b)^2(a+n+b+1)} \right]
\end{aligned}$$

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

$$\begin{aligned}
&\inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathbf{P}|m} \mathbb{E}_{X|\mathbf{P}} [(M_n - \hat{T}_n)^2 | \mathbf{P}] \\
&\geq \sum_{j=1}^m \mathbb{P}(X_{n,j} = 0) \frac{a(n+b)}{(a+n+b)^2(a+n+b+1)} \\
&= \sum_{j=1}^m \mathbb{E}_{\mathbf{p}_j} [(1 - \mathbf{p}_j)^n] \frac{a(n+b)}{(a+n+b)^2(a+n+b+1)} \\
&= \sum_{j=1}^m \frac{\Gamma(b+n)\Gamma(a+b)}{\Gamma(b)\Gamma(a+b+n)} \frac{a(n+b)}{(a+n+b)^2(a+n+b+1)} \quad (\text{A.5})
\end{aligned}$$

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

$$\begin{aligned}
\inf_{\hat{T}_n \leq W} \mathbb{E}_{\mathbf{P}|m} \mathbb{E}_{X|\mathbf{P}} [(M_n - \hat{T}_n)^2 | \mathbf{P}] &\geq m \frac{\Gamma(b+n)\Gamma(1+b)}{\Gamma(b)\Gamma(1+b+n)} \frac{n+b}{(n+b+1)^2(n+b+2)} \\
&= \frac{mb}{(n+b+1)^2(n+b+2)} \\
&= \frac{2mn}{9n^2(3n+1)} = \frac{2m}{9n(3n+1)}
\end{aligned}$$

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

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

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

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

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 determination of (2.5) and (2.6)

The concentration inequalities (2.5)–(2.6) can be proved by 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

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

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 $0 < \lambda < n$*

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
\log \mathbb{E} \left(e^{\lambda(\hat{W}_n - W)} \right) &\leq n \sum_{k \geq 2} \frac{1}{k!} \left(\frac{\lambda}{n} \right)^k \mathbb{E} \hat{W}_n \leq n \sum_{k \geq 2} \left(\frac{\lambda}{n} \right)^k \mathbb{E} \hat{W}_n \\
&= n \cdot \frac{\lambda^2}{n^2} \cdot \frac{1}{1 - \lambda/n} \mathbb{E} \hat{W}_n = \frac{w_n^+ \lambda^2}{2(1 - \lambda/n)}
\end{aligned}$$

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 $\phi(-\lambda/n) \leq \lambda^2/(2n^2)$. $\square$

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

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

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:

$$\begin{aligned} \log \mathbb{E}(e^{t(K_{n,r} - \mathbb{E}(K_{n,r}))}) &= \sum_{j=1}^{\infty} \log \mathbb{E} \exp \{(\mathbb{I}\{X_{n,j} = r\} - \mathbb{E}\mathbb{I}\{X_{n,j} = r\})\} \\ &\leq \phi(|t|) \sum_{j=1}^{\infty} \text{Var}(\mathbb{I}\{X_{n,j} = r\}) = \phi(|t|) \text{Var}(K_{n,r}) \\ &\stackrel{(\text{A.12})}{\leq} \phi(|t|) \mathbb{E}(K_{n,r}) \end{aligned}$$

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 $\text{Var}(K_n) \leq \mathbb{E}(K_n)$. $\square$

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 $1 - \delta$*

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

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 by 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*

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

and

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

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

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

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

$$\begin{aligned}\text{Var}(\hat{M}_n - M_n) &\leq \sum_{j \geq 1} \mathbb{E} \left(\frac{1}{n} \mathbb{I}\{X_{n,j} = 1\} - p_j \mathbb{I}\{X_{n,j} = 0\} \right)^2 \\ &= \sum_{j \geq 1} \left(\left(\frac{1}{n} \right)^2 \mathbb{P}(X_{n,j} = 1) + p_j^2 \mathbb{P}(X_{n,j} = 0) \right) \\ &= \frac{1}{n^2} \sum_{j \geq 1} \mathbb{P}(X_{n,j} = 1) + \sum_{j \geq 1} p_j^2 \mathbb{P}(X_{n,j} = 0) \\ &= \frac{1}{n^2} \mathbb{E}(K_{n,1}) + \sum_{j \geq 1} p_j^2 (1 - p_j)^n \leq \frac{1}{n^2} \mathbb{E}(K_{n,1}) + \sum_{j \geq 1} p_j^2 (1 - p_j)^{n-2} \\ &= \frac{1}{n^2} \mathbb{E}(K_{n,1}) + \frac{2}{n(n-1)} \mathbb{E}(K_{n,2})\end{aligned}$$

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

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 $0 < \lambda < n$*

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

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 $0 < \lambda < (n-1)$

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \log \mathbb{E} \left(e^{-\lambda[D_n - \mathbb{E}(D_n)]} \right) &= \sum_{j \geq 1} \{ \log[E(e^{\lambda p_j W_j})] - \lambda p_j \mathbb{E} W_j \} \\ &\leq \sum_{j \geq 1} \mathbb{E}[e^{\lambda p_j W_j} - 1 - \lambda p_j W_j] = \sum_{j \geq 1} \mathbb{E} \left[\frac{e^{\lambda p_j W_j} - 1 - \lambda p_j W_j}{(\lambda p_j W_j)^2} (\lambda p_j W_j)^2 \right] \\ &\leq \sum_{j \geq 1} \mathbb{E} \left[\frac{e^{\lambda p_j} - 1 - \lambda p_j}{(\lambda p_j)^2} (\lambda p_j W_j)^2 \right] = \sum_{j \geq 1} \mathbb{E} W_j^2 \phi(\lambda p_j) \\ &= \sum_{j \geq 1} \phi(\lambda p_j) \mathbb{E} \left(\mathbb{I}\{X_{n,j} = 0\} - \frac{1}{np_j} \mathbb{I}\{X_{n,j} = 1\} \right)^2 \\ &\leq \sum_{j \geq 1} \phi(\lambda p_j) \mathbb{E} \left(\mathbb{I}\{X_{n,j} = 0\} + \frac{1}{n^2 p_j^2} \mathbb{I}\{X_{n,j} = 1\} \right) \\ &= \sum_{j \geq 1} \phi(\lambda p_j) \left[(1 - p_j)^n + \frac{1}{np_j} (1 - p_j)^{n-1} \right] \\ &= \sum_{j \geq 1} \sum_{k \geq 2} \frac{(\lambda p_j)^k}{k!} \left[(1 - p_j)^n + \frac{1}{np_j} (1 - p_j)^{n-1} \right] \\ &\leq \sum_{k \geq 2} \lambda^k \sum_{j \geq 1} \left(\frac{p_j^k}{k!} e^{-np_j} + \frac{p_j^{k-1}}{nk!} e^{-(n-1)p_j} \right), \end{aligned}$$

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

$$\begin{aligned} \log \mathbb{E} \left(e^{-\lambda[D_n - \mathbb{E}(D_n)]} \right) &\leq \sum_{k \geq 2} \lambda^k \left\{ \frac{1}{n^k} \sum_{j \geq 1} \frac{(np_j)^k}{k!} e^{-np_j} + \frac{1}{kn(n-1)^{k-1}} \sum_{j \geq 1} \frac{[(n-1)p_j]^{k-1}}{(k-1)!} e^{-(n-1)p_j} \right\} \\ &= \sum_{k \geq 2} \lambda^k \left\{ \frac{1}{n^k} \mathbb{E}[K_k(n)] + \frac{1}{kn(n-1)^{k-1}} \mathbb{E}[K_{k-1}(n-1)] \right\} \\ &\leq \sum_{k \geq 2} \left(\frac{\lambda}{n-1} \right)^k \left\{ \mathbb{E}[K_k(n)] + \frac{1}{k} \mathbb{E}[K_{k-1}(n-1)] \right\} \end{aligned}$$

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

hence (A.16) has been now established. $\square$

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:

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

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

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

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

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 $1 - \delta$

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

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

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

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

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

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

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

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. $\square$

Acknowledgements

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.

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**.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	A Good-Turing estimator for feature allocation models
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Published in the Electronic Journal of Statistics

Student Confirmation

Student Name:	Fadhel Ayed
Contribution to the Paper	I was in charge of section 2.2 on the optimality of the estimator and section 2.3 on the connections between GT for species sampling and feature allocation. I also contributed to section 2.1 on the interpretation of the estimator.
Signature <i>Mohamed El Fadhel Ayed</i>	Date 12/07/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor F CARON
Supervisor comments	
Signature 	Date 14/07/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

7

Consistent estimation of rare masses in feature allocation models

Consistent estimation of rare 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:

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 (FS) 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 FS counterpart of the classical problem of estimating rare probabilities in the species sampling (SS) framework, where each individual is endowed with only one feature (or “species”). In this paper we study the problem of consistent estimation of the rare 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 rare 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 rare 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 rare probabilities, let consider the following species sampling (SS) 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 $(Z_1, \dots, Z_n)$ 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 rare probability of order $r \geq 0$ is defined as

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

where $\mathbb{1}_{\{\cdot\}}$ 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 $(Z_1, \dots, Z_n)$. 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 rare 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 (FS). 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 FS framework generalizes the SS framework by allowing each individual in the population to belong to more than one “species”, now called “feature”. Formally, the FS 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 $(Y_1, \dots, Y_n)$ of individuals from the population. In analogy to the SS framework, under the FS framework the missing mass is defined as the sum of the probabilities p_j ’s of unseen features in the observed sample $(Y_1, \dots, Y_n)$.

The Bernoulli product model is arguably the most popular model for estimating the missing mass under the FS framework. It assumes that the i th element of the random sample $(Y_1, \dots, Y_n)$ 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 $(Y_1, \dots, Y_n)$, is distributed according to a Binomial distribution with parameter (n, p_j) , for any $j \geq 1$. In analogy to the SS framework, given the random sample $(Y_1, \dots, Y_n)$ under the Bernoulli product model we define the rare mass of order $r \geq 0$ as

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

That is, $M_{n,r}(Y_1, \dots, Y_n; (p_j)_{j \geq 1})$ is the total mass of features observed with frequency $r \geq 0$ in the sample $(Y_1, \dots, Y_n)$. In particular, we refer to $M_{n,0}(Y_1, \dots, Y_n; (p_j)_{j \geq 1})$ 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 $(Y_1, \dots, Y_n)$ and $(p_j)_{j \geq 1}$ in $M_{n,r}(Y_1, \dots, Y_n; (p_j)_{j \geq 1})$, 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 FS 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 $n \rightarrow +\infty$. 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, it is critical in our strategy the use of a generalized Beta process prior (James, 2017), which allows us to prove non-consistency of $\hat{T}_{n,0}$ by exploiting the posterior distribution of $M_{n,0}$ given $(Y_1, \dots, Y_n)$. 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 FS 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 $\mathbb{R}^+$. 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 are 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 rare 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 rare 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 proof of Theorem 1 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 rare probabilities $P_{n,r}$. It assumes that the (observable) random sample $\mathbf{Z}_n = (Z_1, \dots, Z_n)$ 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

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

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

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

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

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 $q \in \mathcal{Q}$ is a geometric distribution with small enough parameter, then there exists $\varepsilon > 0$ such that

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

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 rare 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 $\mathcal{Q}$ 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 $\nu(dx) = \sum_{j \geq 1} \delta_{q_j}(dx)$ and let $\bar{\nu}(x) = \nu[x, 1]$. Then, following Karlin (1967), we say that $q \in \mathcal{Q}$ is a regularly varying heavy-tailed distribution, with regular variation index $\alpha \in (0, 1)$ if

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

where $\ell(\cdot)$ is a slowly varying function, namely $\ell(ct)/\ell(t) \rightarrow 1$ as $t \rightarrow +\infty$ for all $c > 0$. Ohannessian and Dahleh (2012) showed that if $q \in \mathcal{Q}$ 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 $\varepsilon \in (0, \varepsilon_r)$

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

Then, by combining (5) with the Borel–Cantelli lemma it follows that $\hat{P}_{n,r}/P_{n,r} \rightarrow 1$ almost surely, as $n \rightarrow +\infty$. 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 $q \in \mathcal{Q}$ such that

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

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 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

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

Recall that $X_{n,j}$ denotes the number of times that feature F_j has been observed in the sample $(Y_1, \dots, Y_n)$, 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 $\hat{T}_{n,0} : (\{0, 1\}^\infty)^n \rightarrow \mathbb{R}_+$ of the missing mass $M_{n,0}$ is a measurable map whose argument is the observed sample

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

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

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 $\mathcal{P}$. This means that for any estimator $\hat{T}_{n,0}$ of the missing mass, there exists at least one element $p \in \mathcal{P}$ for which $\hat{T}_{n,0}/M_{n,0}$ does not converge to 1 in probability, as $n \rightarrow +\infty$.

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 $\hat{T}_{n,0}$, it is possible to find an element $p \in \mathcal{P}$ such that for any $\varepsilon \in (0, 1/6)$*

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

for some strictly positive constant C_ε .

3.1 Proof of Theorem 1

The proof of this theorem will build on technical lemmas whose 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 $\varepsilon \in (0, 1/6)$,

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

and therefore there exists a $p \in \mathcal{P}$ 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 $\varepsilon \in (0, 1/6)$,

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

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

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

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

The main idea of the proof is as follows: we lower bound the supremum over $\mathcal{P}$ 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

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

where we have applied reverse Fatou's lemma to exchange the limsup 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 $p \in \mathcal{P}$ 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 $p(\cdot) = \sum_j (1 - e^{-s_j}) \delta_{F_j}(\cdot) \in \mathcal{P}$. We select $\tilde{N}$ to be a completely random measure with Lévy intensity $\nu(ds, dF) = e^{-s}/s ds \mathbb{1}_{(0,1)}(F) dF$. The distribution of $\tilde{N}$ is completely characterized by its Laplace functional defined as,

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

for any measurable function $f : [0, 1] \rightarrow \mathbb{R}^+$. 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 $\mathbf{Y}_n$. Denoted by $F_1^*, \dots, F_{k_n}^*$ the k_n distinct features out of $\mathbf{Y}_n$, the following distributional equality holds

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

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 $\nu_n(ds, dF) = e^{-s n} \nu(ds, dF)$, independent of $(J_\ell, F_\ell^*)_{\ell=1, \dots, k_n}$ 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

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

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 $x \rightarrow 0$, 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

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

which implies

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

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

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

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

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

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

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

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

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

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

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

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 FS

In this section we will establish exponential tail bounds for the rare 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. $\lambda \leq 0$)*

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

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

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

- 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 $\lambda \in \mathbb{R}$*

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

being $\phi(\lambda) = e^\lambda - 1 - \lambda$.

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.

We now prove exponential tail bounds for the moment generating function of the rare mass $M_{n,r}$, more precisely in the following proposition we show 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*

$$\log \mathbb{E} \left[e^{\lambda[M_{n,r} - \mathbb{E}[M_{n,r}]]} \right] \leq \frac{\lambda^2 v_{n,r}^-}{2}. \quad (25)$$

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

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

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

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

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 $\lambda \in [0, n - r)$

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

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

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

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

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

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

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

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

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

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

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

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. $\blacksquare$

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*

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

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

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

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

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

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

where

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

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

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

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

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

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*

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

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*

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

Corollaries 7 and 9 provide us with concentration inequalities for the rare 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 rare 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 mass of rare features $M_{n,r}$ 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 rare 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. Second it is a feasible quantity from 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 SS takes values in $[0, 1]$, the same estimator in the FS 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 somehow 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 $f/g \rightarrow 1$; 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 $\nu(dx) := \sum_{i \geq 1} \delta_{p_i}(dx)$ and define the measure $\bar{\nu}(x) := \nu[x, 1]$, 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 $\bar{\nu}(x) \simeq x^{-\alpha} \ell(1/x)$ as $x \downarrow 0$, where $\ell(t)$ is a slowly varying function, that is $\ell(ct)/\ell(t) \rightarrow 1$ as $t \rightarrow +\infty$ 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 $j \rightarrow \infty$, 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 Gneden 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 $n \rightarrow +\infty$, $\mathbb{E}[K_{n,r}] \simeq \frac{\alpha \Gamma(r-\alpha)}{r!} n^\alpha \ell(n)$ and $\mathbb{E}[K_n] \simeq \Gamma(1-\alpha) n^\alpha \ell(n)$.*

Proof It is worth recalling the notation already used in Section 4

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

roughly speaking $\Phi_{n,r}$ 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 $n \rightarrow +\infty$, 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 (Gneden et al., 2007, Proposition 13) applies also for the feature setting under regularly varying heavy tails, indeed the measure defined by $\nu_r(dp) := p^r \nu(dp)$ is such that

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

Since $\Phi_{n,r} = n^r / r! \int_0^1 e^{-np} \nu_r(dp)$ 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

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

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

$$|\mathbb{E}[K_{n,r}] - \Phi_{n,r}| \leq \frac{c}{n} \max\{\Phi_{n,r}, \Phi_{n,r+2}\} \rightarrow 0, \quad (40)$$

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

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

in other words we have shown that $\Phi_{n,r} \simeq \mathbb{E}[K_{n,r}]$ as $n \rightarrow +\infty$. 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

$$\left| \frac{\mathbb{E}[K_n]}{\Phi_n} - 1 \right| = \frac{|\mathbb{E}[K_n] - \Phi_n|}{\Phi_n} \stackrel{(28)}{\leq} \frac{2}{n} \rightarrow 0$$

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

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

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

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.

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

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

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 $C > 0$

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

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

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

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 $n \rightarrow \infty$ to conclude that for any fixed $\varepsilon > 0$

$$\sum_{n \geq 1} \mathbb{P}(|M_{n,r}/\mathbb{E}[M_{n,r}] - 1| \geq \varepsilon) \lesssim \sum_{n \geq 1} \exp\{-Cn^\alpha \ell(n)\} < +\infty \quad (42)$$

for a suitable constant $C > 0$. Then, 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 rare 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 FS framework some recent results on the consistent estimation of the rare 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 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 $n \rightarrow +\infty$;
- 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 rare 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 rare mass $M_{n,r}$, for $r \geq 0$.

Our work paves the way to explore new research directions in the estimation of the rare 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 rare probabilities $P_{n,r}$ may be reconsidered under the more general FS 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 FS 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.

Appendix A. Technical lemmas

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

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

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

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

$$-M_{n,0}\varepsilon < \hat{T}_{n,0} - M_{n,0} < M_{n,0}\varepsilon. \quad (43)$$

From the lower bound of (43), 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 (43) 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, $\mathbb{P}\left(\left|\frac{\hat{T}_{n,0}}{M_{n,0}} - 1\right| \geq \varepsilon\right) \geq \mathbb{P}\left(\left|\frac{M_{n,0}}{\hat{T}_{n,0}} - 1\right| \geq 2\varepsilon\right)$, proving (11). ■

In the following two lemmas, we denote by $\tilde{N}_n$ a completely random measure with Lévy intensity $\nu_n(ds, dF) = e^{-sn}\nu(ds, dF) = e^{-s(n+1)}/s \, ds \mathbb{1}_{(0,1)}(F) dF$. 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

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

which is the characteristic function of a $\text{Gamma}(1, n+1)$ 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

$$\begin{aligned} \mathbb{P}\left(1 - \frac{\bar{\varepsilon}}{2} \leq \frac{M_{n,0}}{S_n} \leq 1\right) - 1 &\geq \mathbb{P}(S_n \leq \bar{\varepsilon}) - 1 \\ &= (n+1) \int_0^{\bar{\varepsilon}} e^{-x(n+1)} dx - 1 = -e^{-\bar{\varepsilon}(n+1)}, \end{aligned} \tag{44}$$

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

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. *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 log-normal 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**.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Consistent estimation of rare masses in feature sampling
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Submitted to the Journal of Machine Learning Research, and is now under major revision

Student Confirmation

Student Name:	Fadhel Ayed		
Contribution to the Paper	I was in charge of section 3 on the non existence of consistent estimators and section 5 on Consistency of the estimator under regular variation		
Signature	<i>Mohamed El Fadhel Ayed</i>	Date	12/07/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	<i>Professor F Caron</i>		
Supervisor comments			
Signature		Date	<i>14/07/2020</i>

This completed form should be included in the thesis, at the end of the relevant chapter.

Discussion

8

Discussion

Contents

8.1 Summary	189
8.2 Future directions	190

This chapter concludes the dissertation by summarizing the presented work and outlining potential future work directions.

8.1 Summary

This dissertation presents contributions to methodological and theoretical aspects of BNP. Part I, dedicated to modeling aspects, proposes frameworks building on completely random measures with applications to network, text, and financial modeling. These models focus on capturing power-law properties that are empirically observed. In particular, we introduce a novel class of completely random measures that exhibit a double power-law behavior and show that it outperforms classical models in several fields. Part II analyzes the problem of the missing mass and rare events estimation in the species sampling and feature allocation models. We show that one can use BNP beyond modeling purposes to prove lower-bounds on rates of convergence of any estimator.

8.2 Future directions

Our work opens many exciting directions for further research that we briefly discuss below.

The network model presented in Chapter 2 leads to asymptotically dense graphs. To capture sparse networks, we can consider the distinctive notion of infinite exchangeability developed in [Caron and Fox, 2017]. The proposed methodology can also be generalized to non-count data, as long as the likelihood is closed under convolution; one example that could be of particular interest would be to consider Gaussian distributions.

Research on the stochastic gradient descent’s theoretical properties has attracted much attention in the past few years as this algorithm is a building block of the neural network methodologies. One of the subjects of interest is the gradient noise associated with the stochastic gradient. Often, this noise is considered Gaussian, allowing the dynamics of the gradient to be analyzed through a stochastic differential equation driven by a Brownian motion. A new line of work initiated in [Simsekli and Sagun, 2019] challenged the Normal assumption. Based on an empirical investigation of the gradient noise, they show that the distribution of this noise at later iterations is far from Gaussian, and has a heavy tail. Therefore, they replace the Brownian motion by another Lévy process, accounting for such tails, namely the Stable process. One possible direction that we are considering now is to replace the stable process by the more general class of double power-law completely random measures introduced in Chapters 3 and 4, as early empirical experiments seem to indicate that this class gives better fit.

In the context of the missing mass and rare events estimations problems, we believe that the BNP approach could be used to study the rate of consistency of the GT estimator for estimating rare events and show that this rate is somehow optimal. From a methodological perspective, the non-asymptotic confidence intervals can for example be exploited to extend the upper confidence bound multi-armed bandit strategy of [Bubeck et al., 2013].

Bibliography

- [Barndorff-Nielsen et al., 2012] Barndorff-Nielsen, O. E., Mikosch, T., and Resnick, S. I. (2012). Lévy processes: theory and applications. Springer Science & Business Media.
- [Bernardo et al., 2003] Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. Bayesian statistics, 7:733–742.
- [Bertoin, 1996] Bertoin, J. (1996). Lévy processes, volume 121. Cambridge university press Cambridge.
- [Blackwell et al., 1973] Blackwell, D., MacQueen, J. B., et al. (1973). Ferguson distributions via pólya urn schemes. The annals of statistics, 1(2):353–355.
- [Bubeck et al., 2013] 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(Feb):601–623.
- [Caron and Fox, 2017] Caron, F. and Fox, E. B. (2017). Sparse graphs using exchangeable random measures. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 79(5):1295.
- [Clark, 1973] Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. Econometrica: journal of the Econometric Society, pages 135–155.
- [Colwell et al., 2012] Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., 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(1):3–21.
- [Cont et al., 1997] Cont, R., Potters, M., and Bouchaud, J.-P. (1997). Scaling in stock market data: stable laws and beyond. In Scale invariance and beyond, pages 75–85. Springer.
- [Daley and Vere-Jones, 2007] Daley, D. J. and Vere-Jones, D. (2007). An introduction to the theory of point processes: volume II: general theory and structure. Springer Science & Business Media.
- [Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22.
- [Di Benedetto et al., 2020] Di Benedetto, G., Caron, F., and Teh, Y. W. (2020). Non-exchangeable feature allocation models with sublinear growth of the feature sizes. arXiv preprint arXiv:2003.13491.

- [Diebolt and Robert, 1994] Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society: Series B (Methodological), 56(2):363–375.
- [Ferguson, 1973] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. The annals of statistics, pages 209–230.
- [Fisher et al., 1943] Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. The Journal of Animal Ecology, pages 42–58.
- [Ghahramani and Griffiths, 2006] Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the Indian buffet process. In Advances in neural information processing systems, pages 475–482.
- [Ghosal and Van der Vaart, 2017] Ghosal, S. and Van der Vaart, A. (2017). Fundamentals of nonparametric Bayesian inference, volume 44. Cambridge University Press.
- [Gnedin et al., 2007] Gnedin, A., Hansen, B., Pitman, J., et al. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. Probability surveys, 4:146–171.
- [Good, 1953] Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. Biometrika, 40(3-4):237–264.
- [Görür et al., 2006] Görür, D., Jäkel, F., and Rasmussen, C. E. (2006). A choice model with infinitely many latent features. In Proceedings of the 23rd international conference on Machine learning, pages 361–368.
- [Grabchak and Zhang, 2017] Grabchak, M. and Zhang, Z. (2017). Asymptotic properties of Turing’s formula in relative error. Machine Learning, 106(11):1771–1785.
- [Hjort et al., 1990] Hjort, N. L. et al. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. the Annals of Statistics, 18(3):1259–1294.
- [Hjort et al., 2010] Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). Bayesian nonparametrics, volume 28. Cambridge University Press.
- [Hougaard, 1986] Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. Biometrika, 73(2):387–396.
- [James et al., 2017] James, L. F. et al. (2017). Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors. The Annals of Statistics, 45(5):2016–2045.
- [Kingman, 1967] Kingman, J. (1967). Completely random measures. Pacific Journal of Mathematics, 21(1):59–78.
- [Kingman, 1993] Kingman, J. F. C. (1993). Poisson processes. Encyclopedia of biostatistics, 6.
- [Korwar et al., 1973] Korwar, R. M., Hollander, M., et al. (1973). Contributions to the theory of Dirichlet processes. The Annals of Probability, 1(4):705–711.

- [Lijoi et al., 2010] Lijoi, A. et al. (2010). Models beyond the Dirichlet process. Bayesian nonparametrics, 28(80):342.
- [Lijoi et al., 2007] Lijoi, A., Mena, R. H., and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4):715–740.
- [Madan and Seneta, 1990] Madan, D. B. and Seneta, E. (1990). The variance gamma (VG) model for share market returns. Journal of business, pages 511–524.
- [Mossel and Ohannessian, 2019] Mossel, E. and Ohannessian, M. I. (2019). On the impossibility of learning the missing mass. Entropy, 21(1):28.
- [Ohannessian and Dahleh, 2012] Ohannessian, M. I. and Dahleh, M. A. (2012). Rare probability estimation under regularly varying heavy tails. In Conference on Learning Theory, pages 21–1.
- [Pitman, 2003] Pitman, J. (2003). Poisson-Kingman partitions. Lecture Notes-Monograph Series, pages 1–34.
- [Regazzini et al., 2003] Regazzini, E., Lijoi, A., Prünster, I., et al. (2003). Distributional results for means of normalized random measures with independent increments. The Annals of Statistics, 31(2):560–585.
- [Simsekli and Sagun, 2019] Simsekli, U. and Sagun, L. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks. Proceedings of Machine Learning Research, 97.
- [Tankov, 2003] Tankov, P. (2003). Financial modelling with jump processes. CRC press.
- [Teh and Gorur, 2009] Teh, Y. W. and Gorur, D. (2009). Indian buffet processes with power-law behavior. In Advances in neural information processing systems, pages 1838–1846.
- [Thibaux and Jordan, 2007] Thibaux, R. and Jordan, M. I. (2007). Hierarchical Beta processes and the Indian buffet process. In Artificial Intelligence and Statistics, pages 564–571.
- [Zhou, 2015] Zhou, M. (2015). Infinite edge partition models for overlapping community detection and link prediction. In Artificial intelligence and statistics, pages 1135–1143.