Babagnidé François Koladjo[1] / Mesrob I. Ohannessian[2] / Elisabeth Gassiat[3]

A Truncation Model for Estimating Species Richness

[1] ENSPD, Universite de Parakou, Parakou, Benin, E-mail: francois.koladjo@gmail.com
[2] Toyota Technological Institute at Chicago, Chicago, USA
[3] Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, Univ. Paris-Sud, CNRS, 91405 Orsay, France

Abstract:
We propose a truncation model for the abundance distribution in species richness estimation. This model is inherently semiparametric and incorporates an unknown truncation threshold between rare and abundant observations. Using the conditional likelihood, we derive a class of estimators for the parameters in this model by stepwise maximization. The species richness estimator is given by the integer maximizing the binomial likelihood, given all other parameters in the model. Under regularity conditions, we show that our estimators of the model parameters are asymptotically efficient. We recover Chaos lower bound estimator of species richness when the parametric part of the model is single-component Poisson. Thus our class of estimators strictly generalized the latter. We illustrate the performance of the proposed method in a simulation study, and compare it favorably to other widely-used estimators. We also give an application to estimating the number of distinct vocabulary words in French playwright Molière's Tartuffe.

Keywords: species richness, semiparametric model, model selection, abundance distribution

DOI: 10.1515/ijb-2017-0035

Received: May 11, 2017; **Revised:** January 4, 2018; **Accepted:** June 19, 2018

1 Introduction

We consider the "species richness" problem, also known as the problem of estimating the number of species, which arises when a sample of individuals is taken from a population with N classes or species. The usual data set is a series of observed counts $X_1^+, \dots, X_D^+$, with $D \leq N$ being the total number of distinct species observed in the sample and N being the parameter to be estimated. Estimating N using such abundance data is an old problem that has been tackled in several ways, both by parametric models, including Bayesian models [1, 2], and by nonparametric models [3]. Due to their flexibility to account for heterogeneity, the nonparametric approaches are those predominantly considered in the last two decades. This setting contains among others the Chao-type estimators developed by Chao and collaborators (see for example [4–6]), and the likelihood-based nonparametric estimators of which one can cite [7, 8]

Many of these methods, although theoretically founded on a single model, perform the common practice of truncating the data into abundant and rare species. One then assumes that the number of abundant species is adequately represented by the number of distinct such species, whereas the same number leads to an underestimate for the rare species and thus necessitates a correction. Such truncation is generally justified on the basis of avoiding instability. This, however, forces even initially nonparametric models to become effectively parametric, while losing the original hypothesis and the accompanying theoretical guarantees. This motivated us to study this heuristic in a more rigorous light. In particular, we make the following contributions:

– We give an explicit semiparametric model to represent this truncation practice, where the abundant species are represented by an arbitrary abundance distribution whose support is offset away from the rare range. We partially motivate this as arising from the commonly used Poisson mixtures as being inappropriate for modeling more abundant species.

– We show that the practice of pure truncation as described above is justified only when the abundant and rare species have abundance distributions whose supports are disjoint. In this case truncation leads to an efficient estimation of the number of species.

Babagnidé François Koladjo is the corresponding author.

– In general, although pure truncation is not efficient, accounting for the support overlap leads to a hybrid truncation that is a semiparametric procedure which is efficient. We show this by using standard single-parameter families to derive a local minimax bound and a matching (asymptotically) efficient estimator. Coincidentally, we show that this framework recovers several previously suggested estimators as special cases.

– When the abundance threshold is not known, neither pure truncation nor the hybrid approach can be used directly. For this reason, the proper offset should be obtained from data. We present a model selection approach to resolve this problem. Our experiments show that this approach adapts to the true unknown offset, in the sense that the resulting estimator achieves (almost) the same asymptotic performance as when knowing the offset.

– We illustrate this estimator on both synthetic and real data, showing that our more refined analysis leads to practical improvement.

In Section 2, after a brief introduction to the problem, we present our semiparametric truncation model and the proposed likelihood method. We then prove that our method recovers previous estimators in particular situations, see Propositions 1 and 2. We end the section by proposing a model selection method to choose the truncation parameter. Section 3 is devoted to the semiparametric asymptotic analysis of the estimators, see Theorems 1 and 2. The behaviour on finite samples is investigated through a simulation study in Section 4, and we obtained effectively good results when applying our estimator to observational richness of text-data. We discuss our results in Section 5. All the proofs of theoretical results are detailed in the Appendix.

2 Model and estimator

2.1 Problem statement

Assume that N species exist in nature and that each is represented by $X_1, \cdots, X_N$ individuals in a sample. We call X_i the *abundance* of species i in the sample. A classical statistical model of the abundances is to assume that they are independent random variables identically distributed according to a distribution $f_\nu(x)$ for $x \in \mathbb{N}$ and where ν is an index within a class of abundance distributions. One of the more common choices of abundance distribution classes are Poisson mixtures indexed by a mixing distribution ν on $\mathbb{R}_+$:

$$f_\nu(x) = \int \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \mathrm{d}\nu(\lambda), \quad \text{for } x \in \mathbb{N}. \tag{1}$$

Of course, we do not get to access non-observed species, i.e. species for which $X_i = 0$. If we let D denote the number of distinct observed species, i.e. $D = \sum_{i=1}^{N} \mathbf{1}\{X_i > 0\}$, and if we re-index and relabel those species as $X_1^+, \cdots, X_D^+$, then it is easy to show that these observed abundances are independent and identically distributed according to the zero-truncated distribution:

$$f_\nu^+(x) = \frac{f_\nu(x)}{1 - f_\nu(0)}, \quad \text{for } x \in \mathbb{N}_+. \tag{2}$$

The central problem of this paper is that of estimating the number of species N with a functional $\hat{N}(X_1^+, \cdots, X_D^+)$ of the abundances of the observed species. In other words, $\hat{N}$ needs to complement the number of observed species D with an estimate of the number of non-observed species.

As outlined in the introduction, a long line of research has addressed this problem. But we focus here in particular on a sequence of influential papers [6, 9], the methodology of which continues to be used in more recent papers such as [4, 10]. In theory, these results are within the current framework but, in practice, the estimation is done as follows. The data is divided into rare and abundant components according to an *abundance threshold* τ. Although their estimators are derived and analyzed under the general model, the theoretical estimators are fed with only those abundances such that $X_i^+ \leq \tau$, to yield an estimator of the number of rare species $\hat{N}_{\text{rare}}$. For the abundant species, they use the trivial estimator:

$$\hat{N}_{\text{abundant}} = \sum_{i=1}^{D} \mathbf{1}\{X_i^+ > \tau\}.$$

The estimate for the total number of species is then simply the sum of both:

$$\hat{N} = \hat{N}_{\text{rare}} + \hat{N}_{\text{abundant}}.$$

What is the justification behind such truncation? This paper strives to answer this question and to give a more principled model of this common practice, thus leading to a more transparent methodology.

2.2 Truncation model

To motivate the reason behind truncation, note that the justification often given in this line of work [4, 6, 9, 10] is that including the abundant species into the estimator may cause instabilities. In coverage based estimators (i.e. Chao's CV-based estimators) that account for heterogeneity of species abundance, the instability is due to the estimation of the coefficient of variation of abundances. Indeed it is clear that including more abundant species gives larger estimates of the coefficient of variation. But when using the abundance sampling model, we can also interpret observed instability as this model, and in particular the Poisson mixture model, not being a good model for abundant species. In this section, we first give some informal insight as to why this may be the case. We then proceed to present an explicit model to handle this rare-abundant dichotomy.

The abundance model can be traced back to a simple sampling model where individuals are drawn independently and identically (with replacement) from a population, where the frequency of species i is p_i. If m such individuals are drawn, let $Y_1, \cdots, Y_m$ denote their species. In this model, the abundance of species i has therefore a binomial distribution of parameters m and p_i:

$$X_i = \sum_{j=1}^{m} \mathbf{1}\{Y_j = i\} \sim \text{binomial}(m, p_i). \tag{3}$$

If the species are not labeled a priori, which corresponds to a random permutation among the N species, then the distribution of a particular abundance becomes a mixture of binomial distributions, with mixture weights at $(m, p_i)_{i=1,\cdots,N}$. Note that these abundances are not independent as in the abundance model, but are rather exchangeable. Notwithstanding this fact, we can see that the abundance model of eq. (1) effectively replaces this binomial mixture with a Poisson mixture, which cannot be accurate for abundant species.

Indeed, a Poisson distribution with a large mean places much more mass near 0 compared to a corresponding binomial. More precisely, if the model substitutes a binomial mixture with a Poisson mixture, then when an estimator places a mixing mass at a higher abundance, it contributes more to $f_\nu(0)$ than a binomial would. This is then interpreted as evidence of more unseen species than the reality, and thus N is overestimated. This is indeed what is observed with such estimators: with larger values of the truncation τ, the estimate of N tends to increase (see for example the last three columns of Table 2, page 949, and the last two columns of Table 13, page 956, in [10]). That said, simply truncating the data is not a theoretically sound approach since the resulting samples no longer follow the hypothesized model. For example Poisson distributions place a positive mass, even if small, beyond any threshold. There is therefore a need to rigorously model rare species, say with mixtures of Poisson distributions, while capturing the possibility that there may be abundant species that have much less influence on our inference about the rare species.

In this paper, we propose the following semiparametric alternative. Let $\tau \in \mathbb{N}_+$ and let $\mathscr{F}_\tau$ be the family of discrete distributions supported on $\{\tau+1, \tau+2, ...\}$. We assume that abundant species follow a non parametric distribution $F \in \mathscr{F}_\tau$, that rare species follow a parametric distribution R_θ (e.g. we may think of a finite mixture of Poisson distributions) with $\theta \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is an appropriate subset of $\mathbb{R}^k$ for some $k \in \mathbb{N}_+$, and that the proportion of rare species is q, so that abundances follow a distribution f that belongs to a model $\mathscr{P}_\tau$:

$$\mathscr{P}_\tau = \left\{ f_{(q,\theta,F)}(x) = qR_\theta(x) + (1-q)F(x),\ \theta \in \boldsymbol{\Theta},\ F \in \mathscr{F}_\tau,\ q \in (0,1) \right\}. \tag{4}$$

Using eq. (2) and the fact that the nonparametric component vanishes at $x = 0$, this model induces the zero-truncated version as follows:

$$\mathscr{P}_\tau^+ = \left\{ f^+_{(q,\theta,F)}(x) = \frac{f_{(q,\theta,F)}(x)}{1 - qR_\theta(0)}, f_{(q,\theta,F)} \in \mathscr{P}_\tau \right\}. \tag{5}$$

We leave the choice of R_θ open, except for certain identifiability and smoothness assumptions that we later spell out in detail. Thus R_θ is not necessarily a Poisson mixture. The choice of a parametric model for R_θ is justified by the fact that even originally nonparametric models are effectively reduced to parametric classes under the constraint of identifiability from a small (truncated) support.

It is now clear that our model in eq. (5) makes explicit the notion that rare and abundant species may coexist. This allows us to bypass heuristics and suggest estimators with provable performance guarantees. In particular, we may harness the basic theory of semiparametric models to establish the efficiency of likelihood-based estimators, and suggest potential model selection mechanism for the choice of τ. Furthermore, as we make no further assumptions beyond adopting a parametric form for the rare component and dislocating the support of the abundant species away from zero, we have a model that can go beyond a simple justification of truncation. For example, one may think of F as a nonparametric corruption to the data, rather than a legitimate measurement of abundant species, and our analysis and methodology still goes through unaffected.

2.3 Estimator of the number of species

The estimator that we propose for N falls under the category of maximum likelihood (MLE)-type M-estimators. In this section we define and derive the estimator, and in the next section we study some of its asymptotic properties.

The distribution of the abundances is a multinomial distribution, for which the empirical counts of the abundances are sufficient statistics for computing likelihoods. Let $n_x = \sum_{i=1}^{D} \mathbf{1}\{X_i^+ = x\}, x \geq 1$, and notice that $D = \sum_{x\geq 1} n_x$ and $n_0 = N - D$. Notice also that since N is unknown, it appears as a parameter when writing the likelihood. Thus the combined likelihood of N and the rest of the model parameters given the samples can be written as follows:

$$L(N, f|(n_x)_{x\geq 1}) = \frac{N!}{(N-D)! \prod_{x\geq 1} n_x!} f(0)^{N-D} \prod_{x\geq 1} f(x)^{n_x}. \tag{6}$$

One could also consider that, since the species for which the abundance is 0 are not observed, the likelihood is that of a multinomial distribution within the D observed species, that is using the zero-truncated model. After substituting $f^+(x)$ by its expression in $\mathscr{P}_\tau^+$, this likelihood writes

$$L^+(q, \theta, F|(n_x)_{x\geq 1}) = \frac{D!}{\prod_{x\geq 1} n_x!} \prod_{x\geq 1} \underbrace{\left[\frac{qR_\theta(x) + (1-q)F(x)}{1 - qR_\theta(0)}\right]}_{f^+(x)}^{n_x}. \tag{7}$$

Now, it is interesting to note that the likelihood L may be decomposed into the product of two terms, L^+ and another term denoted L_b such that, after substituting $f(x)$ by its expression in $\mathscr{P}_\tau$, L_b is given by

$$L_b(N|D, q, \theta) = \frac{N!}{D!\,(N-D)!} \left[qR_\theta(0)\right]^{N-D} \left[1 - qR_\theta(0)\right]^{D}. \tag{8}$$

The term L_b has a binomial form and may be interpreted as the likelihood of N given the rest of the model parameters and the number of distinct samples D. The term L^+ is the likelihood of the rest of the model parameters, given the samples. This suggests two methods to undertake the maximum likelihood estimation from L. Note that some of the earliest works to suggest such a decomposition were [11, 12] (see also [13, 14] for a more recent treatment).

The first method is to maximize directly the likelihood L over all of (N, q, θ, F). The estimator of N obtained from this method is typically called the *unconditional* maximum likelihood estimator. For example, some nonparametric models with unconditional estimation methods are proposed in [7, 15]. The second method to obtain a maximum likelihood estimator of N is to first maximize the likelihood L^+ from the zero-truncated model $\mathscr{P}_\tau^+$ to derive the estimators of q,θ and F, and then to maximize the binomial likelihood L_b in the parameter N given that q and θ are known. This method is known as the *conditional* maximum likelihood method for estimating N. It should be noted that both conditional and unconditional methods are asymptotically equivalent in a parametric model (see [11, 12]). This results is very important here because our model $\mathscr{P}_\tau$ becomes fully parametric when one replace the nonparametric component F by a parameterized distribution as its pseudo-estimator given at eq. (10). Furthermore, the same procedure as described below for conditional maximum likelihood method leads to the same pseudo-estimator of F for unconditional method. The asymptotic equivalence of the two methods then follows in our semiparametric framework.

We consider here only the conditional maximum likelihood method because it is numerically easier to undertake than the unconditional maximum likelihood. As we shall prove below, the estimators of the parameters θ and q obtained by maximizing L^+ are asymptotically efficient in the semiparametric framework we propose in this work. This result may be seen as a theoretical guarantee to our method. Before we proceed with the

estimation of θ, q , and F, note that maximizing the binomial likelihood L_b in eq. (8) gives us the form of our estimator:

$$\hat{N}(q,\theta) = \frac{D}{1 - qR_\theta(0)}. \tag{9}$$

The final expression for the estimator therefore consists in estimating θ by $\hat{\theta}$ and q by $\hat{q}$, in a manner that we shortly outline, and then substituting in eq. (9) to obtain $\hat{N} = \hat{N}(\hat{q}, \hat{\theta})$. Of course, N is an integer parameter, and we could then take the integer part of the resulting estimate. That said, in what follows we allow ourselves to accept non-integer estimates.

We now proceed to estimate the parameters. We observe first that since F plays no role in the expression for $\hat{N}$, we can treat it as a nuisance parameter. The next observation is that to maximize L^+, we can successively fix some parameters while we maximize over others. Because F is mostly a nuisance parameter, we maximize the likelihood L^+ when q and θ are fixed without further constraining F to be a proper distribution. This approach gives us, at each support point x, the following pseudo-estimator, as a function of θ and q:

$$\hat{F}(q,\theta)(x) = \frac{[1 - q\sum_{k=0}^{\tau} R_\theta(k)]}{(1-q)(D - D_\tau)} n_x - \frac{q}{1-q} R_\theta(x), \tag{10}$$

where $D_\tau = \sum_{x=1}^{\tau} n_x$ denotes the number of species with abundance no greater than τ.

The reason we call $\hat{F}(q,\theta)$ a pseudo-estimator is that it may put negative mass at some of its support points as it is not constrained to be nonnegative. This occurs for example at the non-observed support points of F, that is for a support point x such that $n_x = 0$. Despite this fact, the estimators for θ and q that follow from this choice of $\hat{F}$ are not sensitive to its impropriety.

Replacing F by its pseudo-estimator in the expression for L^+ leads to an objective function for q and θ which may now be maximized in q. This leads to an MLE-type estimator of q, still as a function of θ:

$$\hat{q}(\theta) = \frac{1}{R_\theta(0) + \frac{D}{D_\tau}\sum_{k=1}^{\tau} R_\theta(k)}. \tag{11}$$

Note that q is always non-negative. However, for particular values of θ, D , and D_τ, it could be larger than 1. If this occurs in practice, we simply constrain it to 1 to obtain a valid probability. The consistency result in the next section shows that this is not a concern, asymptotically.

The last step is to find a proper estimator of θ. Consider the following simplifying notation. For a fixed τ, let S_θ^τ denote the truncated version of the density R_θ defined as

$$S_\theta^\tau(x) = \frac{R_\theta(x)}{\sum_{k=1}^{\tau} R_\theta(k)} \text{for } 1 \le x \le \tau. \tag{12}$$

By replacing F and q by their estimators in the conditional likelihood L^+, we can show that we obtain (up to factors that do not depend on θ) the following truncated likelihood:

$$\prod_{x=1}^{\tau} \left\{S_\theta^\tau(x)\right\}^{n_x}. \tag{13}$$

The estimator $\hat{\theta}$ is then simply a maximizer of eq. (13). We can thus see that $\hat{\theta}$ is an MLE of the truncated density S_θ^τ, based on the first τ abundance counts. This completes our estimator construction. Indeed, to estimate N, we first compute $\hat{\theta}$ directly from the samples by maximizing eq. (13), we then calculate $\hat{q}(\hat{\theta})$ using eq. (11), and lastly we substitute both to obtain $\hat{N}(\hat{q}(\hat{\theta}), \hat{\theta})$ using eq. (9).

We conclude by noting that all the derivations we performed were based on the premise that a value of τ was given. As $\hat{q}$ and $\hat{\theta}$ depend on τ, in what follows either we make this explicit by writing $\hat{q}_\tau$ and $\hat{\theta}_\tau$ respectively or keep it implicit when the notation gets encumbered. Similarly we write $\hat{N}_\tau$. We also sometimes use the notation $\hat{q}(\hat{\theta})$ instead of $\hat{q}$ to make it explicit that the estimator of q depends on $\hat{\theta}$.

2.4 Relationship to other estimators

Despite the fact that θ is estimated by truncating the model to the abundance values between 1 and τ, our estimator differs from the traditional truncation with conditional MLE-type estimators often described in the

literature, as overviewed in the introduction. To be precise, assume the same parametric rare-species model is used for R_θ, and recall that in these classical estimators the data is truncated and the conditional MLE is solved using the zero-truncated version of R_θ to obtain $\hat{\theta}$, and then the rare-species count is estimated by:

$$\hat{N}_{\text{rare}} = \frac{D_\tau}{1 - R_{\hat{\theta}}(0)}.$$

The abundant species are then assumed to be represented exactly by what is seen:

$$\hat{N}_{\text{abundant}} = D - D_\tau.$$

The combined estimator is therefore:

$$\begin{aligned} \hat{N}_{\text{classical}} &= \hat{N}_{\text{rare}} + \hat{N}_{\text{abundant}} \\ &= \frac{D_\tau}{1 - R_{\hat{\theta}}(0)} + (D - D_\tau). \end{aligned}$$

The following proposition identifies the condition under which our estimator is equivalent to this classical estimator.

Proposition 1
If all R_θs are supported on $\{0, \dots, \tau\}$, then the two estimators $\hat{N}_\tau$ and $\hat{N}_{\text{classical}}$ are equal.

Proposition 1 means that if the parametric part R_θ and the nuisance parameter in the model $\mathscr{P}_\tau^+$ are supported on disjoint sets, then one can split the data set into rare-species data $(X_i \leq \tau)$ and abundant-species data $(X_i > \tau)$. In this context, inference on rare species is not affected by the estimation of the nuisance parameter F and thus throwing away high-abundance data is justified. On the other hand when R_θ does extend over all integers, then one should not ignore any part of the data, and instead one should perform a hybrid truncation, as suggested by N_τ in order to obtain efficient estimators.

Thus far we have considered the general context for any eligible distribution R_θ. Some particular cases enable us to make simple and concrete connections between $\hat{N}_\tau$ and other popular estimators that come close to falling within our framework. In particular, Chao, in [16], suggests the following popular estimator

$$\hat{N}_{\text{Chao}} = D + n_1^2/2n_2.$$

The following proposition shows that our estimator $\hat{N}_\tau$, for $\tau = 2$ and R_θ corresponding to a pure Poisson distribution, is equal to Chao's $\hat{N}_{\text{Chao}}$. As such, we can interpret our estimator as a generalization of Chao's, where τ is no longer restricted to 2 and where R_θ may be more general than a pure Poisson distribution.

Proposition 2
Assume that $\tau = 2$ and $R_\theta(x) = \theta^x e^{-\theta}/x!$ for all $x \geq 0$. Then $\hat{N}_\tau = \hat{N}_{\text{Chao}}$.

It is worth noting that Zelterman [17] explicitly considers the pure Poisson model with access only to the first two counts $n_1 \neq 0$ and n_2, and suggested $\hat{\theta}_{\text{Zelterman}} = 2n_2/n_1$ as an estimator for θ, showing certain robustness properties under heterogeneity in the true model (see [17] for more details). It is indeed straightforward to verify that for $\tau = 2$ and a pure Poisson model for R_θ, we have that our estimator maximizing the truncated likelihood of eq. (13) corresponds to that of Zelterman: $\hat{\theta}_\tau = \hat{\theta}_{\text{Zelterman}}$.

An effective proof of Proposition 2 is given in the work of Böhning et al in [18] within an alternative framework: using the conditional expectation of f_0. In the Appendix, we propose a simpler proof that is more in line with our framework, by plugging-in $\hat{\theta}_{\text{Zelterman}}$, which as noted is the correct conditional ML estimate, into our expression for $\hat{N}_\tau$.

We end by noting that if the abundant species are not taken into account, we would have the estimator of N given by $\hat{N}_{\text{Zelterman}} = D/[1 - \exp(-\hat{\theta}_{\text{Zelterman}})]$. This estimator is in the spirit of pure-truncation, and would clearly deviate from $\hat{N}_\tau$ (the denominator has no q factor). Since we establish the latter to be consistent within our model, then it follows that the former is not (see also [19] for a more quantitative comparison between Chao's and Zelterman's estimators of N.)

2.5 Choice of τ via model selection

To end the discussion of our estimator of N, we stress once again that $\hat{N}_\tau$ depends on the integer truncation parameter τ, which delimits the zone of influence of the abundant species through the support of the nuisance

parameter F. When τ is not known, we need a procedure to estimate this parameter. This is effectively a model selection problem, which we now address using the Goldenshluger-Lepski (G-L) method as inspiration. The G-L method was introduced in [20] in the context of bandwidth selection for kernel density estimation. In the current paper we use it heuristically, without formal proofs. Experimental evidence, however, suggests that the method is very effective.

The principle of the method is as follows. As our estimator is of the form $\hat{N} = D/(1 - \hat{P}(0))$, we focus on the problem of estimating $P(0) = qR_\theta(0)$. Let us drop the (0) argument from the notation, to make the exposition clearer. Assume that we have a known upper bound $\tau_{\max}$ on the largest value τ could take, and let $\tau_{\min}$ be the least τ that enables the necessary identifiability assumptions. If we relax the requirement that F is positive on its support, we have successively smaller nested models as τ varies from $\tau_{\min}$ to $\tau_{\max}$. Each of these models has a corresponding version of our estimator, that we denote by $\hat{P}_\tau$. The (squared) *bias* of each model is $\mathsf{bias}_\tau = (\mathbf{E}[\hat{P}_\tau] - P)^2$. The variance of each model is $\mathsf{var}_\tau = \mathbf{E}[(\hat{P}_\tau - \mathbf{E}[\hat{P}_\tau])^2]$. The mean squared error *risk* decomposes as usual into the sum of bias and variance, $\mathsf{risk}_\tau = \mathbf{E}[(\hat{P}_\tau - P)^2] = \mathsf{bias}_\tau + \mathsf{var}_\tau$. Now observe the following:

- For $\tau \leq \tau_0$, the consistency result of Theorem 1 tells us we are asymptotically unbiased.
- For $\tau = \tau_0$Theorem 2 shows that we are efficient and therefore asymptotically we have the least variance.
- For $\tau < \tau_0$, the estimator becomes inefficient and the variance may be higher. Intuitively, this is because less of the data is used to estimate θ when the truncation is stricter.
- For $\tau > \tau_0$, Theorem 1 tells us that we may have a non-vanishing bias. However, the variance itself may be lower simply because more F-corrupted data is used to converge to an incorrect value of θ.

The inevitable bias-variance tradeoff thus manifests itself in this framework, and the best compromise in terms of risk will be achieved at the correct model class τ_0. If accurate proxies $\widehat{\mathsf{bias}}_\tau$ and $\widehat{\mathsf{var}}_\tau$ are available, then we may empirically select a model $\hat{\tau}$ near τ_0, by minimizing

$$\hat{\tau} = \operatorname{argmin}_\tau \left(\widehat{\mathsf{bias}}_\tau + \widehat{\mathsf{var}}_\tau \right). \tag{14}$$

The bootstrap method is one effective way for estimating var_τ. In its simplest version, bootstrap consists in resampling D points from the data and computing an estimator $\tilde{P}_\tau$ from the resampled data. Then this is repeated a number of times, say $j = 1, \cdots, M$, and the variance is estimated as:

$$\widehat{\mathsf{var}}_\tau = \frac{1}{M} \sum_{j=1}^{M} (\tilde{P}_{\tau,j} - \hat{P}_\tau)^2. \tag{15}$$

While the resampling process of the bootstrap is good at quantifying the *relative* (to $\hat{P}_\tau$) variability of the resampled estimators, it offers no *absolute* reference point, crucial for estimating the bias. Luckily, as we have argued, the larger model classes have small bias and can themselves be used as a reference point. The Goldenshluger-Lepski method suggests the following method to obtain a bias proxy:

$$\widehat{\mathsf{bias}}_\tau = \max_{\tau' \leq \tau} \left[(\hat{P}_{\tau'} - \hat{P}_\tau)^2 - \widehat{\mathsf{var}}_{\tau'} \right]_+, \tag{16}$$

where $[\cdot]_+$ stand for the non-negative part. The justification and behavior for this bias proxy needs to be rigorously established, as is done for kernel width selection in [20]. For our heuristic use, we provide simply the intuition behind it. This formula can be interpreted by noticing that the maximum of $(\mathbf{E}[\hat{P}_{\tau'}] - \mathbf{E}[\hat{P}_\tau])^2$ over $\tau' \leq \tau$ is indeed approximately the bias since, as we described, the larger models are (asymptotically) unbiased. But because we only have access to $(\hat{P}_{\tau'} - \hat{P}_\tau)^2$ instead of $(\mathbf{E}[\hat{P}_{\tau'}] - \mathbf{E}[\hat{P}_\tau])^2$, and since the larger models have higher variance, we place a conservative confidence bound on the τ' end using $\widehat{\mathsf{var}}_{\tau'}$ in order not to overestimate the bias.

Equations (14) (the selection of $\hat{\tau}$), (15) (the bootstrap variance proxy), and (16) (the bias proxy) completely specify a heuristic model selection procedure for estimating the integer truncation parameter τ.

3 Analysis of the estimator

3.1 The semiparametric framework

We now analyze the convergence and optimality of our estimator in the context of efficient estimation, when the model contains nuisance parameters. We do so particularly in order to handle the nonparametric component

F within our semiparametric model. In the absence of such nuisance parameters, efficiency may be defined in terms of attaining the Cramér-Rao bound. In regular parametric models, the Cramer-Rao bound is the variance of the score function, itself (often) defined as the derivative of the log-likelihood, and efficient estimators are at first order empirical means of the score function. The nuisance parameters, however, can lead to unavoidable loss in the accuracy of any estimator. The notion of efficiency can then be extended by assessing new lower bounds to the variance of the parameters of interest. We provide the details in Section A.4 below, and describe here what is useful to state our results.

One can define a set $\dot{\mathscr{P}}_F^+$ of score functions relatively to the nonparametric part of the model, built using one dimensional submodels (see Section A.4 for details). Then, if $\dot{\ell}_{(q,\theta)}$ is the usual score function (given by the partial derivative and gradient with respect to q and θ respectively of the log-likelihood in the full model), the *efficient score function* related to (q, θ) is then defined component-wise as $\tilde{\ell}_{(q,\theta)} = \dot{\ell}_{(q,\theta)} - \boldsymbol{\Pi}_F \dot{\ell}_{(q,\theta)}$, where $\boldsymbol{\Pi}_F$ is the orthogonal projection onto the closure of the linear space spanned by $\dot{\mathscr{P}}_F^+$. The efficient score functions play the same role for efficient estimators (if they exist) as the ordinary score functions for the maximum likelihood estimators in a parametric model with no nuisance parameter. Namely, they lead to the best asymptotic variance for any estimator. The corresponding efficient Fisher information $\tilde{I}_{(q,\theta)}$ is a matrix whose components are the variances and covariances of the various components of the vector of efficient score functions.

As such, this leads to what we shall give as formal definition of the properties of consistency and efficiency:

Definition 1.

As $N \to \infty$, an estimator sequence $T_D = (\hat{q}, \hat{\theta})$ is:

- *Consistent, if $T_D \to (q, \theta)$ in probability.*
- *Efficient (asymptotically), if*

$$\sqrt{D}\,(T_D - (q, \theta)) = \frac{1}{\sqrt{D}} \sum_{i=1}^{D} \tilde{I}^{-1}_{(q,\theta)} \tilde{\ell}_{(q,\theta)}(X_i^+) + o_P(1).$$

Note that the typical asymptotics for estimator sequences rely on increasing sample size. The sample size in our problem is D, as the samples consist of the positive (observed) abundances $X_1^+, \cdots, X_D^+$. Thus, the sample size is a random quantity. Despite this, it is clear that as $N \to \infty$, we also have that $D \to \infty$ in probability, and we therefore think of the two asymptotic notions interchangeably.

One of the challenges is that in many models the efficient score is not amenable to be used in the same way as the ordinary score because the orthogonal projection $\boldsymbol{\Pi}_F$ might not be available in closed form. In Proposition 3 (stated and proved in Section A.4), we show that such a closed form can be obtained in our model, and give the expressions that ensue for the efficient score functions for estimating the parameters θ and q in the model $\mathscr{P}_\tau^+$.

3.2 Consistency and efficiency

In what follows, when the true model lies within the hypothesized class $(\mathscr{P}_\tau)_{\tau \geq 1}$, we refer to the true parameters by θ_0, q_0 , and F_0, and to the true truncation by τ_0. We first list some regularity assumptions that we have recourse to throughout.

Assumptions

1. [Compactness] $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^k$.
2. [Identifiability] The parameter θ is identifiable from the truncated density S_θ^τ, as defined by eq. (12).
3. [Continuity] For all x in $\{1, \dots, \tau\}$,$\theta \mapsto R_\theta(x)$ is a continuous function of θ, and $R_\theta(x) \geq \delta > 0$ for all θ in $\boldsymbol{\Theta}$ and $x \leq \theta_0$.

Let us now move to the main results of this section, the consistency and efficiency of $\hat{q}_\tau$ and $\hat{\theta}_\tau$ whenever $\tau \leq \tau_0$, and some further properties that give more insight into these estimators. We begin with the consistency result stated below as Theorem 1.

Theorem 1.

Under Assumptions 1–3, as N tends to infinity, the following results hold:

(i) If $\tau \leq \tau_0$, then $\hat{\theta}_\tau$ and $\hat{q}_\tau$ converge in probability to θ_0 and q_0 respectively.

(ii) If $\tau > \tau_0$, then $\hat{\theta}_\tau$ converges in probability to the set of maximizers of $M^\tau(\theta) = \sum_{x=1}^{\tau} f^+(x) \log S_\theta^\tau(x)$.

The results in Theorem 1 are remarkable since they ensure the consistency of $\hat{\theta}_\tau$ and $\hat{q}_\tau$ for a fixed τ, as long as it is smaller than or equal to its true value τ_0 and identifiability holds. If, however, one chooses τ greater than τ_0, then the proposed estimators may not be consistent. (This leads to the challenge of choosing τ via model selection when τ_0 is unknown, as described in Section 2.5). We now complement this consistency result with efficiency properties.

Theorem 2.
Consider Assumptions 1-3, and assume further that $\theta \mapsto R_\theta(x)$ is $\mathscr{C}^2$ for all $x \in \mathbb{N}$, that θ_0 is an interior point of $\boldsymbol{\Theta}$, and that the efficient Fisher information is non-singular at (q_0, θ_0). Then, $(\hat{q}_{\tau_0}, \hat{\theta}_{\tau_0})$ is asymptotically efficient at (q_0, θ_0).

Remark 1
The estimators $\hat{q}_\tau$ and $\hat{\theta}_\tau$ have the following properties:

(i) $\hat{q}_\tau$ depends on the observations x_i no greater than τ and on the number of those x_i that are greater than τ.

(ii) $\hat{\theta}_\tau$ depends only on the observations x_i no greater than τ.

These follow either from direct inspection or from the proof of Lemma 2 in the Appendix. In particular, $\hat{\theta}$ solves the efficient score eq. (17) which depends only on abundances x_i no greater than τ. Now, from eq. (42), for a given estimator $\hat{\theta}$ of θ, the estimator $\hat{q}(\hat{\theta})$ depends on the x_i greater than τ only through their cardinal $D - D_\tau$ and the property follows.

Theorem 2 asserts the efficiency of $\hat{\theta}_\tau$ and $\hat{q}_\tau$, and through them of the corresponding estimator of the total number of species $\hat{N}_\tau$. Remark 1 also sheds light on the fact that the latter depends only on: (1) the threshold τ, (2) the number of observed species D and (3) on the abundances of rare species (those that are not greater than τ). In other words, as in the case of pure truncation, the abundant species contribute only through their cardinality. That said, $\hat{N}_\tau$ distinguishes itself by using this cardinality to estimate how to weigh appropriately the respective contributions of both the rare and abundant species, using the parameter q. Notice that the fact that q is positive is crucial to be able to carry out inference with this model.

4 Simulations and experiments

To illustrate the impact of truncation on our ability to estimate the number of species, we give some numerical simulations and experiments. To make our theoretical work concrete and results easily reproducible, we consider simple parametric families. In particular we look at a single Poisson distribution and a Gamma-Poisson mixture, which gives rise to the negative binomial distribution. In Section 4.1, we perform synthetic experiments for both, and use this to illustrate the heuristic method of selecting the best truncation. In Section 4.2, we consider real data in the form of literary texts, and confine ourselves to the negative binomial model. In order to be able to compare to a known ground truth, we adapt our number of species framework to the very related observational richness problem, and show that the choice of truncation has a significant impact on estimation accuracy.

4.1 Number of species simulations

4.1.1 Algorithms to compute $\hat{\theta}_\tau$

As we take R_θ to be a parametric family, many of the EM-style MLE algorithms for parameter estimation in such frameworks can be adapted to zero- to τ-truncated versions of the distributions. This is all that's needed since, for a fixed value of τ, computing $\hat{\theta}_\tau$ amounts to maximizing eq. (13), which is equivalent to solving

$$\sum_{x=1}^{\tau} \frac{\dot{R}_\theta(x)}{R_\theta(x)} n_x - D_\tau \frac{\sum_{k=1}^{\tau} \dot{R}_\theta(k)}{\sum_{k=1}^{\tau} R_\theta(k)} = 0. \tag{17}$$

For example, when R_θ is a Poisson distribution, it is not difficult to check that eq. (17) becomes exactly

$$\frac{\bar{X}^\tau}{\theta} = \frac{\sum_{k=0}^{\tau} R_\theta(k-1)}{\sum_{k=1}^{\tau} R_\theta(k)}, \text{with } \bar{X}^\tau = \frac{1}{D_\tau} \sum_{x=1}^{\tau} x n_x \tag{18}$$

leading to the fixed point equation $\theta = \bar{X}\tau \frac{\mathbf{P}_\theta(\tau)-\exp(-\theta)}{\mathbf{P}_\theta(\tau-1)}$ in which $\mathbf{P}_\theta$ stands for the cumulative distribution function of the Poisson model with parameter θ. This is equivalent to moment-matching and the solution $\hat{\theta}_\tau$ could be found numerically by performing a bisection search, for example. Similar parameter searches can be performed for the truncated negative binomial distribution that we consider in this section. In a more complex model where R_θ is a finite mixture of Poisson distributions, that is when $R_\theta(x) = \sum_{j=1}^{J} \pi_j R_{\theta_j}(x)$ for all $x \geq 0$, we can derive an EM algorithm for the truncated MLE similarly to the classical Poisson mixture. We do not elaborate this further, except to mention that each EM iteration entails the solution of fixed point equations, as in eq. (18), for each Poisson component.

Design To investigate the performance of the new estimator and compare it to other existing estimators, we conducted a set of experiments with synthetic data.

In the first set of these experiments, we take the abundances of rare species to be distributed according to a single Poisson distribution with parameter θ and the nuisance distribution (of abundant species) is the uniform distribution on $\tau^*, \ldots, \tau_{max}$. The resulting distribution has density $qR_\theta(x) + (1-q)U(x)$ with $0 < q < 1$ and U the aforementioned uniform distribution. Now, for any fixed $N \in \{200, 1000, 5000, 10000\}$, we generate a sample of size N from the Bernoulli model with parameter $q \in \{0.4, 0.6, 0.8\}$, then generate the corresponding counts observations according to the Poisson or uniform model. The parameters τ^* and τ_{max} are fixed equal 10 and 40 respectively whereas θ ranges over $\{0.6, 1, 1.5\}$. The observed zero-truncated counts are used to compute our new estimator $\hat{N}_{\hat{\tau}}$ and some other existing estimators with which $\hat{N}_{\hat{\tau}}$ will be compared.

To show that the results extend to other parametric families, we perform a second set of experiments, where we take the abundance of rare species to be distributed according to a Gamma-Poisson mixture, which leads naturally to the negative binomial distribution. In particular, in this case θ is two-dimensional, consisting of real parameters $r > 0$ and $s > 0$, and in eq. (1) the distribution ν_θ is the Gamma distribution with parameter $\theta = (r, s)$. This results in R_θ being the negative binomial distribution with parameters r and $p = 1/(1+s)$. We fix $p = 0.8$ and take r to vary over the range $\{0.5, 1, 2\}$. Larger values of N are needed to learn this model even in the absence of nonparametric noise. We consider the range of $N \in \{10000, 20000, 50000\}$, and we generate a sample of size N from the Bernoulli model with parameter $q \in \{0.4, 0.6, 0.8\}$, then generate the corresponding counts observations according to the negative binomial or uniform model. That is, the observational model is as before, $qR_\theta(x) + (1-q)U(x)$.

Risk approximation using G-L method We use the simulations as an opportunity to illustrate the G-L method and show the quality of the risk estimation by the proposed proxy in the selection rule. As displayed in Figure 1, the proxy $\widehat{\text{bias}}_\tau + \widehat{\text{var}}_\tau$ provides a good approximation of the true risk when $\widehat{\text{var}}_\tau$ is estimated by a bootstrap procedure as in eq. (15). Note that in this numerical example we calculate the risk, bias proxy, and variance proxy for N instead of $P(0)$. The approximation is remarkably accurate especially in the region where the estimator $\hat{N}_\tau$ is asymptotically unbiased (that is for $\tau \leq \tau^*$). Its remains satisfactory, but not overly so, for some τ greater than τ^*. This indicates that the bootstrap procedure is a good choice to estimate $\widehat{\text{var}}_\tau$. Note that Figure 1 corresponds to the results of simulations of the single Poisson model with parameters $q = 0.6$,$\theta = 1$, and $N = 1000$. We obtain similar results for all other parameter choices.

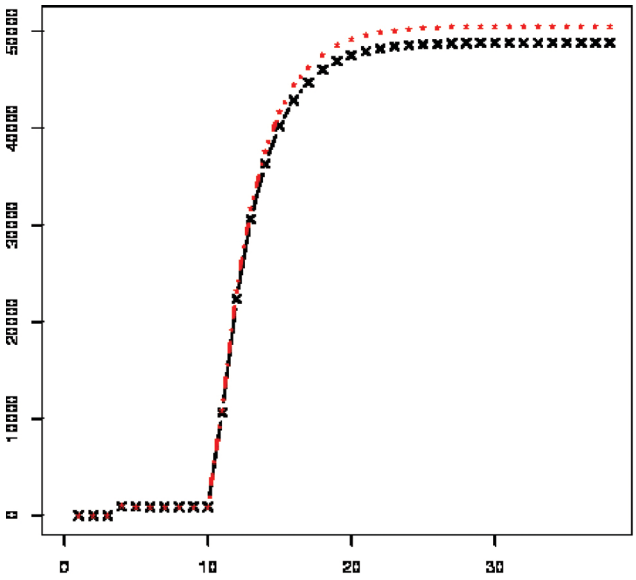

Figure 1: Estimated risk of the estimator of N as a function of τ (Lines X, in black) and its proxy $\widehat{\text{bias}}_\tau + \widehat{\text{var}}_\tau$ (dotdash $*$, in red) from the G-L method.

Performances of $\hat{N}_{\hat{\tau}}$ We focus on the performance of $\hat{N}_{\hat{\tau}}$ by calculating its Monte-Carlo mean (denoted "Mean" in the tables of results) and the renormalized standard error ($\frac{S_e}{N}$) based on 1000 samples. We also investigate the bootstrap-based confidence interval for N by providing the estimated non-coverage probabilities

$$\text{Inf} = \frac{1}{1000}\sum_{j=1}^{1000} \mathbf{1}_{[N<N_{\text{inf}}^{(j)}]}$$

and

$$\text{Sup} = \frac{1}{1000}\sum_{j=1}^{1000} \mathbf{1}_{[N>N_{\text{sup}}^{(j)}]},$$

where $I^{(j)} = [N_{\text{inf}}^{(j)}, N_{\text{sup}}^{(j)}]$ is the 95 per cent bootstrap-based confidence interval using the estimated model from the j^{th} Monte-Carlo sample. For the single Poisson model, the results are summarized in Table 1. It is clear that the renormalized S_e decreases when θ grows and increases as q becomes larger. As the small values of θ characterize small abundances and as a high value of q means that there is a large number of rare species in community (according to the simulated model), the observed variation of S_e suggests that a high number of rare species will be estimated with larger variance. We can also notice that S_e decreases with N in all simulated configurations showing the accuracy of the method when N becomes larger. As the large values of N describe the asymptotic regime of the estimators $\hat{\theta}_\tau$ and $\hat{q}_\tau$, we believe that the observed accuracy is related to the asymptotic efficiency of those estimators which improves the variance and then the mean square error (MSE) of $\hat{N}_{\hat{\tau}}$ as will be seen later. Table 2 summarizes the results for the Gamma-Poisson mixture model, with very comparable observations. Note that both sets of experiments show that we cannot rely on bootstrap confidence intervals as true intervals for the estimator. While the bootstrap is adequate in estimating the variability of the estimator, it does not accurately convey its location. It exhibits a clear skew to smaller values, which could be explained by the fact that resampling from the base distribution reduces the number of distinct observations. Therefore more principled methods are needed to go beyond point estimates in species richness estimation. One such avenue is through the use of concentration inequalities, [21].

Table 1: Performance of $\hat{N}_{\hat{\tau}}$ for single Poisson distributions. Inf and Sup are given in percentage (%).

		$\theta = 0.6$				$\theta = 1$				$\theta = 1.5$			
q	N	**Mean**	$\frac{Se}{N}$	**Inf (%)**	**Sup (%)**	**Mean**	$\frac{Se}{N}$	**Inf**	**Sup**	**Mean**	$\frac{Se}{N}$	**Inf**	**Sup**
0.4	200	192	0.116	1.5	26.3	200	0.058	2.5	7.2	199	0.036	2.2	11.7
	1000	1005	0.043	2.9	3.5	1001	0.024	3.6	4.6	1000	0.014	3.1	4.2
	5000	5003	0.018	3.0	3.4	4999	0.011	3.0	6.6	5001	0.006	3.3	3.7
	10000	10002	0.013	3.5	4.4	10002	0.007	3.3	4.3	10002	0.005	3.4	4.6
0.6	200	199	0.133	1.8	11.7	199	0.073	3.1	9.1	198	0.042	2.0	12.7
	1000	1003	0.055	3.3	5.0	1001	0.030	3.9	4.1	1000	0.017	2.9	2.7
	5000	5003	0.023	4.1	3.5	5001	0.013	3.5	2.8	5000	0.008	2.7	4.3
	10000	10009	0.017	4.3	3.7	10005	0.009	4.0	3.9	9999	0.006	3.3	4.0
0.8	200	192	0.160	2.5	15.5	195	0.079	1.5	13.3	196	0.048	1.1	17.0
	1000	1005	0.063	4.2	5.0	1002	0.034	3.7	4.7	999	0.021	3.5	6.5
	5000	5017	0.027	5.2	3.6	5000	0.015	3.9	4.0	4997	0.009	3.1	4.1
	10000	10001	0.019	2.9	4.6	9999	0.011	3.3	4.6	9998	0.006	3.2	4.4

Table 2: Performance of $\hat{N}_{\hat{\tau}}$ for Gamma-Poisson mixtures ($p = 0.8$). Inf and Sup are given in percentage (%).

		$r = 0.5$				$r = 1$				$r = 2$			
q	N	**Mean**	$\frac{Se}{N}$	**Inf**	**Sup**	**Mean**	$\frac{Se}{N}$	**Inf**	**Sup**	**Mean**	$\frac{Se}{N}$	**Inf**	**Sup**
0.4	10,000	9,854	0.038	1.3	10.4	9,955	0.012	2.2	14.6	9,998	0.003	2.9	9.0
	20,000	19,413	0.020	0.0	24.0	19,867	0.008	1.1	25.0	19,981	0.002	1.3	15.5
	50,000	48,359	0.013	0.0	64.0	49,561	0.005	0.0	50.0	49,933	0.001	0.0	39.0
0.6	10,000	9,618	0.042	0.4	25.0	9,883	0.015	0.8	16.3	9,986	0.003	1.3	13.3
	20,000	19,222	0.035	0.4	35.4	19,823	0.011	1.1	27.0	19,964	0.002	1.3	27.9
	50,000	47,792	0.018	0.0	71.0	49,319	0.005	0.0	72.0	49,885	0.002	0.4	53.2

0.8	10,000	9,561	0.053	0.7	23.1	9,843	0.016	0.3	27.0	9,973	0.004	0.7	23.3
	20,000	18,770	0.031	0.0	49.0	19,623	0.011	0.1	50.7	19,968	0.003	3.2	21.4
	50,000	46,812	0.019	0.0	86.0	49,128	0.006	0.0	76.0	49,816	0.002	0.0	80.0

Comparison with other estimators We end the simulations by comparing the proposed estimator of the number of species to other existing one in literature. We focus entirely on the single Poisson model, which represents the ground truth assumption of many of theses estimators. We consider Chao's estimator $\hat{N}_{Ch_0}$ defined as lower bound for N and proposed in [16], the coverage based estimator $(\hat{N}_{CL})$ proposed in [6] by Chao and Lee, the estimator $\hat{N}_{CB}$ of N using the expected proportion of duplicate species in the sample (by Chao and Bunge in [4]), the nonparametric MLE $\hat{N}_{WL_0}$ of N using a penalized likelihood (by Wang and Lindsay in [10]) and $\hat{N}_{LB}$: an extension of Chao's estimator proposed by Lanutheang and Böhning in [22]. The criteria used for this comparison (Mean, rMAE: relative Mean Absolute Error and rMSE: relative Mean Square Error) are computed and presented in Table 3. The six estimators display a good performance in all simulated configurations and $\hat{N}_{\hat{\tau}}$ seems to better estimate N than all other methods. This is quantified in Table 3, by the remarkably small value of *rMSE* as compared to the others. This shows that, despite our results being about the asymptotic efficiency of $\hat{\theta}_\tau$ and $\hat{q}_\tau$, we can expect finite-sample improvements for the estimator $\hat{N}_{\hat{\tau}}$, when N is moderately large. Also note that all six estimators become less reliable for very small value of θ or large value of q explaining thus the common difficulty for these approaches to better approximate N in the case of a large number of rares species, which touches upon the inherent problems of unidentifiability [14].

Table 3: Comparison of $\hat{N}_{\hat{\tau}}$ with five other estimators of N using 1000 monte-carlo samples. $\hat{N}_{Ch_0}$: Chao's estimator as lower bound on N in [16]; $\hat{N}_{CL}$: The coverage based estimator of N by Chao and Lee in [6]; $\hat{N}_{CB}$: Estimator of N using the expected proportion of duplicate species in the sample (by Chao and Bunge in [4]); $\hat{N}_{WL_0}$ Nonparametric MLE of N using a penalized likelihood (by Wang and Lindsay in [10]) and $\hat{N}_{LB}$ is an extension of Chao's estimator proposed by Lanutheang and Böhning in [22].

		$\theta = 0.6$			$\theta = 1$			$\theta = 1.5$		
q	**Est**	**Mean**	**rMAE**	**rMSE**	**Mean**	**rMAE**	**rMSE**	**Mean**	**rMAE**	**rMSE**
0.4	$\hat{N}_{\hat{\tau}}$	1005	0.034	0.185	1001	0.019	0.058	1000	0.011	0.019
	$\hat{N}_{Ch_0}$	1010	0.045	0.341	1002	0.026	0.108	1001	0.016	0.041
	$\hat{N}_{CL}$	1015	0.040	0.286	1007	0.023	0.084	1004	0.013	0.029
	$\hat{N}_{CB}$	1054	0.132	23.651	1004	0.035	0.227	1002	0.018	0.051
	$\hat{N}_{WL_0}$	1041	0.058	0.731	1024	0.035	0.292	1017	0.023	0.146
	$\hat{N}_{LB}$	1026	0.092	1.717	1022	0.046	0.482	1014	0.028	0.162
0.6	$\hat{N}_{\hat{\tau}}$	1003	0.043	0.298	1001	0.024	0.088	1000	0.014	0.029
	$\hat{N}_{Ch_0}$	1007	0.056	0.522	1003	0.031	0.160	1002	0.020	0.065
	$\hat{N}_{CL}$	1015	0.051	0.434	1008	0.027	0.125	1005	0.017	0.045
	$\hat{N}_{CB}$	1037	0.119	3.956	1005	0.043	0.315	1002	0.022	0.080
	$\hat{N}_{WL_0}$	1044	0.072	1.122	1034	0.047	0.510	1025	0.032	0.288
	$\hat{N}_{LB}$	1045	0.113	2.789	1031	0.057	0.704	1018	0.034	0.250
0.8	$\hat{N}_{\hat{\tau}}$	1005	0.051	0.401	1002	0.027	0.118	999	0.017	0.045
	$\hat{N}_{Ch_0}$	1009	0.062	0.621	1006	0.037	0.218	1003	0.023	0.088
	$\hat{N}_{CL}$	1020	0.058	0.553	1011	0.032	0.169	1006	0.020	0.065
	$\hat{N}_{CB}$	1038	0.128	3.719	1007	0.051	0.433	1003	0.026	0.111
	$\hat{N}_{WL_0}$	1062	0.088	1.550	1046	0.060	0.835	1031	0.040	0.405
	$\hat{N}_{LB}$	1059	0.126	3.452	1041	0.069	1.054	1019	0.038	0.294

4.2 Observational richness in text data

Rather than estimating the absolute number of species, an important extension of the species richness problem is concerned with estimating the number of distinct species to be observed in a sample larger than the current sample of individuals. Indeed, the abundance data $X_1^+, \cdots, X_D^+$ are ostensibly obtained by performing a sampling of individuals. If the said sample is enlarged, then how do the new abundances relate to the original ones? In the words of [23], in a pure Poisson abundance model: "Obviously, [the parameter λ] will be proportional to the size of the sample taken [...]". This is most easily seen in the individual sampling model of eq. (3): when the binomial size parameter is changed from m to $m' = \gamma m$, the parameters of the corresponding Poisson mixture are changed from $\lambda = mp_j$ to $\lambda' = m'p_j = \gamma\lambda$.

Generally in a Poisson mixture model, therefore, a γ factor increase in the sample size is equivalent to a γ dilation of the mixture distribution. Let $\mathbf{E}^{\gamma}[D]$ denote the expected number of distinct symbols in the enlarged sample, and thus $\mathbf{E}^{1}[D] = \mathbf{E}[D] = N(1 - qR_{\theta}(0))$. The observational richness estimation problem can thus be concretely stated as the problem of estimating $\mathbf{E}^{\gamma}[D]$, based on $X_1^+, \cdots, X_D^+$.

One application of the observational richness problem is to forecast the vocabulary of an author, from a portion of their text. This was popularized in the work of [24], who applied this methodology to the complete works of William Shakespeare. The problem goes back to the work of [25], who approached it from an empirical Bayesian perspective, without any specific parametrization. The earlier work of [23] also implicitly addressed the same problem.

Here, we restrict ourselves to the context of a parametric Poisson mixture abundance model for R_{θ}, that is as in eq. (1), with $\nu = \nu_{\theta}$ appropriately parametrized by θ. We require the family of such densities ν_{θ} to be closed under dilation, which means that for all θ and $\gamma > 0$, there exists a density $\nu_{\theta^{\gamma}}$ equal to the dilation of ν_{θ} by a factor γ, that is there exists a parameter θ^{γ} such that, for all measurable subsets A, $\nu_{\theta^{\gamma}}(A) = \nu_{\theta}(A/\gamma)$. Furthermore, we assume that for fixed γ, the transformation $\theta \mapsto \theta^{\gamma}$ is continuous in the sense that if a sequence $\theta_i \to \theta$ then the sequence $\theta_i^{\gamma} \to \theta^{\gamma}$. Note that for discrete mixtures, the scaling simply shifts the supports by γ, and for continuous mixtures it expands and scales the density by γ, and the requirement in either case is for the resulting density to remain an element of the parametric family.

As we focus primarily on text data, the Gamma-Poisson mixture family is very well-suited. Recall that in this case θ is two-dimensional, consisting of real parameters $r > 0$ and $s > 0$, and ν_{θ} is the Gamma distribution with parameter $\theta = (r, s)$. Then, the distribution $f_{\nu_{\theta}}$ is the negative binomial distribution with parameters r and $p = 1/(1+s)$. To dilate the Gamma distribution, it is easy to see that one simply scales $s' = \gamma s$. This corresponds to a transformation of the negative binomial parameter $p' = 1/(1 + \gamma(1-p)/p)$.

This paper's framework applies to this problem as follows. If the rare abundances are well modeled by a Gamma-Poisson mixture while the abundant ones are not, then our framework allows us to efficiently learn the parameters q and θ. By continuity, for fixed γ we also have an efficient estimator of θ^{γ}. Since N is assumed to stay constant, we then have

$$N = \frac{\mathbf{E}[D]}{1 - qR_{\theta}(0)} = \frac{\mathbf{E}^{\gamma}[D]}{1 - qR_{\theta^{\gamma}}(0)}.$$

We could therefore use our estimates $\hat{q}_{\tau}$ and $\hat{\theta}_{\tau}$ to evaluate $\hat{\theta}_{\tau}^{\gamma}$ and thus to estimate $\mathbf{E}^{\gamma}[D]$ as follows:

$$\widehat{\mathbf{E}^{\gamma}[D]}_{\tau} = \frac{1 - \hat{q}_{\tau} R_{\hat{\theta}_{\tau}^{\gamma}}(0)}{1 - \hat{q}_{\tau} R_{\hat{\theta}_{\tau}}(0)} D.$$

The data we look at is French playwright Molière's *Tartuffe* play, which we gradually observe a portion of and try to estimate the number of distinct vocabulary words. Thus, the scale γ is the ratio of the total text size to the size of the observed text, varying from 0 to 100%. For this problem, we illustrate $\widehat{\mathbf{E}^{\gamma}[D]}_{\tau}$ for various choices of τ and also the G-L selected $\hat{\tau}$ in Figure 2. Note how quickly the result becomes an accurate estimate of the vocabulary. But most importantly, note how sub-optimal choices of the truncation can adversely affect the performance of the estimator.

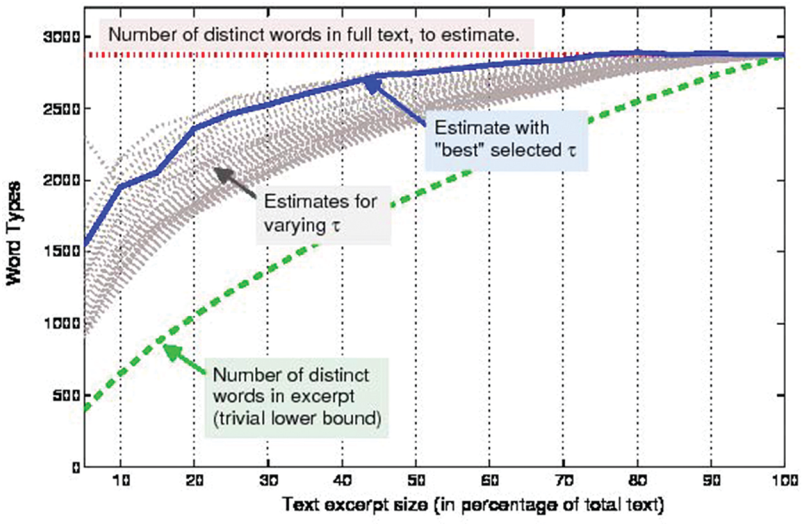

Figure 2: Estimating the vocabulary growth in Molière's *Tartuffe* play.

5 Conclusion

In this paper, we revisited the species richness estimation problem and studied a commonly followed practice of truncating the data into rare and abundant species. We proposed a semiparametric framework to model such a truncation as a parametric component well-suited to model rare species and a nonparametric nuisance component to cover the abundant species in an agnostic manner. We showed that asymptotic efficiency in this framework requires handling the truncation more delicately. This is in particular true if the rare species model has a significant overlap with the abundant species. Finally, we proposed a heuristic method to learn a good truncation threshold from data.

Several possible avenues of investigation may be proposed. We already mentioned the importance of going beyond point estimates. One would also like to relax some assumptions about rare or/and abundant species. Concerning rare species, the parametric assumption is not fully satisfying and it would be interesting to know whether semiparametric efficiency can be obtained in a larger context, or if our estimator has some robustness properties. Concerning abundant species, it is not clear whether the assumption that they are truly located entirely away from zero is needed. In particular, it is important to handle the situation when such a dichotomy arises from an underlying binomial mixture model. Some recent approaches to species richness have successfully used Chebyshev polynomials as a fitting model, see for example [26], and one would like to understand the relationship between such fits and mixtures of Poissons. Finally, one would hope that a truncation threshold that automatically conforms to the underlying model could make the most of the available data and thus give a fundamental theoretical edge, perhaps in the form of adaptive rates.

Appendix

A Proofs

A.1 Proof of Proposition 1

For any fixed τ, the classical conditional MLE satisfies

$$\hat{N}_{\text{classical}} = D + \frac{D_\tau R_{\hat{\theta}_\tau}(0)}{1 - R_{\hat{\theta}_\tau}(0)}.$$

Let us consider now the new conditional MLE proposed in this work:

$$\begin{aligned}\hat{N}_\tau =& \frac{D}{1 - \hat{q}_\tau R_{\hat{\theta}_\tau}(0)} \\ =& D + \frac{D\hat{q}_\tau R_{\hat{\theta}_\tau}(0)}{1 - \hat{q}_\tau R_{\hat{\theta}_\tau}(0)}.\end{aligned}$$

But using eq. (11), we have

$$\frac{\hat{q}_\tau R_{\hat{\theta}_\tau}(0)}{1 - \hat{q}_\tau R_{\hat{\theta}_\tau}(0)} = \frac{D_\tau R_{\hat{\theta}_\tau}(0)}{D\sum_{k=1}^{\tau} R_{\hat{\theta}_\tau}(k)}$$

from which we get

$$\hat{N}_\tau = D + \frac{D_\tau R_{\hat{\theta}_\tau}(0)}{\sum_{k=1}^{\tau} R_{\hat{\theta}_\tau}(k)}. \tag{19}$$

Now, if R_θ is supported on $\{0, \dots, \tau\}$, then $\sum_{k=1}^{\tau} R_{\hat{\theta}_\tau}(k)$ equals $1 - R_{\hat{\theta}_\tau}(0)$ in the last expression which finally gives $\hat{N}_\tau = \hat{N}_{\text{classical}}$. □

A.2 Proof of Proposition 2

For τ equals 2 and R_θ corresponding to the Poisson distribution with parameter θ, it is not difficult to see that $\hat{\theta}_\tau = \hat{\theta}_{\text{Zelterman}} = 2n_2/n_1$. Then,

$$\begin{aligned}\hat{N}_\tau =& D + \frac{D_\tau R_{\hat{\theta}_\tau}(0)}{\sum_{k=1}^{\tau} R_{\hat{\theta}_\tau}(k)} \\ =& D + \frac{D_\tau}{\hat{\theta}_\tau + \hat{\theta}_\tau^2/2}\end{aligned}$$

and using the fact that $D_\tau = n_1 + n_2$, we get $\hat{N}_\tau = D + n_1^2/2n_2$. □

A.3 Proof of Theorem 1

We first prove (i), and fix $\tau \leq \tau_0$. We use the fact that $\hat{\theta}$ is the maximum likelihood estimator in the model with density S_θ^τ. Recall that $\sum_{x=1}^{\tau} \frac{n_x}{D_\tau} = 1$. Note that maximizing the likelihood of eq. (13) amounts to maximizing in θ the criterion $\mathscr{L}_D(\theta) = \sum_{x=1}^{\tau} \frac{n_x}{D_\tau} \log S_\theta^\tau(x)$ which as N tends to infinity converges almost surely to $\mathscr{L}(\theta) = \sum_{x=1}^{\tau} S_{\theta_0}^\tau(x) \log S_\theta^\tau(x)$ when $\tau \leq \tau_0$. Moreover, we have

$$|\mathscr{L}_D(\theta) - \mathscr{L}(\theta)| \leq \sum_{x=1}^{\tau} \left| \frac{n_x}{D_\tau} - S_{\theta_0}^\tau(x) \right| \, |\log S_\theta^\tau(x)|.$$

On the right hand side of this inequality, $\frac{n_x}{D_\tau} - S_{\theta_0}^\tau(x)$ converges almost surely, thus in probability, to zero. Also, $|\log S_\theta^\tau(x)|$ is bounded since $R_\theta(x) \geq \delta > 0$, for all $\theta \in \Theta$ and all $x \leq \tau$. We conclude that

$$\sup_{\theta \in \Theta} |\mathscr{L}_D(\theta) - \mathscr{L}(\theta)| \to 0 \quad \text{in probability.}$$

It is easy to see that

$$\mathscr{L}(\theta) - \mathscr{L}(\theta_0) = \sum_{x=1}^{\tau} S_{\theta_0}^\tau(x) \log \frac{S_\theta^\tau(x)}{S_{\theta_0}^\tau(x)}$$

attains uniquely its maximum (equals zero) at θ_0 since the true model $S_{\theta_0}^\tau$ is identifiable, as assumed. We then obtain

$$\sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} \mathscr{L}(\theta) < \mathscr{L}(\theta_0), \tag{20}$$

where $d(\theta, \theta_0)$ is the Euclidean distance between θ and θ_0. As $\hat{\theta}$ maximizes $\mathscr{L}_D$, we have $\mathscr{L}_D(\hat{\theta}) \geq \mathscr{L}_D(\theta_0) - \mathrm{o}_{\mathbb{P}}(1)$. This, together with the condition in eq. (20) and the above convergence in probability, entails that $\hat{\theta}$ converges in probability to θ_0 as N tends to infinity. This result holds from Theorem 5.7 in [27].

To end part (i) of the theorem, recall that from eq. (11):

$$\hat{q}(\hat{\theta}) = \frac{1}{R_{\hat{\theta}}(0) + \frac{D}{D_\tau} \sum_{k=1}^{\tau} R_{\hat{\theta}}(k)}.$$

We then observe from the law of large numbers that as N tends to infinity, $\frac{D}{D_\tau} = \frac{D/N}{D_\tau/N}$ converges almost surely to $\frac{1 - q_0 R_{\theta_0}(0)}{q_0 \sum_{k=1}^{\tau} R_{\theta_0}(k)}$ when $\tau \leq \tau_0$. Recall that we assume R_θ to be continuous in θ for each x. Thus using the continuous map theorem and the convergence in probability of $\hat{\theta}$ to θ_0, we find that $R_{\hat{\theta}}(0)$ and $\sum_{k=1}^{\tau} R_{\hat{\theta}}(k)$ converge in probability to $R_{\theta_0}(0)$ and $\sum_{k=1}^{\tau} R_{\theta_0}(k)$ respectively when $\tau \leq \tau_0$. We finally obtain the convergence in probability of $\hat{q}(\hat{\theta})$ to

$$\hat{q}(\theta_0) = \frac{1}{R_{\theta_0}(0) + \left\{ \frac{1 - q_0 R_{\theta_0}(0)}{q_0 \sum_{k=1}^{\tau} R_{\theta_0}(k)} \right\} \sum_{k=1}^{\tau} R_{\theta_0}(k)} = q_0,$$

using once again the continuous map theorem. This ends the proof of part (i) of Theorem 1.

Similar arguments to what we have given here can be used to prove part (ii) of the Theorem, namely that if $\tau > \tau_0$, then $\hat{\theta}$ converges in probability to the set of maximizers of $M^\tau(\theta) = \sum_{x=1}^{\tau} f^+(x) \log S_\theta^\tau(x)$, in the sense that the probability of falling in an ϵ-dilation of this set tends to 1 as $N \to \infty$.

A.4 Efficient score functions and efficient fisher information

We now build up some notation. Denote by $\text{supp}(F)$ the support of F, that is the set of integers x such that $F(x) > 0$. In particular, $\text{supp}(F)$ contains only integers x such that $x > \tau$. Let $\mathcal{G}$ denote the set of measurable functions G defined on $\text{supp}(F)$ by

$$\mathcal{G} = \left\{ G : \text{supp}(F) \mapsto \mathbb{R} \ : \ \sum_{x \in \text{supp}(F)} F(x)G(x) = 0 \text{and} \sum_{x \in \text{supp}(F)} F(x)G^2(x) < \infty \right\}. \tag{21}$$

For a given G in $\mathcal{G}$, a real number a and a vector b of dimension k let us define $q_t = q + at, \theta_t = \theta + bt$ and $F_t = F(1 + tG)$. This parametrization of F, q and θ defines a path (a one-dimensional sub-model) $f_t^+ = f^+_{(q_t,\theta_t,F_t)}$ in the model $\mathscr{P}^+$. To simplify the notation, we let f^+ stand for $f^+_{(q,\theta,F)}$ and f for $f_{(q,\theta,F)}$. Recall the definition of score functions.

Definition 2.
A differentiable path is a map $t \mapsto f_t^+$ from a neighborhood $[0, \varepsilon)$ of 0 to $\mathscr{P}^+$ with $f_0^+ = f^+$ such that, for some measurable real valued function g, one has

$$\sum_{x \geq 1} \left(\frac{\sqrt{f_t^+(x)} - \sqrt{f^+(x)}}{t} - \frac{1}{2} g(x) \cdot \sqrt{f^+(x)} \right)^2 \to 0 \text{as } t \to 0. \tag{22}$$

The one-dimensional sub-model $\{f_t^+, t \in [0, \varepsilon)\}$ is then said to be differentiable in quadratic mean at f^+ with score functiong.

A more useful way to determine the score function of a model such as $\{f_t^+, t \in [0, \varepsilon)\}$ is to take the derivative with respect to t of the log-likelihood at $t = 0$, that is

$$g = \left. \frac{\mathrm{d}}{\mathrm{d}t} \right|_{t=0} \log f_t^+. \tag{23}$$

We will use a dot-notation to indicate differentiation with respect to a parameter. Recall first the parametric score function $\dot{\ell}_q$ and the parametric vector score function $\dot{\ell}_\theta$ which are the partial derivative and gradient with respect to q and θ respectively of the log-likelihood in the full model $\mathscr{P}^+$. We have respectively

$$\dot{\ell}_q = \frac{R_\theta - F}{f} + \frac{R_\theta(0)}{1 - qR_\theta(0)} \text{and } \dot{\ell}_\theta = \frac{q\dot{R}_\theta}{f} + \frac{q\dot{R}_\theta(0)}{1 - qR_\theta(0)}, \tag{24}$$

with $\dot{R}_\theta$ the gradient function of the density R_θ.

In the model defined in eq. (5), a straightforward calculation shows that the score function g of the one-dimensional sub-model is such that

$$g = a\dot{\ell}_q + \langle b, \dot{\ell}_\theta \rangle + \frac{(1-q)FG}{f}$$

where a and b are the scaling scalar and k-dimensional vector of the parametrizations q_t and θ_t respectively, and where $\langle \cdot, \cdot \rangle$ denotes the usual inner product.

Now, we recall briefly the notions of tangent set and efficient score function for the model considered here. The maximal tangent set to the model $\mathscr{P}^+$ at f^+ is the set of all score functions of a one-dimensional sub-model. We denote it $\dot{\mathscr{P}}^+$, and in our case it is given by

$$\dot{\mathscr{P}}^+ = \left\{ g = a\dot{\ell}_q + \langle b, \dot{\ell}_\theta \rangle + \frac{(1-q)FG}{f}; (a, b) \in \mathbb{R}^{k+1}, \text{and } G \in \mathcal{G} \right\}. \tag{25}$$

Consider again the path $t \mapsto f^+_{(q,\theta,F_t)}$ related to the model $\mathscr{P}^+$, but now with the parameters q and θ fixed. Then the tangent set at f^+ for the nonparametric part of the model in eq. (5) is denoted and given by

$$\dot{\mathscr{P}}_F^+ = \left\{ h = \frac{(1-q)FG}{f}, \ G \in \mathcal{G} \right\}. \tag{26}$$

The *efficient score function* related to a given component α of the parameter vector $(q, \theta_1, \cdots, \theta_k)$ is then defined component-wise as $\tilde{\ell}_\alpha = \dot{\ell}_\alpha - \Pi_F \dot{\ell}_\alpha$, where Π_F is the orthogonal projection onto the closure of the linear space spanned by $\dot{\mathscr{P}}_F^+$.

The expressions of the efficient score functions are given in the following proposition, the coefficients of the efficient Fisher information matrix are displayed in the proof of this proposition.

Proposition 3

The efficient score functions for estimating the parameters q and θ are given for $x \geq 1$ by

$$\tilde{\ell}_q(x) = \frac{1}{q}\mathbf{1}\{x \leq \tau\} - \frac{\sum_{k=0}^{\tau} R_\theta(k)}{1 - q\sum_{k=0}^{\tau} R_\theta(k)}\mathbf{1}\{x > \tau\} + \frac{R_\theta(0)}{1 - qR_\theta(0)} \tag{27}$$

and

$$\tilde{\ell}_\theta(x) = \frac{\dot{R}_\theta(x)}{R_\theta(x)}\mathbf{1}\{x \leq \tau\} - \frac{q\sum_{k=0}^{\tau} \dot{R}_\theta(k)}{1 - q\sum_{k=0}^{\tau} R_\theta(k)}\mathbf{1}\{x > \tau\} + \frac{q\dot{R}_\theta(0)}{1 - qR_\theta(0)} \tag{28}$$

respectively. The efficient Fisher information $\tilde{I}$ is a matrix of order $(k+1)$ with coefficients given by eqs. (33)–(35).

Proof Let $\overline{lin}(\dot{\mathscr{P}}_F^+)$ denote the closure of the linear space spanned by $\dot{\mathscr{P}}_F^+$ in $\mathbb{L}^2(f^+)$. To reduce clutter, let $\dot{\ell}$ refer to a particular component $\dot{\ell}_\alpha$. We first give a closed form expression of the orthogonal projection $\Pi_F\dot{\ell}$.

First observe that for every score function $\dot{\ell}$ in the model $\mathscr{P}^+$, the projection $\Pi_F\dot{\ell}$ is an element of the subspace $\overline{lin}(\dot{\mathscr{P}}_F^+)$ so that it must be a linear combination (or a limit thereof) of elements of the form $\frac{(1-q)FG}{f}$, $G \in \mathscr{G}$. Since the latter all vanish on the set $\{1, \ldots, \tau\}$, so does $\Pi_F\dot{\ell}$.

Next, let $\tilde{h}$ be any $\mathbb{L}^2(f^+)$-integrable function that is orthogonal to the space $\overline{lin}(\dot{\mathscr{P}}_F^+)$:

$$\sum_{x>\tau} \tilde{h}(x)f^+(x) = 0 \text{and} \sum_{x>\tau} \tilde{h}^2(x)f^+(x) < \infty.$$

In particular, note that such an $\tilde{h}$ is orthogonal to elements of $\dot{\mathscr{P}}_F^+$ itself. These, once again, have the form $\frac{(1-q)FG_0}{f}$ for some $G_0 \in \mathscr{G}$. By design, let us choose G_0 such that $G_0(x_1) = F(x_2)$ and $G_0(x_2) = -F(x_1)$ for x_1, x_2 in the support of F and $G_0(x) = 0$ elsewhere. It is easy to verify that such a choice does indeed lie within $\mathscr{G}$. On the other hand, the orthogonality of $\tilde{h}$ and $\frac{(1-q)FG_0}{f}$ in $\mathbb{L}^2(f^+)$ implies that:

$$\sum_{x\geq 1} \frac{F(x)G_0(x)}{f(x)}\tilde{h}(x)f^+(x) = 0,$$

or equivalently

$$\sum_{x\geq 1} F(x)G_0(x)\tilde{h}(x) = F(x_1)F(x_2)\left(\tilde{h}(x_1) - \tilde{h}(x_2)\right) = 0.$$

As $F(x)$ is strictly positive over its support, this implies that $\tilde{h}(x_1) - \tilde{h}(x_2) = 0$. Thus all such $\tilde{h}$ must be constant on the support of F.

Now let us specialize $\tilde{h}$ to the components of the efficient score function, by writing them as $\tilde{\ell} = \dot{\ell} - \Pi_F\dot{\ell}$. Since we have thus determined that $\Pi_F\dot{\ell}$ vanishes on $x \leq \tau$ and $\tilde{\ell}$ is constant over $x > \tau$, we have therefore established:

$$(\Pi_F\dot{\ell})(x) = \begin{cases} 0 \text{ if } 1 \leq x \leq \tau, \\ \dot{\ell}(x) - c(\dot{\ell}) \text{if } x > \tau, \end{cases} \tag{29}$$

where $c(\dot{\ell})$ is a constant depending on $\dot{\ell}$. To obtain the expression of $c(\dot{\ell})$, we can once again use the fact that $\Pi_F\dot{\ell}$ is a linear combination of $(1-q)FG/f$, $G \in \mathscr{G}$, or a limit thereof, in addition to the fact that $\sum_{x>\tau} F(x)G(x) = 0$ for all such G , to write:

$$\sum_{x>\tau} (\dot{\ell}(x) - c(\dot{\ell}))f(x) = 0.$$

We thus get

$$c(\dot{\ell}) = \frac{\sum_{x>\tau} \dot{\ell}(x)f(x)}{\sum_{x>\tau} f(x)}. \tag{30}$$

Now, we can easily compute the efficient score functions:

$$\tilde{\ell}(x) = \begin{cases} \dot{\ell}(x) \text{if } 1 \leq x \leq \tau, \\ c(\dot{\ell}) \text{if } x > \tau. \end{cases} \tag{31}$$

Using the expressions of $\dot{\ell}_q$ and $\dot{\ell}_\theta$ in eq. (24), we explicitly get $\tilde{\ell}_q$ and $\tilde{\ell}_\theta$. We start with $\tilde{\ell}_q$. We have:

$$\begin{aligned} \sum_{x>\tau} \dot{\ell}_q(x) f(x) =& \sum_{x>\tau} R_\theta(x) - 1 + \frac{R_\theta(0)}{1 - qR_\theta(0)} \sum_{x>\tau} f(x) \\ =& \frac{R_\theta(0)}{1 - qR_\theta(0)} \sum_{x>\tau} f(x) - \sum_{x=0}^{\tau} R_\theta(x). \end{aligned}$$

We then determine $c(\dot{\ell}_q)$ from eq. (30),

$$\begin{aligned} c(\dot{\ell}_q) =& \frac{R_\theta(0)}{1 - qR_\theta(0)} - \frac{\sum_{x=0}^{\tau} R_\theta(x)}{\sum_{x>\tau} f(x)} \\ =& \frac{R_\theta(0)}{1 - qR_\theta(0)} - \frac{\sum_{x=0}^{\tau} R_\theta(x)}{1 - q\sum_{x=0}^{\tau} R_\theta(x)} \end{aligned}$$

and since $\frac{R_\theta(x) - F(x)}{f(x)} = \frac{1}{q}$ for all $x \leq \tau$, we finally obtain $\tilde{\ell}_q(x)$ as

$$\tilde{\ell}_q(x) = \frac{1}{q}\mathbf{1}\{x \leq \tau\} - \frac{\sum_{x=0}^{\tau} R_\theta(x)}{1 - q\sum_{x=0}^{\tau} R_\theta(x)}\mathbf{1}\{x > \tau\} + \frac{R_\theta(0)}{1 - qR_\theta(0)}.$$

Moving on to $\dot{\ell}_\theta$, from eq. (24) and using the fact that $\sum_{x\in\mathbb{N}} \dot{R}_\theta(x) = 0$, we have:

$$\begin{aligned} \sum_{x>\tau} \dot{\ell}_\theta(x) f(x) =& q \sum_{x>\tau} \dot{R}_\theta(x) + \frac{q\dot{R}_\theta(0)}{1 - qR_\theta(0)} \sum_{x>\tau} f(x) \\ =& \frac{q\dot{R}_\theta(0)}{1 - qR_\theta(0)} \sum_{x>\tau} f(x) - q\sum_{x=0}^{\tau} \dot{R}_\theta(x). \end{aligned}$$

Then

$$c(\dot{\ell}_\theta) = \frac{q\dot{R}_\theta(0)}{1 - qR_\theta(0)} - \frac{q\sum_{x=0}^{\tau} \dot{R}_\theta(x)}{1 - q\sum_{x=0}^{\tau} R_\theta(x)}$$

and

$$\tilde{\ell}_\theta(x) = \frac{\dot{R}_\theta(x)}{R_\theta(x)}\mathbf{1}\{x \leq \tau\} - \frac{q\sum_{x=0}^{\tau} \dot{R}_\theta(x)}{1 - q\sum_{x=0}^{\tau} R_\theta(x)}\mathbf{1}\{x > \tau\} + \frac{q\dot{R}_\theta(0)}{1 - qR_\theta(0)}.$$

The efficient Fisher information matrix has coefficients defined as

$$\tilde{I}_q = \sum_{x\geq 1} \tilde{\ell}_q^2(x) f^+(x), \tilde{I}_{q\theta_j} = \sum_{x\geq 1} \tilde{\ell}_q(x)\tilde{\ell}_{\theta_j}(x) f^+(x) \text{and } \tilde{I}_{\theta_i\theta_j} = \sum_{x\geq 1} \tilde{\ell}_{\theta_i}(x)\tilde{\ell}_{\theta_j}(x) f^+(x) \tag{32}$$

for all $i, j = 1, \ldots, k$. Recall that when we write $\tilde{\ell}_\theta$, we are referring to a vector of score functions, whereas $\tilde{\ell}_{\theta_j}$ stands for the j^{th} coordinate of $\tilde{\ell}_\theta$. The computation of these coefficients leads to

$$\tilde{I}_q = \frac{1}{1 - qR_\theta(0)} \left\{ \frac{1}{q} \sum_{x=1}^{\tau} R_\theta(x) + \frac{[\sum_{x=0}^{\tau} R_\theta(x)]^2}{1 - q\sum_{x=0}^{\tau} R_\theta(x)} - \frac{[R_\theta(0)]^2}{1 - qR_\theta(0)} \right\} \tag{33}$$

$$\tilde{I}_{\theta_i\theta_j} = \frac{q}{1 - qR_\theta(0)} \left\{ \sum_{x=1}^{\tau} \frac{[\dot{R}_\theta^i(x)][\dot{R}_\theta^j(x)]}{R_\theta(x)} + \frac{q[\sum_{x=0}^{\tau} \dot{R}_\theta^i(x)][\sum_{x=0}^{\tau} \dot{R}_\theta^j(x)]}{1 - q\sum_{x=0}^{\tau} R_\theta(x)} - \frac{q[\dot{R}_\theta^i(0)][\dot{R}_\theta^j(0)]}{1 - qR_\theta(0)} \right\} \tag{34}$$

$$\tilde{I}_{q\theta_j} = \frac{q}{1 - qR_\theta(0)} \left\{ \frac{1}{q} \sum_{x=1}^{\tau} \dot{R}_\theta^j(x) + \frac{[\sum_{x=0}^{\tau} R_\theta(x)][\sum_{x=0}^{\tau} \dot{R}_\theta^j(x)]}{1 - q\sum_{x=0}^{\tau} R_\theta(x)} - \frac{\dot{R}_\theta^j(0) R_\theta(0)}{1 - qR_\theta(0)} \right\} \tag{35}$$

with $\dot{R}_\theta^j$ the partial derivative of R_θ with respect to the j^{th} coordinate of θ.

A.5 Proof of Theorem 2

We first state and prove two lemmas that will be used for the proof of Theorem 2.

As usual, let α be a component of the parameters vector (q, θ), denote by α_0 the true value of α (if it exists), and let $\mathscr{V}(\alpha_0)$ be a closed neighborhood of α_0. We denote by $\mathscr{H}_\alpha$ the subset of $\mathbb{L}^2(f^+)$ defined by

$$\mathscr{H}_\alpha = \left\{\tilde{\ell}_\alpha, \text{with } \alpha \in \mathscr{V}(\alpha_0)\right\}. \tag{36}$$

Lemma 1

Let $\hat{\alpha}$ be a consistent estimator of α_0. If $\theta \mapsto R_\theta(x)$ is twice continuously differentiable for every $x \leq \tau$ and Assumptions 1–3 hold, then $\mathscr{H}_\alpha$ is a Donsker class with square integrable envelope that contains $\tilde{\ell}_{\hat{\alpha}}$ with probability that tends to one.

Proof We adapt the method used in Example 19.7 from [27]. Recall that a δ-bracket is a subset $[u_1, u_2]$ of $\mathbb{L}^2(f^+)$ such that $\|u_2 - u_1\|_{\mathbb{L}^2(f^+)} < \delta$. The bracketing number $N\big(\delta, \mathscr{H}_\alpha, \mathbb{L}^2(f^+)\big)$ is the minimum number of δ-brackets needed to cover $\mathscr{H}_\alpha$ and the bracketing entropy is the logarithm of this quantity. To show that $\mathscr{H}_\alpha$ is Donsker, we establish the sufficient condition that the square root entropy integral

$$\int_0^1 \sqrt{\log N\left(\gamma, \mathscr{H}_\alpha, \mathbb{L}^2(f^+)\right)}\, d\gamma \tag{37}$$

is finite. (See Theorem 19.5 in [27].)

We begin by establishing continuity properties of the parametric efficient score functions. From the differentiability of R_θ in θ and the expressions given in Proposition 3, it is evident that $\tilde{\ell}_\alpha$ is always a differentiable function of α. Let us denote these derivatives by $\dot{\tilde{\ell}}_\alpha$. For $\alpha = q$ and $\alpha = \theta_j$ we can respectively compute these as

$$\dot{\tilde{\ell}}_q = \left\{\left(\frac{R_\theta(0)}{1 - qR_\theta(0)}\right)^2 - \frac{1}{q^2}\right\}\mathbf{1}\{x \leq \tau\} + \left\{\left(\frac{R_\theta(0)}{1 - qR_\theta(0)}\right)^2 - \left(\frac{\sum_{k=0}^{\tau} R_\theta(k)}{1 - q\sum_{k=0}^{\tau} R_\theta(k)}\right)^2\right\}\mathbf{1}\{x > \tau\}$$

and

$$\dot{\tilde{\ell}}_{\theta_j} = \left\{\frac{\ddot{R}^j_\theta(x)}{R_\theta(x)} - \left(\frac{\dot{R}^j_\theta(x)}{R_\theta(x)}\right)^2 + \frac{q\ddot{R}^j_\theta(0)}{1 - qR_\theta(0)} + \left(\frac{q\dot{R}^j_\theta(0)}{1 - qR_\theta(0)}\right)^2\right\}\mathbf{1}\{x \leq \tau\} +$$

$$\left\{\frac{q\ddot{R}^j_\theta(0)}{1 - qR_\theta(0)} + \left(\frac{q\dot{R}^j_\theta(0)}{1 - qR_\theta(0)}\right)^2 - \frac{q\sum_{k=0}^{\tau}\ddot{R}^j_\theta(k)}{1 - q\sum_{k=0}^{\tau} R_\theta(k)} - \left(\frac{q\sum_{k=0}^{\tau}\dot{R}^j_\theta(k)}{1 - q\sum_{k=0}^{\tau} R_\theta(k)}\right)^2\right\}\mathbf{1}\{x > \tau\}.$$

By inspection, we find that $\dot{\tilde{\ell}}_q$ is always continuous itself, and that $\dot{\tilde{\ell}}_{\theta_j}$ is also continuous provided that $\theta \mapsto \mathbb{R}_\theta$ is in $\mathscr{C}^2$ and $R_\theta(x) \geq \eta > 0$ for all $x \leq \tau$, as assumed. These conditions also imply that $\dot{\tilde{\ell}}_\alpha$ have a finite $\mathbb{L}^2(f^+)$-norm and that these functions are Lipschitz-continuous on $\mathscr{V}(\alpha_0)$. We thus have a non-negative bounded V such that

$$|\tilde{\ell}_{\alpha_2}(x) - \tilde{\ell}_{\alpha_1}(x)| \leq V|\alpha_2 - \alpha_1| \text{for every } \alpha_1, \alpha_2 \in \mathscr{V}(\alpha_0). \tag{38}$$

Now, from this Lipschitz condition, it follows that if $|\alpha - \alpha_1| < \epsilon$ then $\tilde{\ell}_{\alpha_1} - \epsilon V \leq \tilde{\ell}_\alpha \leq \tilde{\ell}_{\alpha_1} + \epsilon V$. This means that we need as many ϵ-balls (a ball with radius $\epsilon/2$) to cover $\mathscr{V}(\alpha_0)$ as we need δ-brackets ($\delta = 2\epsilon V$) to cover $\mathscr{H}_\alpha$. Since the number n_0 of ϵ-balls needed to cover $\mathscr{V}(\alpha_0)$ is such that

$$n_0 \leq C\frac{\text{diam}[v(\alpha_0)]}{\epsilon} \vee 1,$$

with C a constant depending only on $\mathscr{V}(\alpha_0)$, it follows that the bracketing number is

$$N\big(\delta, \mathscr{H}_\alpha, \mathbb{L}^2(f^+)\big) \leq \left\{C\text{diam}[v(\alpha_0)]\frac{2V}{\delta}\right\} \vee 1. \tag{39}$$

Thus the bracketing entropy is of order smaller than $\log(1/\delta)$, whose square root is integrable near 0. This establishes the sufficient condition of eq. (37), and thus $\mathscr{H}_\alpha$ is indeed Donsker.

To complete the other claims of the proof, note that for all α in $\mathscr{V}(\alpha_0)$,$\tilde{\ell}_\alpha$ has a finite $\mathbb{L}^2(f^+)$-norm and that $|\tilde{\ell}_\alpha(x)| \leq U$ for some $U < \infty$, for all $x \geq 1$. The boundedness of $\tilde{\ell}_\alpha$ is obtained from the expression of $\tilde{\ell}_q$ and $\tilde{\ell}_{\theta_j}$.

The constant function U is a square integrable envelope for $\mathscr{H}_\alpha$. We use the continuity of the map $\alpha \mapsto \tilde{\ell}_\alpha(x)$ and consistency of $\hat{\alpha}$ to show that $\lim_{N\to\infty} \mathbb{P}[|\tilde{\ell}_{\hat{\alpha}}(x) - \tilde{\ell}_{\alpha_0}(x)| > \epsilon] = 0$ for all $x \geq 1$. This proves that $\mathscr{H}_\alpha$ contains $\tilde{\ell}_{\hat{\alpha}}$ with probability that tends to one and the lemma holds.

The result in Lemma 1 holds for $\mathscr{H}_q$ and $\mathscr{H}_{\theta_j}$ for all $j = 1, \ldots, k$ and thus also for their union $\mathscr{H}$. We conclude that $\mathscr{H}$ is a Donsker class with square integrable envelope that contains $(\tilde{\ell}_{\hat{q}}, \tilde{\ell}_{\hat{\theta}})$ with probability that tends to one.

Lemma 2

$\hat{\theta}_\tau$ and $\hat{q}_\tau$ solve the efficient score equations:

$$\sum_{i=1}^{D} \tilde{\ell}_q(x_i) = 0, \text{and } \sum_{i=1}^{D} \tilde{\ell}_\theta(x_i) = 0. \tag{40}$$

Proof Note that the efficient score equation $\sum_{i=1}^{D} \tilde{\ell}_q(x_i) = 0$ can be written as

$$\frac{D_\tau}{q} + \frac{DR_\theta(0)}{1 - qR_\theta(0)} - \frac{(D - D_\tau)\sum_{k=0}^{\tau} R_\theta(k)}{1 - q\sum_{k=0}^{\tau} R_\theta(k)} = 0. \tag{41}$$

The zero notation here refers to the null vector of $\mathbb{R}^k$.

Recalling that $\hat{q}_\tau = q(\hat{\theta}_\tau)$ where, for any θ, $\hat{q}(\theta)$ is given by

$$\hat{q}(\theta) = \frac{D_\tau}{D\sum_{k=1}^{\tau} R_\theta(k) + D_\tau R_\theta(0)}, \tag{42}$$

we see that for $\theta = \hat{\theta}_\tau$ and $q = \hat{q}_\tau$ the efficient score equation $\sum_{i=1}^{D} \tilde{\ell}_q(x_i) = 0$ is verified.

Likewise, recall that if one sets to zero all the partial derivatives of the logarithm of the likelihood in eq. (13), one has eq. (17), that is

$$\sum_{x=1}^{\tau} \frac{\dot{R}_\theta(x)}{R_\theta(x)} n_x - D_\tau \frac{\sum_{k=1}^{\tau} \dot{R}_\theta(k)}{\sum_{k=1}^{\tau} R_\theta(k)} = 0.$$

This equality is equivalent to $\sum_{i=1}^{D} \tilde{\ell}_\theta(x_i) = 0$ with q replaced by $\hat{q}(\theta)$ in $\tilde{\ell}_\theta$. Thus again, for $\theta = \hat{\theta}_\tau$ and $q = \hat{q}_\tau$ the efficient score equation $\sum_{i=1}^{D} \tilde{\ell}_\theta(x_i) = 0$ is verified.

We now prove the asymptotic efficiency of the estimators. Note that all results in this proof are stated under the restriction $\tau \leq \tau_0$ when necessary.

By Lemma 2, $\hat{\theta}$ and $\hat{q}(\hat{\theta})$ are such that

$$\frac{1}{\sqrt{D}} \sum_{i=1}^{D} \tilde{\ell}_{\hat{\theta}}(x_i) = 0 \text{and } \frac{1}{\sqrt{D}} \sum_{i=1}^{D} \tilde{\ell}_{\hat{q}}(x_i) = 0. \tag{43}$$

Efficient score functions, as score functions, are centered. Thus, for any parameter (q, θ, F), one has

$$\sum_{x\geq 1} \tilde{\ell}_\theta(x) f^+_{(q,\theta,F)}(x) = 0 \text{ and } \sum_{x\geq 1} \tilde{\ell}_q(x) f^+_{(q,\theta,F)}(x) = 0.$$

As $\tilde{\ell}_\theta$ and $\tilde{\ell}_q$ are free of F, and as this is also true for the plug-in estimators $\tilde{\ell}_{\hat{\theta}}$ and $\tilde{\ell}_{\hat{q}}$, the previous equations imply that

$$\sum_{x\geq 1} \tilde{\ell}_{\hat{\theta}}(x) f^+_{(\hat{q},\hat{\theta},F)}(x) = 0 \text{ and } \sum_{x\geq 1} \tilde{\ell}_{\hat{q}}(x) f^+_{(\hat{q},\hat{\theta},F)}(x) = 0. \tag{44}$$

The asymptotic efficiency of $(\hat{q}, \hat{\theta})$ follows from Theorem 25.54 in [27]. As assumptions, this theorem needs the assertions of Theorem 1 (consistency) and Lemma 1 (Donsker property), in addition to the following two convergence properties pertaining to the "plug-in" score functions. In particular, we need to show that our estimator $(\hat{q}, \hat{\theta})$ satisfies:

$$\|\tilde{\ell}_{(\hat{q},\hat{\theta})} - \tilde{\ell}_{(q_0,\theta_0)}\|^2_{\mathbb{L}^2(f^+)} = o_{\mathbb{P}}(1), \tag{45}$$

and

$$\|\tilde{\ell}_{(\hat{q},\hat{\theta})}\|^2_{\mathbb{L}^2(\hat{f}^+)} = \sum_{x\geq 1} \|\tilde{\ell}_{(\hat{q},\hat{\theta})}(x)\|_2^2 \hat{f}^+(x) = O_{\mathbb{P}}(1), \tag{46}$$

where f^+ stands for $f^+_{(q_0,\theta_0,F)}$, $\hat{f}^+$ stands for the parametric plug-in $f^+_{(\hat{q},\hat{\theta},F)}$, and $\tilde{\ell}_{(q_0,\theta_0)}$ and $\tilde{\ell}_{(\hat{q},\hat{\theta})}$ are the stacked vectors of $(k+1)$ components, $(\tilde{\ell}_{q_0}, \tilde{\ell}_{\theta_0})$ and $(\tilde{\ell}_{\hat{q}}, \tilde{\ell}_{\hat{\theta}})$ respectively.

To establish eqs. (45) and (46), we can use for each parameter α the continuity properties of $\tilde{\ell}_\alpha$ per component, as in the proof of Lemma 1. In particular, note first that for all $x > \tau_0$, we have that $\tilde{\ell}_\alpha(x)$ is constant. Therefore, for each parameter α we need only to account for the convergence of $\tilde{\ell}_{\hat{\alpha}}(x)$ to $\tilde{\ell}_\alpha(x)$ for $x = 1, \cdots, \tau_0 + 1$, all of which happen (in probability), by continuity.

It follows that for each x, $\|\tilde{\ell}_{(\hat{q},\hat{\theta})}(x) - \tilde{\ell}_{(q_0,\theta_0)}(x)\|^2$

converges to 0 in probability, and since we have only finitely many distinct values, the convergence is uniform for all x. eq. (45) is thus immediate.

On the other hand, $\hat{f}^+(x)$ converges to $f^+(x)$ in probability for each x. By finiteness, it follows that $\sum_{x\le\tau_0}\hat{f}^+(x)$ converges to $\sum_{x\le\tau_0} f^+(x)$, and consequently $\sum_{x>\tau_0}\hat{f}^+(x)$ converges to $\sum_{x>\tau_0} f^+(x)$. By using once again the fact that $\tilde{\ell}$ is constant beyond τ, the convergence reduces to finitely many convergences, and thus $\sum_{x\ge1}\|\tilde{\ell}_{(q_0,\theta_0)}(x)\|_2^2\hat{f}^+(x)$ converges to $\sum_{x\ge1}\|\tilde{\ell}_{(q_0,\theta_0)}(x)\|_2^2 f^+(x)$. We can therefore write:

$$\begin{aligned}\sum_{x\ge1}\|\tilde{\ell}_{(\hat{q},\hat{\theta})}(x)\|_2^2\hat{f}^+(x) &\le 2\sum_{x\ge1}\|\tilde{\ell}_{(\hat{q},\hat{\theta})}(x) - \tilde{\ell}_{(q_0,\theta_0)}(x)\|_2^2\hat{f}^+(x) + 2\sum_{x\ge1}\|\tilde{\ell}_{(q_0,\theta_0)}(x)\|_2^2\hat{f}^+(x)\\ &= o_{\mathbb{P}}(1) + O_{\mathbb{P}}\left(\|\tilde{\ell}_{(q_0,\theta_0)}\|^2_{\mathbb{L}^2(f^+)}\right),\end{aligned}$$

which completes the proof of eq. (46) and the theorem.

References

[1] Barger K, Bunge J. Bayesian estimation of the number of species using noninformative priors. Biometrical J. 2008;50:1064–1076.
[2] Bunge J, Barger K. Parametric models for estimating the number of classes. Biomet J. 2008;50:971–982.
[3] Wang J-P. Estimating species richness by a poisson-compound gamma model. Biometrika. 2010;97:727–740.
[4] Chao A, Bunge J. Estimating the number of species in a stochastis abundance model. Biometrics 2002;58:531–539.
[5] Chao A, Jost L. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology. 2012;93:2533–2547.
[6] Chao A, Lee S-M. Estimating the number of classes via sample coverage. J Am Stat Assoc. 1992;87:210–217.
[7] Norris JL, Pollock KH. Nonparametric MLE under two closed capture-recapture models with heterogeneity. Biometrics. 1996;52:639–649.
[8] Norris JL, Pollock KH. Nonparametric MLE for poisson species abundance models allowing for heterogeneity between species. Environ Ecol Stat. 1998;5:391–402.
[9] Chao A, Yang MC. Stopping rule and estimation for recapture debugging with unequal failure rates. Biometrika 1993;80:193–201.
[10] Wang J-P, Lindsay GB. A penalized nonparametric maximum likelihood approach to species richness estimation. J Am Stat Assoc. 2005;100:942–959.
[11] Sanathanan L. Estimating the size of a multinomial population. Ann Math Stat. 1972;43:142–152.
[12] Sanathanan L. Estimating the size of a truncated sample. J Am Stat Assoc. 1977;72:669–672.
[13] Mao CX, Lindsay BG. Tests and diagnostics for heterogeneity in the species problem. Comput Stat Data Anal. 2003;41:389–398.
[14] Mao, CX, Lindsay BG. Estimating the number of classes. Ann Stat. 2007;35:917–930.
[15] Böhning D, Schün D. Nonparametric maximum likelihood estimation of population size based on the couting distribution. J. Royal Stat Soc. 2005;54:721–737.
[16] Chao A. Nonparametric estimation of the number of classes in a population. Scand J Statist 1984;11:265–270.
[17] Zelterman D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. J Stat Plann Inference. 1988;18:225–237.
[18] Böhning D, Vidal-Diez A, Lerdsuwansri R, Viwatwongkasem C, Arnol M. A generalization of Chao's estimator for covariate information. Biometrics. 2013;69:1033–1042.
[19] Böhning D, van der Heijden PG. A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. Ann Appl Stat. 2009;3:595–610.
[20] Goldenshluger A, Lepski O. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. Ann Stat. 2011;39:1608–1632.
[21] Ben Hamou A, Boucheron S, Ohannessian MI. Concentration inequalities in the infinite Urn scheme for occupancy counts and the missing mass, with applications. Bernoulli. 2017;23:249–287.
[22] Lanumteang K, Böhning D. An extension of Chao's estimator of population size based on the first three capture frequency counts. Comput Stat Data Anal. 2011;55:2302–2311.
[23] Fisher RA, Corbet AS, Williams CB. The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecolo. 1943;12:42–58. http://www.jstor.org/stable/1411.
[24] Efron B, Thisted R. Estimating the Number of Unsen Species: How Many Words Did Shakespeare Know? Biometrika. 1976;63:435–447. http://www.jstor.org/stable/2335721.
[25] Good IJ, Toulmin GH. The number of new species, the increase in population coverage, when a sample is increased. Biometrika 1956;43:45–63, http://www.jstor.org/stable/2333577.

[26] Orlitsky A, Suresh AT, Wu Y. Optimal prediction of the number of unseen species. PNAS 2016;113.
[27] van der Vaart AW. Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press.