Nonparametric species richness estimation under convexity constraint

Cécile Durot, Sylvie Huet, François Koladjo, Stéphane Robin

Environmetrics 00, 1–21

DOI: 10.1002/env.XXXX

Nonparametric species richness estimation under convexity constraint

Cécile Durot[a], Sylvie Huet[b*], François Koladjo[bc] and Stéphane Robin[de]

Summary: We consider the estimation of the number N of present species in a given area at a given time, based on the abundances of species that have been observed. We adopt a nonparametric approach where the true abundance distribution p is only supposed to be convex. A definition for convex abundance distributions is proposed. A least-squares estimate of the truncated version of p under the convexity constraint is used. Two estimators of N are deduced, the asymptotic distribution of which are derived. We propose three different procedures, including a bootstrap one, to obtain a confidence interval for N and a standard error for its estimator. The performances of the estimators are assessed in a simulation study and compared with competitors. The proposed method is illustrated on several examples.

Keywords: Abundance distribution; Convex abundance distribution; Least squares estimator; Nonparametric estimation; Species richness estimation.

1. INTRODUCTION

Species richness is one of the oldest ways to evaluate the diversity of species in a given area. We refer to species richness as the number of species present in a given area at a given time. In that sense, it provides a simple measure of biodiversity. This measure has been used in a wide range of domains such as conservation biology (Margules *et al.*, 2002), metagenomics

[a] *UFR SEGMI, Université Paris Ouest Nanterre La Défense, F-92001, Nanterre, France*
[b] *UR1404 MaIAGE, INRA, F-78350 Jouy-en-Josas, France*
[c] *CIMPA-Chaire UNESCO, FAST, UAC, 072BP50 Cotonou, Bénin*
[d] *UMR518 MIA, AgroParisTech, F-75005 Paris, France*
[e] *UMR518 MIA, INRA, F-75005 Paris, France*
[*] *Correspondence to: UR1404 MaIAGE, INRA, F-78350 Jouy-en-Josas, France. E-mail: sylvie.huet@jouy.inra.fr*

This paper has been submitted for consideration for publication in *Environmetrics*

(Allen *et al.*, 2013) or entomology (Longino *et al.*, 2002). Classically, the inference on species richness aims at estimating the number of present species based on the abundances observed for a series of species. The problem is then to estimate the number of unobserved species that are present. The problem traces back to Fisher *et al.* (1943), who first proposed to estimate the distribution of abundance in a Gamma-Poisson framework. Various sampling theoretic frameworks have been considered for modelling observations of species abundance in a population, see for example the review given by Bunge and Fitzpatrick (1993) or more recently by Bunge *et al.* (2014) in the context of microbial diversity estimation. The generic problem can be stated as follows. Suppose that the population is composed of N species and for $i = 1 \dots N$, denote by A_i the abundance (that is the number of observed individuals) of species i and by S_j the number of species with abundance j in a sample. The total number of observed species is $D = \sum_{j \geqslant 1} S_j$ whereas S_0 is the number of unobserved species. The total number of species is $N = S_0 + D$ and, because D is observed, the estimation of N amounts to the estimation of S_0. We will denote by n the sample size: $n = \sum_i A_i = \sum_j jS_j$.

A first approach consists of considering that the n individuals are sampled from an infinite population composed of N species in proportions $w_1, \dots, w_N$. In this setting, Harris (1959) considered the problem of estimating the sample coverage $\sum_i w_i I(A_i \geqslant 1)$ and provided an approximation for the expected number of unobserved species $E(S_0)$. Inspired by this approximation, Chao (1984) proposed an estimator of a lower bound for N and illustrated that her estimator can be considered as an estimator for N if n is large and most of the information is concentrated on the triplet (D, S_1, S_2). Chao and Lee (1992) introduced an estimator based on the estimation of both the expected sample coverage and the variation coefficient of the w_i's. Chao and Lin (2012) also considered lower bound estimators in nonparametric models under very general sampling models.

A second approach is to assume that the A_i's are independent variates with the same distribution $p = (p_0, p_1, \dots, p_n)$, called the *abundance distribution*. In this setting,

Sanathanan (1972) pointed out that if p were known, then the maximum likelihood estimator of N would be

$$\widehat{N}_p = \lfloor D/(1-p_0) \rfloor, \tag{1}$$

where $\lfloor x \rfloor$ denotes the integer part of x. Postulating a parametric assumption on p in order to make p_0 identifiable, Sanathanan (1972) computed the asymptotic distribution of both the maximum likelihood estimator (MLE) and the so-called conditional MLE of N as $N \to \infty$.

Most authors adopting the point of view of independent A_i's with common distribution p assumed that A_i is distributed as a Poisson with expectation λ_i, the λ_i's being independent variables from some distribution ω over $(0, \infty)$ that is called a mixing distribution. Therefore,

$$P(A_i = j) = p_j(\omega) = \int_0^\infty \frac{\lambda^j \exp(-\lambda)}{j!} d\omega(\lambda) \tag{2}$$

and such a setting is called the Poisson mixture setting. It is generally referred to as parametric if a parametric assumption is formulated on ω, and nonparametric otherwise.

In the parametric Poisson mixture setting, Chao and Bunge (2002) provided a consistent estimator when ω is a Gamma distribution. An extension was proposed by Lanumteang and Böhning (2011). Such mixtures are implemented in the CatchAll program (Bunge *et al.*, 2012).

General results about the MLE in the nonparametric Poisson mixture setting can be found in Laird (1978) and Lindsay (1995). For Poisson mixture truncated at zero, Mao and Lindsay (2007) proved that the model is identifiable only if ω has no mass at zero. Norris and Pollock (1998) developed the MLE of both N and ω, using an Expectation-Maximization (EM) algorithm similar to the one of Norris and Pollock (1996). Böhning and Schön (2005) considered an alternative approach using a nested EM algorithm to compute the MLE of N and ω. Both authors proposed bootstrap approaches to obtain confidence

intervals. Wang and Lindsay (2005) pointed out the numerical instability of earlier estimation methods and proposed to use a penalized log-likelihood function to stabilize the estimation procedure. More recently, Wang (2010) considered a continuous estimator for ω – a 'smooth nonparametric MLE' – to better capture the information of species abundance near zero. Unfortunately, there are no asymptotic results on the aforementioned estimators in the nonparametric Poisson mixture setting. In some sense, Mao and Lindsay (2007) proved that no limiting distribution theory could be achievable in this setting. To be more specific, as a consequence of (1), estimating N amounts to estimating the odds $p_0/(1-p_0)$. Mao and Lindsay (2007) proved the discontinuity of the odds as a function of ω, from which they derived that the odds has no locally unbiased and locally informative estimator. They proved that asymptotically valid (as $D \to \infty$) confidence intervals for the odds are necessarily one-sided, which means that only lower bounds (for the odds as well as for N) can be calculated.

In this paper, we propose a new nonparametric approach for estimating N. We assume that the abundances A_i are independent with common distribution p but in contrast to the Poisson mixture setting, we do not assume any parametric or semi-parametric form for p. Our approach is built upon two basic facts.

(i) Most real observed abundance distributions appear to be convex.

(ii) The convexity assumption is sufficient to solve the identifiability issue raised by the estimation of S_0.

Observation (i) results from empirical evidence, which we illustrate in Figure 1. This figure presents four typical abundance distributions which all display a convex shape, together with several estimators discussed below. This general observation was also acknowledged in Bunge *et al.* (2014) (Section 3.2), whose general description of abundance distributions is fully consistent with a convex shape. Furthermore the mixture representation suggested in this reference is consistent with the one we propose hereafter.

Observation (*ii*) is less intuitive and results from the characterization of any convex distribution as a mixture of triangular distributions (see Subsection 2.2). The first component of the mixture corresponds to a Dirac mass at 0, which corresponds to absent species. Hence, we define a *convex abundance distribution* as a convex distribution from which this first component is absent. The weights of all other components of the mixture can be inferred resulting in estimators for S_0 and N.

[Figure 1 about here.]

2. MODEL

2.1. The observations

As only species that are present in the sample can be counted, species for which $A_i = 0$ are not observed. Thus, we only observe the zero-truncated counts $X_1, \ldots, X_D$, where X_i is the abundance of the i-th observed species in the sample. Here, $D \sim \text{Bin}(N, 1 - p_0)$, and conditionally on D, $X_1, \ldots, X_D$ are i.i.d. random variables with distribution p^+ defined by

$$p_j^+ = \frac{p_j}{1 - p_0}, \text{ for all integers } j \geqslant 1, \tag{3}$$

(see Supporting Information). We aim at estimating N, the total number of species.

2.2. The assumption of a convex abundance distribution

To make N identifiable, we propose a nonparametric modelling of p, assuming that p is a discrete convex abundance distribution, as defined below.

A discrete distribution p on $\mathbb{N}$ is convex if $p_j - p_{j-1} \leqslant p_{j+1} - p_j$ for all $j \geqslant 1$. Consequently p is non-increasing, otherwise $\sum_{j \in \mathbb{N}} p_j$ would be infinite. Moreover, it follows from Theorem 7 in Durot *et al.* (2013) that p can be decomposed into a mixture of triangular distributions,

and that this mixture is unique. More precisely, we have

$$p_j = \sum_{\ell \geqslant 1} \pi_\ell T_\ell(j), \tag{4}$$

for all integers $j \geqslant 0$, where

$$\pi_\ell = \frac{\ell(\ell+1)}{2}\left(p_{\ell+1} - 2p_\ell + p_{\ell-1}\right) \text{ for all integers } \ell \geqslant 1 \tag{5}$$

and where T_ℓ is the triangular distribution defined by

$$T_\ell(j) = \frac{2(\ell - j)_+}{\ell(\ell+1)} \text{ for all } j \in \mathbb{N}$$

where $x_+ = \max\{x, 0\}$. Our interpretation of the mixture (4) is that the set of species is separated into groups, each species having probability π_ℓ to belong to the group ℓ of species, and the abundance distribution of all species in the group ℓ is the triangular distribution T_ℓ. As the first component T_1 is a Dirac mass at 0, it refers to species for which the only abundance that could be observed is 0. This group simply defines absent species, and therefore π_1 has to be zero in an abundance distribution. This leads to the following definition:

Definition of a convex abundance distribution: *The distribution p on $\mathbb{N}$ is a convex abundance distribution if there exist positive weights π_ℓ, $\ell \geqslant 2$ such that $p_j = \sum_{\ell \geqslant 2} \pi_\ell T_\ell(j)$ for all integers $j \geqslant 0$.*

In the following, we assume that the abundance distribution p is a convex abundance distribution. It then follows from (5) that $p_2 + p_0 - 2p_1 = 0$, or equivalently, that

$$\frac{1}{1-p_0} = 2p_1^+ - p_2^+ + 1, \tag{6}$$

where p^+ is the zero-truncated distribution defined by (3). The distribution p^+ is identifiable since we observe $X_1, \ldots, X_D$ which are i.i.d. with distribution p^+ conditional on D. Therefore,

it follows from (6) that $1-p_0$ is identifiable and because $D \sim \mathrm{Bin}(N, 1-p_0)$, we conclude that N also is identifiable. This shows that our assumption is sufficient to avoid identifiability problems. The precise construction of the estimates is the aim of the following section.

3. ESTIMATING THE NUMBER OF SPECIES

In order to estimate N, we first build an estimator for $1/(1-p_0)$. Because of (6), we consider estimators of the form $2\widehat{p}_1^+ - \widehat{p}_2^+ + 1$, where $\widehat{p}^+$ is an estimator for p^+ and we estimate N by

$$\widehat{N} = D\left(2\widehat{p}_1^+ - \widehat{p}_2^+ + 1\right). \tag{7}$$

We present two different estimators for N based on two different estimators of p^+.

3.1. Estimators based on the empirical estimator of p^+

For all $j \geqslant 1$, the empirical estimator (which is the more commonly used estimator for a discrete distribution) of p_j^+ is $f_j = S_j/D$. Using this estimator in (7) leads to the estimator

$$\widehat{N}^f = 2S_1 - S_2 + D. \tag{8}$$

One can derive from the central limit theorem (see Supporting Information) that

$$\frac{\widehat{N}^f - N}{\sqrt{6S_1}} \text{ converges in law to } \mathcal{N}(0,1) \tag{9}$$

as $N \to \infty$. Note that this calculation is feasible only if S_1 is strictly positive, which happens with probability that tends to one if the underlying distribution is non-increasing (which is the case under the convexity assumption) and not concentrated at zero.

3.2. Estimators based on the constrained least-squares estimator of p^+

The estimator (8) exploits the convexity assumption only through the identity (6). But it follows from (3) that p^+ is convex if p is convex. We might obtain better estimates by incorporating this information into our estimation procedure, so instead of the empirical estimator, we consider here a convex estimator of p^+: we consider the constrained least-squares estimator $\widehat{p}^+$ defined as the unique solution to the following optimisation problem:

$$Q(\widehat{p}^+) = \inf_{q\in\mathcal{C}} Q(q), \qquad \text{where } Q(q) = \sum_{j\geqslant 1}(q_j - f_j)^2 \tag{10}$$

and where $\mathcal{C}$ denotes the set of all convex sequences q on $\mathbb{N}$ satisfying $\sum_{j\geqslant 1} q_j^2 < \infty$. Durot *et al.* (2013) proved that $\widehat{p}^+$ exists, has a finite support, and is a probability mass function. They provided an algorithm, based on the support reduction algorithm of Groeneboom *et al.* (2008), for computing $\widehat{p}^+$ in a finite number of steps.

3.3. Asymptotic distribution of the number of species' estimator

Let $\widehat{N}$ be defined by (7) where $\widehat{p}^+$ is taken from (10). To compute the limiting distribution of $\widehat{N}$ we need the following definitions and notation. A knot of p^+ is an integer $j \geqslant 2$ such that $p_j^+ - p_{j-1}^+ < p_{j+1}^+ - p_j^+$; a double-knot of p^+ is an integer $j \geqslant 2$ such that both j and $j+1$ are knots of p^+. Notice that p^+ has at least one knot since it is a convex probability mass function on $\mathbb{N}$. However, double-knots of p^+ may not exist.

- Let τ be the maximum of the support of p^+ if it is finite, and $\tau = \infty$ otherwise.
- Let κ be the smallest double-knot of p^+ if it exists, and $\kappa = \infty$ otherwise.
- Let $k \geqslant \min\{\tau + 1, \kappa\}$, let $\mathcal{J}$ be the set of all knots of p^+ that are smaller than k and set $\mathcal{I} = \{1, k\} \cup \mathcal{J}$. We assume that either p^+ has a finite support, or p^+ has at least one double knot. This amounts to assume that $\min\{\tau, \kappa\} < \infty$.

- Let W be a centered Gaussian vector in $\mathbb{R}^k$ with covariance matrix Γ defined as $\Gamma_{jj} = p_j^+(1-p_j^+)$ and $\Gamma_{jj'} = -p_j^+ p_{j'}^+$ for all $1 \leqslant j, j' \leqslant k$ and $j \neq j'$.
- Denoting by $1 = i_1 < i_2 < \ldots < i_I = k$ the points in $\mathcal{I}$, let

$$\mathcal{C}^{\mathcal{I}} = \left\{ q \in \mathbb{R}^k \text{ such that } q \text{ is convex on } \{i_{j-1}, \ldots, i_j\} \text{ for all } j = 2, \ldots, I \right\}.$$

Let $W^{\mathcal{I}} = (W_1^{\mathcal{I}}, \ldots, W_k^{\mathcal{I}})$ be the minimizer of $\sum_{j=1}^k (q_j - W_j)^2$ over $q = (q_1, \ldots, q_k)$ in $\mathcal{C}^{\mathcal{I}}$. One can show that

$$\frac{\widehat{N} - N}{\sqrt{D}} \text{ converges in law to } 2W_1^{\mathcal{I}} - W_2^{\mathcal{I}} + \frac{\sqrt{p_0}}{1-p_0}\mathcal{Z}$$

as $N \to \infty$, where $\mathcal{Z} \sim N(0,1)$ is independent of $W^{\mathcal{I}}$ (see Supporting Information).

4. CONFIDENCE INTERVALS

4.1. Estimation based on the empirical frequencies

If N is large and if the quantities Np_0 and $N(1-p_0)$ are not too small, then it follows from (9) that the distribution of $(\widehat{N}^f - N)/\sqrt{6S_1}$ can be approximated by that of a standard Gaussian variable. This leads to the following confidence interval

$$\mathrm{CI}^f = \left[\widehat{N}^f - z_{1-\alpha/2}\sqrt{6S_1}\ ,\ \widehat{N}^f + z_{1-\alpha/2}\sqrt{6S_1} \right], \tag{11}$$

where $\alpha \in (0,1)$ is fixed and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard Gaussian distribution. According to (9), the asymptotic level of the interval is $1-\alpha$.

4.2. The plug-in procedure on $\widehat{N}$

The limiting distribution of $\widehat{N}$ given in Section 3.3 depends on p^+ through $k, \mathcal{I}$, the covariance matrix Γ of the Gaussian vector W, and the variance $p_0/(1-p_0)^2$ of the additional Gaussian

variable. We estimate p_0 by $1 - 1/(2\widehat{p}_1^+ - \widehat{p}_2^+ + 1)$ with $\widehat{p}^+$ being the constrained least-squares estimator of p^+, and we estimate all unknown quantities depending on p^+ by similar quantities with p^+ replaced by $\widehat{p}^+$. For simplicity, we consider $k = \min\{\tau + 1, \kappa\}$ and we estimate k and $\mathcal{I}$ as follows. Let $\widehat{s}$ be the first double knot of $\widehat{p}^+$ if it exists. In the case where such a double knot does not exist we use the convention that $\widehat{s} = \infty$. Let $\widehat{\tau}$ be the maximum of the support of $\widehat{p}^+$. From Theorem 1 in Durot *et al.* (2013) we know that $\widehat{\tau}$ is finite. Therefore, $\widehat{k} = \min\{\widehat{s}, \widehat{\tau} + 1\}$ is finite and $\mathcal{I}$ is estimated by $\widehat{\mathcal{I}}$, the set consisting of 1 and the knots of $\widehat{p}^+$ before $\widehat{k}$. Then, the estimated quantiles of the random variable $2W_1^{\widehat{\mathcal{I}}} - W_2^{\widehat{\mathcal{I}}} + \sqrt{\widehat{p}_0}/(1 - \widehat{p}_0)\mathcal{Z}$, say $\widehat{\lambda}_{1-\alpha/2}$ and $\widehat{\lambda}_{\alpha/2}$, are obtained by simulation. The calculation of $W^{\mathcal{I}}$ is done using the algorithm proposed by Dykstra (1983) for restricted least squares regression; see Balabdaoui *et al.* (2014) for more details. Then we consider the confidence interval for N given by

$$\mathrm{CI} = \left[\widehat{N} - \widehat{\lambda}_{1-\alpha/2}\sqrt{D}, \widehat{N} - \widehat{\lambda}_{\alpha/2}\sqrt{D}\right]. \tag{12}$$

The obvious advantage of the interval (11) as compared to (12) lies in its computational simplicity, due to the fact that it is based on the empirical frequencies rather than on the constrained estimator $\widehat{p}^+$. Moreover its level tends to $1 - \alpha$ as $N \to \infty$. In contrast, the asymptotic level of the confidence interval (12) is not known, due to the unknown asymptotic behaviour of $\widehat{\mathcal{I}}$. However, it is not clear in the general case which of those two intervals has better length or coverage probability. Such a comparison can easily be performed only in the special case where $\min\{\tau + 1, \kappa\} = 2$. In that case, $\widehat{N}$ and $\widehat{N}^f$ have the same limiting distribution (see the comments just below Corollary 1 in Supporting Information), so the difference between the two intervals relies on the way we estimate the unknown parameters in the limiting distribution, and on the chosen center $\widehat{N}_f$ or $\widehat{N}$ for the interval. The two intervals will be compared in the next section. An alternative to the plug-in method is to use a bootstrap procedure for estimating the quantiles of the limiting distribution of $\widehat{N}$.

4.3. The bootstrap procedure on $\widehat{N}$

The bootstrap procedure consists of generating a bootstrap sample as follows: draw a binomial random variable D^* with parameters $\widehat{N}$ and $1-\widehat{p}_0$, then draw a D^*-sample $(X_1^*,\ldots,X_{D^*}^*)$ from $\widehat{p}^+$. We calculate the statistics $f_j^* = \sum_{i=1}^{D^*} I(X_i^*=j)/D^*$ for all j in the support of $\widehat{p}^+$ and the bootstrap estimator of p^+ by minimizing $\sum_{j\geqslant 1}(q_j - f_j^*)^2$ over $q\in\mathcal{C}$. Finally we get $\widehat{p}_0^*$, the bootstrap estimator of p_0, and $\widehat{N}^* = D^*/(1-\widehat{p}_0^*)$. For a fixed $\beta\in(0,1)$, the β-quantile ζ_β of $(\widehat{N}-N)/\sqrt{D}$ is estimated by the β-quantile ζ_β^* of the distribution of $(\widehat{N}^*-\widehat{N})/\sqrt{D^*}$. Finally the bootstrap confidence interval for N is

$$\mathrm{CI}^* = \left[\widehat{N} - \sqrt{D}\zeta_{1-\alpha/2}^* \ ,\ \widehat{N} - \sqrt{D}\zeta_{\alpha/2}^*\right]. \tag{13}$$

Note that in contrast to the confidence interval in (11), no asymptotic results are available for the confidence interval in (13).

5. STANDARD ERROR

Following (9), the standard error of $\widehat{N}^f$ is estimated by $\mathrm{SE}^f = \sqrt{6S_1}$. Applying the plug-in procedure of Section 4.2, the standard error of $\widehat{N}$ is estimated as the standard error of the random variable $2W_1^{\widehat{\mathcal{I}}} - W_2^{\widehat{\mathcal{I}}} + \sqrt{\widehat{p}_0}/(1-\widehat{p}_0)\mathcal{Z}$ and is estimated by simulation. It will be denoted SE. Finally applying the bootstrap procedure of Section 4.3, the bootstrap estimate of the standard error of $\widehat{N}$, denoted SE^*, is the standard error of $\widehat{N}^*$ estimated by simulation.

6. SIMULATION STUDY

We designed a simulation study to assess the performance of the estimators $\widehat{N}$ and $\widehat{N}^f$, and the confidence intervals CI^f, CI and CI^*. We compare our procedure to other methods.

6.1. Simulation design

Convex abundance distribution. We assumed a Poisson mixture setting with a Gamma mixing distribution ω. We were motivated by the fact that all methods to which we will compare (Norris and Pollock 1996; 1998; Wang and Lindsay, 2005; 2008; Wang, 2010; Lanumteang and Boehning, 2011) are either based on this assumption, or are proved to give consistent estimators of N under this assumption, or consider a statistical modeling that covers this distribution. Precisely, p is a Gamma-Poisson distribution that takes the form

$$p_j = \frac{\Gamma(j+\nu)}{\Gamma(\nu)j!}\mu^\nu(1-\mu)^j \tag{14}$$

for some unknown $\nu > 0$ and $\mu \in (0,1)$. Note that for such distributions, $p_0 = \mu^\nu$ and $p_j^+ = p_j/(1-\mu^\nu)$ for $j \geqslant 1$. We focused on the case where p is a convex abundance distribution, which is satisfied (see Supporting Information) when

$$\nu > 1 \quad \text{and} \quad 1-\mu = \frac{2\nu - \sqrt{2\nu(\nu-1)}}{\nu(\nu+1)}.$$

Simulation parameters and evaluation criteria. To cover a wide range of possible applications, we chose N in the set $\{50, 100, 200, 400, 800, 1500, 3000, 5000, 10000\}$. We considered $\nu \in \{1.01, 1.05, 1.1, 1.3, 1.5, 1.75\}$ corresponding to the following values of p_0: $\{0.073, 0.160, 0.218, 0.330, 0.382, 0.420\}$. All confidence intervals were computed at level $\alpha = 0.05$.

All simulation results were based on 1000 samples including the computation of the quantiles $(\zeta^*_{\alpha/2}, \zeta^*_{1-\alpha/2})$ used in the bootstrap procedure and the computation of the quantiles $(\widehat{\lambda}_{\alpha/2}, \widehat{\lambda}_{1-\alpha/2})$ used in the plug-in procedure. The simulations were carried out with R (R Core Team, 2014). The R functions are available in the Supporting Information.

The accuracy of an estimate $\widehat{N}$ was measured in terms of bias, standard error and root mean

squared error of prediction. The bias of an estimator $\widehat{N}$ was estimated by $\texttt{bias} = \widehat{N}_\bullet - N$, where $\widehat{N}_\bullet = \sum_s \widehat{N}_s/1000$ with $\widehat{N}_s$ being the estimate of N at simulation s. The standard error of $\widehat{N}$ was estimated by $\texttt{se} = \sqrt{\sum_s (\widehat{N}_s - \widehat{N}_\bullet)^2/1000}$, and the root mean squared error of prediction by $\texttt{RMSEP} = \sqrt{\texttt{bias}^2 + \texttt{se}^2}$. For the sake of comparison between the different values of N, we will rather consider these quantities divided by N.

The quality of a confidence interval $[\mathrm{B_{inf}}, \mathrm{B_{sup}}]$ was measured in terms of left and right non-coverage probabilities, defined by $P(N < \mathrm{B_{inf}})$ and $P(N > \mathrm{B_{sup}})$ respectively. They were estimated by $\sum_s I(N < \mathrm{B}_{s,\mathrm{inf}})/1000$ and $\sum_s I(N > \mathrm{B}_{s,\mathrm{sup}})/1000$ respectively, where $\mathrm{B}_{s,\mathrm{inf}}$ and $\mathrm{B}_{s,\mathrm{sup}}$ stand for the bounds obtained at simulation s. These quantities give information on the assymetry of the confidence intervals.

6.2. Comparison of $\widehat{N}$ and $\widehat{N}^f$

The relative bias $\texttt{bias}/N$, standard error $\texttt{se}/N$ and root mean squared error of prediction $\texttt{RMSEP}/N$ for both estimators $\widehat{N}$ and $\widehat{N}^f$ are shown on Figure 2. For $p_0 = 0.218$ (respectively $N = 200$), we display these quantities versus N (respectively versus p_0). The graphs for other values of p_0 and N, being similar to those two, are omitted.

While $\widehat{N}^f$ is nearly unbiased, $\widehat{N}$ tends to over estimate N for small values of N and p_0. However, $\widehat{N}$ has a smaller relative standard error and a smaller relative $\texttt{RMSEP}$. Both estimators become more accurate as N increases and as p_0 decreases. Indeed, as p_0 is small, almost all of the species have been observed, leading to a smaller relative standard error.

[Figure 2 about here.]

6.3. Comparison of confidence intervals

Figure 3 shows the estimated non-coverage probabilities of the confidence intervals and are compared with 2.5%. Note however that the standard error of the estimated non-coverage probabilities based on 1000 simulations equals $\sqrt{0.025(1-0.025)/1000} \approx 0.5\%$,

so the estimated non-coverage probabilities are expected to lie typically within 1.5% and 3.5%.

[Figure 3 about here.]

- *Lower bound of the confidence intervals.* For the smallest values of p_0, the estimated values of $P(N < \mathrm{B_{inf}})$ are smaller than 2.5%, which means that the lower bounds of the confidence intervals are too conservative. This tendancy vanishes as N increases. The three methods are nearly equivalent.
- *Upper bound of the confidence intervals.* For the largest values of p_0, the estimated values of $P(N > \mathrm{B_{sup}})$ are greater than 2.5%, and tend to decrease with N up to 2.5%. This means that as N is small, the upper bounds of the confidence intervals are too small. For the smallest values of p_0, the interval CI^f based on the empirical procedure gives very high values of the non-coverage probability as N is small. As N increases, the estimated $P(N > \mathrm{B_{sup}})$ fluctuates around 4%. To verify if these values tend to 2.5% as N increases, we completed the simulation study by considering $N = 10^5$ and $N = 5 \times 10^5$, with $p_0 = 0.218$. The estimated values of the non-coverage probabilities $P(N > \mathrm{B_{sup}})$ were equal to 2.6% and 2.4% respectively, for the interval CI based on the plug-in procedure.

6.4. Comparison of estimated standard errors

The estimated standard errors SE^f, SE and SE^* (Figure 4) are compared with the standard error $\sqrt{6Np_1}$ (see Supplementary Information) of the estimator $\widehat{N}^f$ and with the standard error of $\widehat{N}$. As expected the estimator of the standard error of $\widehat{N}^f$ is nearly unbiased. The estimator SE over-estimates the standard error of $\widehat{N}$ even when N is large. However the estimator SE^* based on the bootstrap procedure is less biased than SE as soon as p_0 is not too small. In all cases the medians of the distributions of SE and SE^* are not larger than the median of the distribution of SE^f, nor than the standard error of $\widehat{N}^f$.

[Figure 4 about here.]

6.5. Method Recommendations

The confidence interval based on the empirical procedure is very easy to calculate and gives reasonable results provided that N is large and p_0 not too small. But, the standard error of $\widehat{N}^f$ is higher than that of $\widehat{N}$.

In some cases, the confidence intervals based on the plug-in procedure are better (in terms of non-coverage probabilities) than those based on the empirical procedure, especially when p_0 is small. The bootstrap procedure does not improve the coverage of the confidence intervals in the simulation study. However the bootstrap estimators of the standard error of $\widehat{N}$ are less biased than the estimators based on the plug-in method. Therefore we recommend using the bootstrap procedure to obtain an estimator with small variability and a good estimate of this variability.

The empirical procedure may be used when one needs results requiring quick calculation. Obviously the computation time for the other methods is higher than for the empirical one, and depends on both N and ν. In our simulation study, the worse case for the plug-in procedure was for $N = 800$ and $\nu = 1.01$, where the mean computation time was 70 s using an algorithm written in `R` on a 64 bits processor with 48 GB of RAM; whereas the bootstrap method needed 150 s.

6.6. Comparison with other methods

We will denote by `Emp` the method that uses the estimator N^f and the confidence interval CI^f defined in (8) and (11) respectively, and we will denote by `CvxPi` the method that uses the estimator $\widehat{N}$ based on the convex least-squares estimator, and the confidence interval CI defined in (12). Because we consider a non-parametric point of view, we focused our comparison on methods that do not involve a parametric distribution of the abundance.

This includes the methods proposed by Chao (1984), Chao and Lee (1992), Chao and Bunge (2002), and Lanumteang and Böhning (2011) denoted `Chao84`, `ChaoLee`, `ChaoBunge`, `LB` respectively. We also consider the following methods based on the maximum likelihood estimation of N and p^+ under the assumption of a Poisson mixture model, or a Poisson-compound Gamma model: `unpmle` (for unconditional nonparametric maximum likelihood estimator) proposed by Norris and Pollock (1996, 1998), `pnpmle` (for penalized nonparametric maximum likelihood estimator) proposed by Wang and Lindsay (2005, 2008) and `pcg` (for Poisson-compound Gamma estimator) proposed by Wang (2010). The simulations were carried out using the library `SPECIES` (Wang, 2011). The last three methods require the choice of a cutoff value, denoted by t, since only the less abundant species are used in these procedures. The behaviour of the procedures as well as the estimation of N may strongly depend on the choice of t. We chose t according to the authors' recommendations. Nevertheless, it appeared that it was not possible to carry out a simulation study for these methods because of numerical problems possibly coming from a bad choice of t: either the algorithm failed to converge or the R function did not return to the main program. Because these numerical problems occurred in a large number of runs, these methods are not considered in our simulation study. But we will come back to the comparison with these methods in Section 7.

The methods `Emp` and `CvxPi` outperformed the other methods in almost all situations in terms of the `RMSEP` (Table 1). The non-coverage probabilities were not too far from 2.5%, and typically much closer to 2.5% than the other methods. For the other methods, the behaviour of the bias and standard error depended strongly on the considered case. The confidence intervals were typically shifted to the left with boundaries smaller than N. This behaviour was less marked for the `LB` method.

[Table 1 about here.]

6.7. Robustness to convexity

The convexity assumption on p^+ seems reasonable when looking at the observed zero-truncated abundance distributions in several examples. Nevertheless, convexity of p^+ does not imply the convexity of p. To evaluate the robustness of our procedure to convexity of p, we carried out a simulation study considering distributions p' defined as follows:

$$p'_0 = (p_0 + p_1)/2, \text{ and } p'_j = p_j \frac{1 - p'_0}{1 - p_0} \text{ for } j \geqslant 1$$

where the probabilities p_j, for $j \geqslant 0$ are defined in (14). We chose $p'_0 = 0.0707$ leading to small deviation from convexity, and $p'_0 = 0.354$ leading to a larger deviation from convexity.

Our procedures lead to positively biased estimators ($E(\widehat{N}) \geqslant N$) (Table 2), which was expected because p'_0 is smaller that $2p'_1 - p'_2$ especially in the case where $p'_0 = 0.354$. In contrast, the standard errors of our estimators were not affected by the lack of convexity. The confidence intervals were shifted to the right with boundaries larger than N in the case $N = 5000$ and $p'_0 = 0.354$. The behaviour of the other methods depended strongly on the values of (N, p'_0). Based on this simulation study, we conclude that our methods still apply in case of small discrepancy of the true abundance distribution to convexity.

[Table 2 about here.]

7. ILLUSTRATION ON PUBLIC DATASETS

7.1. Testing convexity of p^+

First, we want to check whether the convexity assumption of p^+ is reasonable. Recall that p^+ is convex if $\Delta_j \geqslant 0$ for all $j \geqslant 2$, where $\Delta_j = p^+_{j+1} - 2p^+_j + p^+_{j-1}$. Hence, we propose to reject the hypothesis that p^+ is convex if one of the empirical estimators $\Delta^f_j = f_{j+1} - 2f_j + f_{j-1}$ is smaller than some negative threshold. The test procedure is defined as follows.

Let τ^f be the maximum of the support of the empirical distribution f, $\tau^f = \max_{i=1,\dots D} X_i$ and let Γ^f be the matrix with components $\Gamma^f_{jj'} = -f_j f_{j'}$ if $j \neq j'$ and $\Gamma^f_{jj} = f_j(1-f_j)$ for $1 \leqslant j, j' \leqslant \tau^f$. Let us introduce the matrix A whose lines A_j^T satisfy $\Delta^f_{j+1} = A_j^T f$ for $j = 1, \dots, \tau^f - 1$, and let the matrix M^f be defined as the square-root of the matrix $A\Gamma^f A^T$: $M^f M^f = A\Gamma^f A^T$. Finally, for $\alpha > 0$ let q^f_α be the α-quantile of the conditional distribution of $\min_{1\leqslant j\leqslant \tau^f-1} \sum_{j'=1}^{\tau^f-1} M^f_{jj'} \mathcal{Z}_{j'}$ given $X_1, \dots, X_D$, where the $\mathcal{Z}_{j'}, j' = 1, \dots \tau^f - 1$ are independent centered Gaussian variates. The threshold q^f_α is calculated by simulation, and the rejection region for testing that p^+ is convex is defined as $\left\{\sqrt{D}\min_{1\leqslant j\leqslant \tau^f-1} \Delta^f_j \leqslant q^f_\alpha\right\}$. It is proved that this test is approximatively of level α (see Supplementary Information).

For all data sets presented in the introduction, the assumption of convexity of p^+ is not rejected, the smallest p-value being equal to 0.58.

7.2. Estimating the number of species

For each of the data sets, we implemented all of the methods of Subsection 6.6. For the methods `unpmle`, `pnpmle` and `pcg`, we chose the cutoff value t according to the authors recommendations under the condition that the algorithm did converge, and to the goodness-of-fit of the empirical frequencies. For the `ChaoBunge` procedure we chose $t = 10$ according to the authors recommendations. In cases where the estimation of N was negative, we decreased t such that the resulting estimator was positive. The values of t are given in Figure 1.

In the Microbial and Tomato data, where the empirical distribution is convex, the `unpmle` estimators and the `pcg` estimator (for the Tomato data) are almost equal to the empirical distribution, as it is obviously the case for our estimator (Figure 1). However the estimated values of N in the Tomato data, differ a lot from one method to another: from 4439 for our estimator, to 7417 for the `unpmle` estimator, up to 13960 for the `pcg` estimator (Table 3). This indicates that the assumptions underlying each of the methods have strong consequences on the final estimates. We feel that the definition for convex abundance distribution we propose

is both consistent with the shape of observed distributions and less restrictive in terms of assumptions.

In the Malayan Butterfly datasets, all estimators are nearly convex for the less abundant species. The estimated values of N are less variable than for the two preceding examples (Table 3) and our method gives the highest value.

In the Bird example, the empirical distribution is non convex with $f_1 < f_2$. Nevertheless, the estimator of p^+ based on the `pcg` method is convex. The estimators based on the two non-parametric Poisson mixture methods are non convex but $\widehat{p}_1^+$ and $\widehat{p}_2^+$ are far above f_1 and f_2. It is not easy to decide based on Figure 1 which estimate of p^+ is preferred.

[Table 3 about here.]

SUPPLEMENTARY MATERIALS

Supporting Information referenced in Sections 1, 3, 6 and 7.1 is available with this paper at the journal website on Wiley Online Library.

ACKNOWLEDGEMENTS

REFERENCES

Allen HK, Bunge J, Foster JA, Bayles DO, Stanton TB. 2013. Estimation of viral richness from shotgun metagenomes using a frequency count approach. *Microbiome* **1**(1): 5.

Balabdaoui F, Durot C, Koladjo F. 2014. On asymptotics of the discrete convex lse of a pmf. *arXiv preprint arXiv:1404.3094* .

Böhning D, Schön D. 2005. Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(4): 721–737.

Bunge J, Fitzpatrick M. 1993. Estimating the number of species: a review. *Journal of the American Statistical Association* **88**(421): 364–371.

Bunge J, Willis A, Walsh F. 2014. Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* **1**: 427–445.

Bunge J, Woodard L, Böhning D, Foster JA, Connolly S, Allen HK. 2012. Estimating population diversity with catchall. *Bioinformatics* **28**(7): 1045–1047.

Chao A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**: 265–270.

Chao A, Bunge J. 2002. Estimating the number of species in a stochastic abundance model. *Biometrics* **58**(3): 531–539.

Chao A, Lee SM, 1992. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* **87**(417): 265–270.

Chao A, Lin CW. 2012. Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. *Biometrics* **68**(3): 912–921.

Durot C, Huet S, Koladjo F, Robin S. 2013. Least-squares estimation of a convex discrete distribution. *Computational Statistics & Data Analysis* **67**: 282–298.

Dykstra RL. 1983. An algorithm for restricted least squares regression. *Journal of the American Statistical Association* **78**(384): 837–842.

Fisher RA, Steven-Corbet A, Williams CB. 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* **12**(1): 42–58.

Groeneboom P, Jongbloed G, Wellner JA. 2008. The support reduction algorithm for computing non-parametric function estimates in mixture models. *Scandinavian Journal of Statistics. Theory and Applications* **35**(3): 385–399.

Harris B. 1959. Determining bounds on integrals with applications to cataloging problems. *The Annals of Mathematical Statistics* **30**(417): 210–217.

Laird N. 1978. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73**(364): 805–811.

Lanumteang K, Böhning D. 2011. An extension of Chao's estimator of population size based on the first three capture frequency counts. *Computational Statistics & Data Analysis* **55**(7): 2302 – 2311.

Lindsay BG. 1995. Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, volume 5, Institute of Mathematical Statistics, i–163.

Longino JT, Coddington J, Colwell RK. 2002. The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology* **83**(3): 689–702.

Mao CX, Lindsay BG. 2007. Estimating the number of classes. *The Annals of Statistics* : 917–930.

Margules CR, Pressey R, Williams P. 2002. Representing biodiversity: data and procedures for identifying priority areas for conservation. *Journal of biosciences* **27**(4): 309–326.

Norris JL, Pollock KH. 1996. Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics* **52**(4): 639–649.

Norris JL, Pollock KH. 1998. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics* **5**(4): 391–402.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Sanathanan L. 1972. Estimating the size of a multinomial population. *The Annals of Mathematical Statistics* **43**(1): 142–152.

Wang JP. 2010. Estimating species richness by a poisson-compound gamma model. *Biometrika* **97**(3): 727–740.

Wang JP. 2011. SPECIES: An R package for species richness estimation. *Journal of Statistical Software* **40**(9): 1–15.

Wang JP, Lindsay BG. 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100**(471): 942–959.

Wang JP, Lindsay BG. 2008. An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology* **5**(1): 30–45.

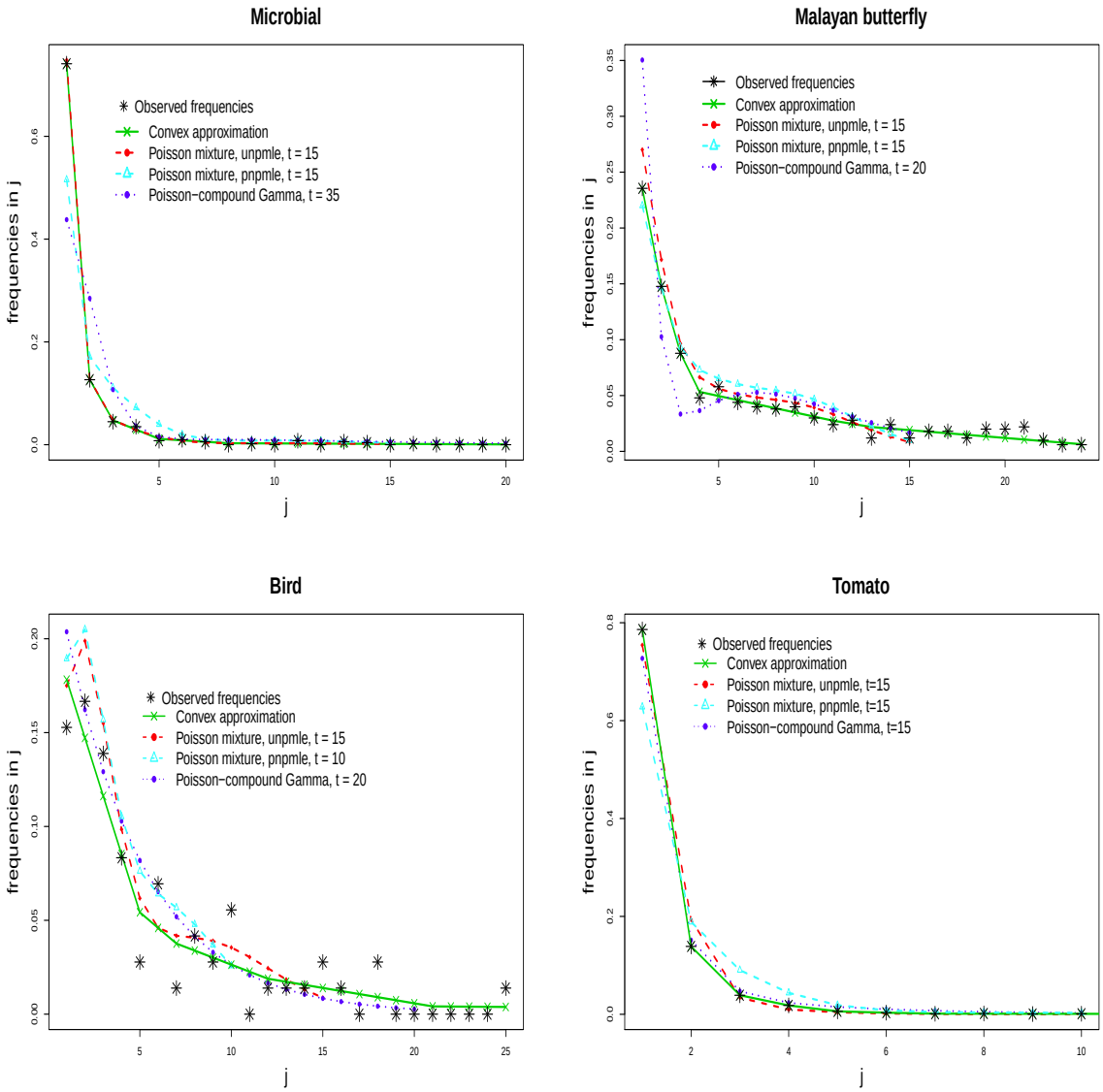

Figure 1. Estimation of the zero-truncated distribution p^+. The green curve is the projection of the empirical frequencies onto the set of convex distributions, the red and the light blue curves are the estimated distribution under a nonparametric Poisson mixture, and the dark blue curve is the estimation obtained under the Poisson-compound Gamma model. For the last three methods the cutoff value t is given.

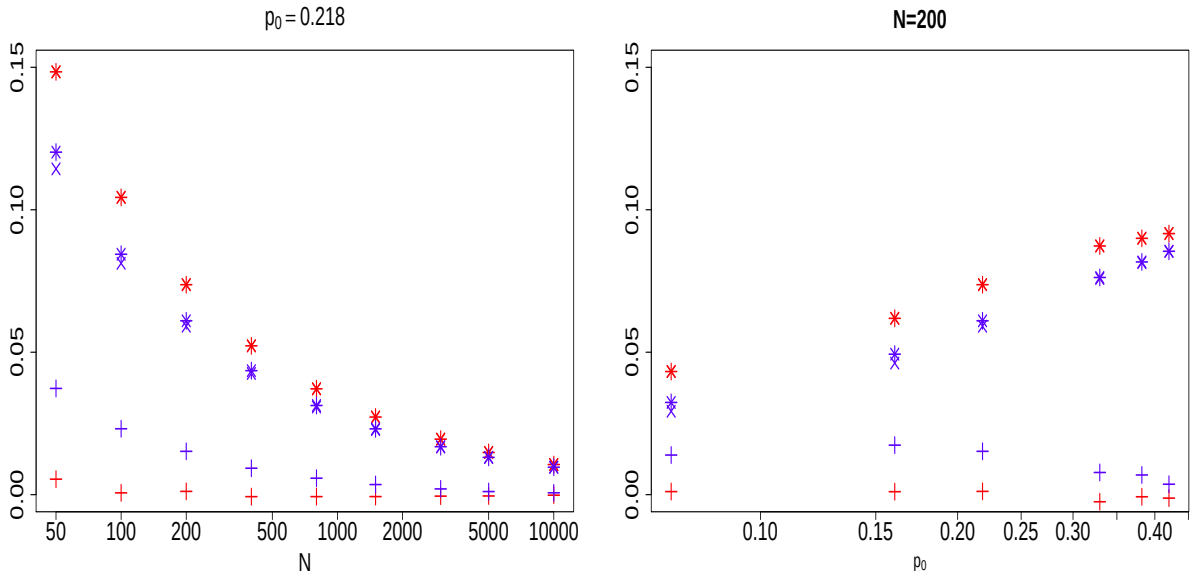

Figure 2. Relative bias $+$, relative standard errors $\times$ and relative `RMSEP` $*$ versus N for $p_0 = 0.218$ on the left side, versus p_0 for $N = 200$ on the right side. The red color is for $\widehat{N}^f$ and the blue one for $\widehat{N}$.

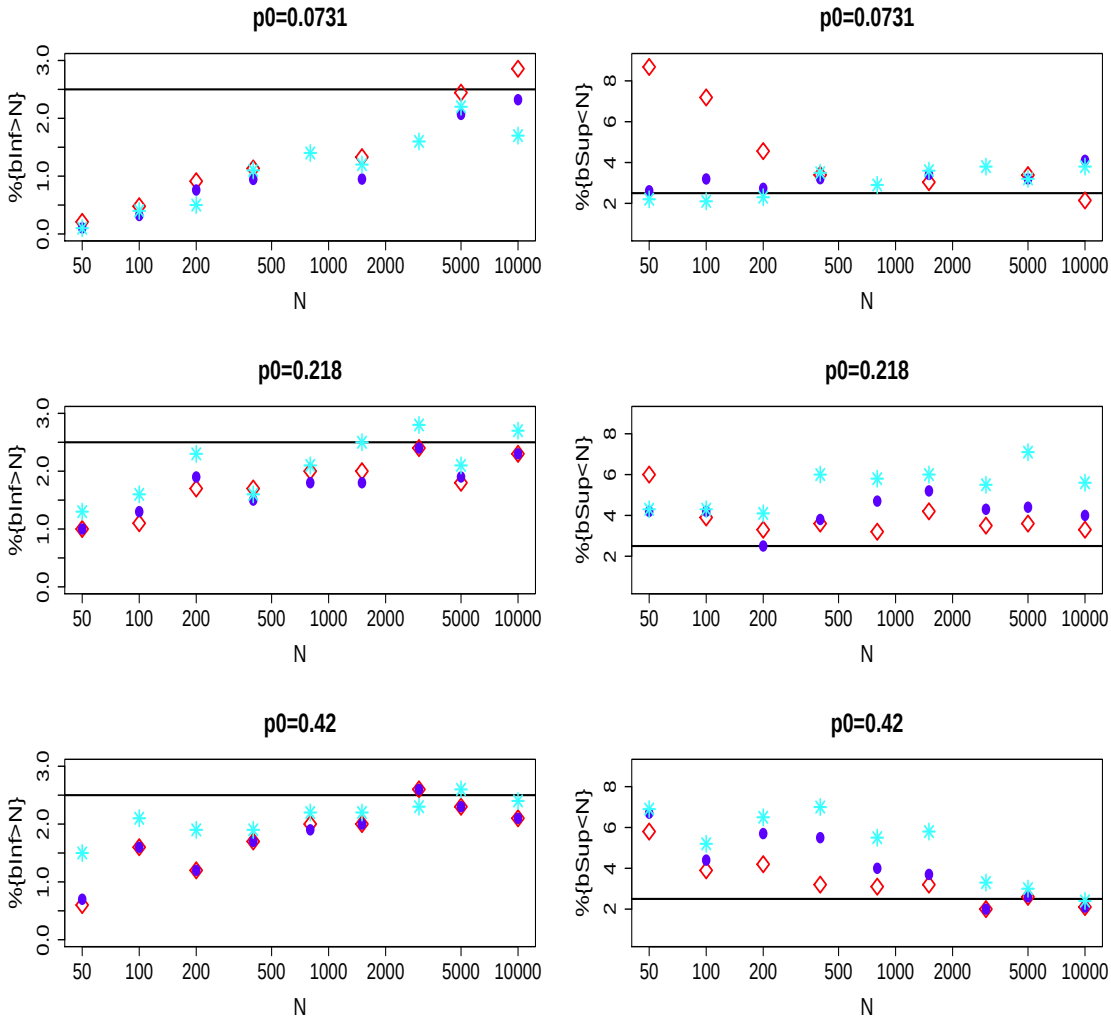

Figure 3. Comparison of confidence intervals. On the left: estimated values of $P(N < \mathrm{B_{inf}}) \times 100\%$ versus N for three values of p_0. The legend is the following : $\diamond$ is for CI^f (Equation (11)), $\bullet$ is for CI (Equation (12)), and $*$ is for the bootstrap confidence interval CI^* (Equation (13)). The horizontal line represents the wanted probability 2.5%. Right: estimated values of $P(N > \mathrm{B_{sup}}) \times 100\%$.

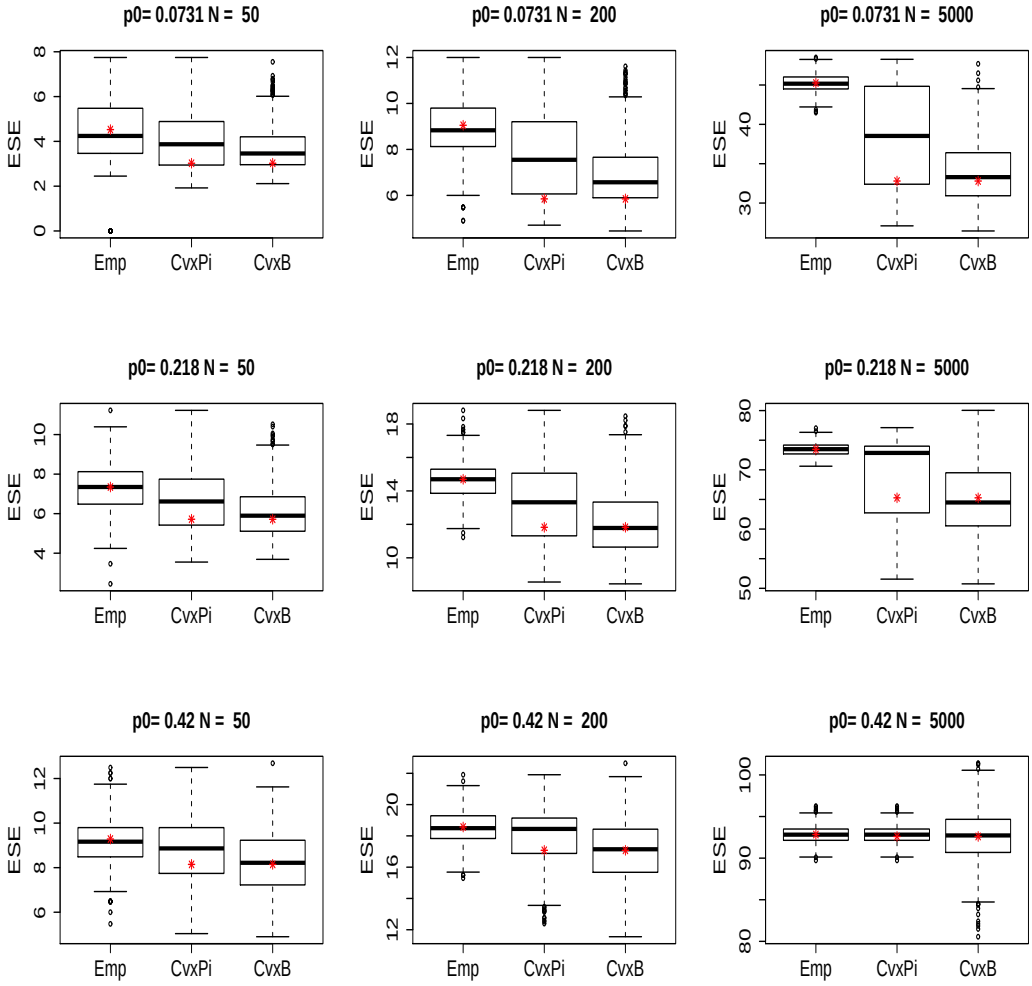

Figure 4. Comparison of estimated standard errors (`ESE`) when p_0 and N vary. For each graphic, from the left to the right : the boxplots represent the distributions of SE^f, SE and SE^* respectively denoted `Emp`, `CvxPi`, `CvxB`. The distribution are estimated on the basis of 1000 simulations. The red stars $*$ are the standard errors of $\widehat{N}^f$ and $\widehat{N}$.

Table 1. Comparison of methods `Emp`, `CvxPi`, `ChaoBunge`, `Chao84`, `ChaoLee`, `LB` defined in Section 6.6, for estimating N: relative bias, relative standard error, relative `RMSEP` and non-coverage probabilities ($\times 100\%$) are reported, for $N = 100$ and $N = 5000$, considering $p_0 = 0.073$ and $p_0 = 0.42$.

$N = 100$, $p_0 = 0.073$						
	`Emp`	`CvxPi`	`ChaoBunge`	`Chao84`	`ChaoLee`	`LB`
`bias/N`	0.002	0.020	-0.041	-0.020	-0.042	0.310
`se/N`	0.061	0.041	0.033	0.057	0.031	2.240
`RMSEP/N`	0.061	0.046	0.052	0.061	0.052	2.260
$P(B_{\text{inf}} > N)$	0.8	0.6	0.0	1.0	0.0	0.0
$P(B_{\text{sup}} < N)$	6.7	2.7	38.0	6.9	40.0	22.0
$N = 100$, $p_0 = 0.42$						
	`Emp`	`CvxPi`	`ChaoBunge`	`Chao84`	`ChaoLee`	`LB`
`bias/N`	0.0007	0.0110	0.1230	-0.1240	-0.0600	0.2200
`se/N`	0.133	0.121	1.690	0.158	0.194	1.200
`RMSEP/N`	0.133	0.122	1.690	0.201	0.203	1.220
$P(B_{\text{inf}} > N)$	1.6	1.6	0.0	0.4	1.0	0.0
$P(B_{\text{sup}} < N)$	3.9	4.4	6.5	13.6	8.0	20.0
$N = 5000$, $p_0 = 0.073$						
	`Emp`	`CvxPi`	`ChaoBunge`	`Chao84`	`ChaoLee`	`LB`
`bias/N`	0.0001	0.0025	-0.0450	-0.0360	-0.0450	-0.0160
`se/N`	0.0091	0.0066	0.0044	0.0060	0.0045	0.0140
`RMSEP/N`	0.0091	0.0070	0.0450	0.0360	0.0450	0.0220
$P(B_{\text{inf}} > N)$	2.5	2.2	0.0	0.0	0.0	0.0
$P(B_{\text{sup}} < N)$	2.8	2.8	100.0	100.0	100.0	33.0
$N = 5000$, $p_0 = 0.42$						
	`Emp`	`CvxPi`	`ChaoBunge`	`Chao84`	`ChaoLee`	`LB`
`bias/N`	-0.0002	-0.0002	-0.0080	-0.1530	-0.0850	-0.0530
`se/N`	0.018	0.018	0.040	0.019	0.025	0.054
`RMSEP/N`	0.018	0.018	0.041	0.154	0.089	0.076
$P(B_{\text{inf}} > N)$	2.3	2.3	0.9	0.0	0.0	0.2
$P(B_{\text{sup}} < N)$	2.6	2.6	4.3	100.0	89.0	23.0

Table 2. Comparison of methods `Emp`, `CvxPi`, `ChaoBunge`, `Chao84`, `ChaoLee`, `LB` defined in Section 6.6, for estimating N: relative bias, relative standard error, relative `RMSEP` and non-coverage probabilities ($\times 100\%$) are reported, for $N = 100$ and $N = 5000$, considering $p'_0 = 0.0707$ and $p'_0 = 0.354$.

$N = 100$, $p'_0 = 0.0707$						
	`Emp`	`CvxPi`	`ChaoBunge`	`Chao84`	`ChaoLee`	`LB`
`bias/N`	0.004	0.023	-0.039	-0.014	-0.086	0.520
`se/N`	0.063	0.048	0.034	0.067	0.026	3.120
`RMSEP/N`	0.063	0.049	0.051	0.069	0.090	3.160
$P(B_{\text{inf}} > N)$	0.4	0.2	0.0	2.0	45.0	0.0
$P(B_{\text{sup}} < N)$	7.0	3.7	38.0	6.7	38.0	23.0
$N = 100$, $p'_0 = 0.354$						
	`Emp`	`CvxPi`	`ChaoBunge`	`Chao84`	`ChaoLee`	`LB`
`bias/N`	0.110	0.130	0.074	-0.025	-0.188	0.288
`se/N`	0.130	0.120	4.440	0.160	0.102	0.770
`RMSEP/N`	0.18	0.17	4.44	0.16	0.21	0.82
$P(B_{\text{inf}} > N)$	9.2	9.3	0.7	1.4	96.0	0.0
$P(B_{\text{sup}} < N)$	0.4	0.6	3.3	3.4	2.0	12.0
$N = 5000$, $p'_0 = 0.0707$						
	`Emp`	`CvxPi`	`ChaoBunge`	`Chao84`	`ChaoLee`	`LB`
`bias/N`	0.003	0.005	-0.042	-0.033	-0.049	-0.013
`se/N`	0.0109	0.0070	0.0040	0.0060	0.0040	0.0150
`RMSEP/N`	0.010	0.009	0.043	0.034	0.049	0.019
$P(B_{\text{inf}} > N)$	5.4	5.0	0.0	0.0	0.0	0.0
$P(B_{\text{sup}} < N)$	1.7	1.2	100.0	100.0	100.0	26.0
$N = 5000$, $p'_0 = 0.354$						
	`Emp`	`CvxPi`	`ChaoBunge`	`Chao84`	`ChaoLee`	`LB`
`bias/N`	0.110	0.110	0.110	-0.055	-0.027	-0.055
`se/N`	0.019	0.019	0.040	0.019	0.023	0.055
`RMSEP/N`	0.120	0.120	0.110	0.058	0.035	0.078
$P(B_{\text{inf}} > N)$	100	100	79	0	54	10
$P(B_{\text{sup}} < N)$	0.0	0.0	0.0	74.0	0.1	0.0

Table 3. For each estimation methods, the first line gives the estimate of N, and its estimated standard error (parenthesis). The second line gives the 95% estimated confidence intervals. The values of the parameter t are given in Figure 1 for the three last methods. For the `ChaoBunge` method, $t = 10$ for the Butterfly and Bird data sets, and $t = 3$ for the two others (pointed out with *). The confidence interval could not be calculated for the `pcg` method because of convergence difficulties. Methods `LB` to `unpmle` do not provide standard errors.

	Microbial	Butterfly	Bird	Tomato
`Emp`	1211 (47.8) [1117, 1305]	782 (26.6) [730, 834]	82 (8.1) [66, 98]	4439 (92.7) [4257, 4621]
`CvxPi`	1211 (47.8) [1117, 1305]	782 (26.6) [730, 834]	87 (6.5) [70, 95]	4439 (92.7) [4257, 4621]
`CvxB`	1211 (49.2) [1117, 1306]	782 (24) [734, 825]	87 (6.5) [70, 96]	4439 (93.7) [4250, 4617]
`ChaoBunge`	2269* (854) [1213, 4821]	757 (32.4) [698, 826]	80 (5) [72, 92]	7166* (1157) [5330, 9947]
`ChaoLee`	2511 (391) [1878, 3434]	737 (24) [693, 787]	80 (4.7) [72, 91]	9554 (1034) [7778, 11858]
`Chao84`	1631 (183) [1117, 1305]	714 (22.7) [730, 834]	77 (4) [71, 96]	5888 (340) [4257, 4621]
`LB`	3987 [915, 7060]	754 [629, 878]	78 [65, 93]	11520 [7047, 15993]
`pcg`	3000	744	86	13960
`pnpmle`	2035 [1523, 2758]	724 [686, 843]	79 [73, 100]	7257 [5899, 9167]
`unpmle`	2169 [1620, $> 10^6$]	722 [687, 913]	76 [74, 86]	7417 [6009, $> 10^7$]