ESTIMATING A PROBABILITY MASS FUNCTION WITH UNKNOWN LABELS[1]

BY DRAGI ANEVSKI[2], RICHARD D. GILL AND STEFAN ZOHREN[3]

Lund University, Leiden University and University of Oxford

In the context of a species sampling problem, we discuss a nonparametric maximum likelihood estimator for the underlying probability mass function. The estimator is known in the computer science literature as the high profile estimator. We prove strong consistency and derive the rates of convergence, for an extended model version of the estimator. We also study a sieved estimator for which similar consistency results are derived. Numerical computation of the sieved estimator is of great interest for practical problems, such as forensic DNA analysis, and we present a computational algorithm based on the stochastic approximation of the expectation maximisation algorithm. As an interesting byproduct of the numerical analyses, we introduce an algorithm for bounded isotonic regression for which we also prove convergence.

1. Introduction. Assume we have a random sample that is drawn from an infinite population of species. The goal of this paper is to, based on the random sample, estimate the unknown relative frequencies of all the species in the population.

Probably the most well-known estimator in the context of species sampling is the *naive estimator*, which is the vector of relative frequencies of the species observed in the sample. The problem of this estimator is that it assigns zero probability to any new species which have not yet been observed in the sample. However, when the relative frequencies are very small it is very likely that when sampling a new element this will be a new, so far unobserved species. Such a situation arises, for example, in forensic DNA analysis when the Y-STR profile of the suspect is not present in the database. This makes it necessary to go beyond the naive estimator and consider estimators for the unknown relative frequencies of all the species in the population.

The first to have studied problems in this setting is apparently Fisher et al. [9], who assumed that the members of each separate species are caught according to separate Poisson processes with different intensities and allowing for the processes to be dependent.

The first to use a nonparametric approach is Good [10], who presented an approximate formula for the expectation of the population frequency. Good attributes the formula to Alan Turing. His approximation becomes better for larger sample sizes but it is not clear from the results in his paper if the formula is asymptotically correct. As a consequence, he is also able to give an estimate of the coverage, the sum of the population frequencies of the species observed in the sample, leading to what is known as the *Good estimator* or Good Turing estimator for the probability mass of the unobserved species, which is given by the number of species observed exactly once in a sample divided by the sample size. Next, Good and Toulmin [11] study a similar setting but for the case when there is a second sample drawn from the population, which can then be thought of as an enlargement of the original sample. As an application, Efron and Thisted [6] used the result by Fisher et al. [9] and Good and Toulmin [11] to estimate the number of words known by Shakespeare based on the observed word frequencies in his works. Later work has been concerned with the bias, confidence intervals as well as asymptotic normality of the Good estimator (e.g., [7, 8, 27]); see also Mao and Lindsay [15] for an application to DNA analysis in this context.

One sees that the naive estimator and the Good estimator are complementary in the sense that the former gives an estimate for the probability distribution of the already observed species, while the latter gives an estimate for the total probability mass of all unobserved species. One would like to combine both these estimators and extend the tail of the naive estimator over the region of unobserved species. A proposal for such an estimator has been made in [1, 18–20] for a similar problem in a computer science setting. In [19], they introduced what they call the *high profile estimator* and what we refer to as the *pattern maximum likelihood estimator* (PML) which is explained in detail below. For small models, this estimator can be obtained analytically [1, 19] and for bigger models a Monte Carlo expectation maximisation (EM) algorithm was proposed in [18]. In [20], they have also claimed, without complete proof, consistency results for the PML, and discussed the general problem of modelling and estimation of the distribution over "large alphabets" when there is a small sample. Their work has been the main motivation for the research presented here. In particular, our goals have been to give a full consistency proof, as well as an extension of their model together with its numerical implementation.

We can state the basic estimation problem of the high profile estimator or PML in a simplified manner as follows: Given $N_1, \ldots, N_K$, a set of absolute frequencies, N_i denoting the number of times a species i is observed, and ordered (by us) in decreasing order. There is another order, provided by nature, which orders the species in how frequent they are in nature, modelled by a set of decreasing probabilities $\theta_1, \theta_2, \ldots$ that sum to one, where θ_α denotes how frequent the αth most frequent species is. We can view our data N as an ordering of an underlying data set $X_{\alpha_1}, \ldots, X_{\alpha_k}$ (for some indices $\alpha_i, i = 1, \ldots, n$). There is an unobserved map, which takes the order provided by us to the order provided by nature, which we

denote by χ and which is a bijection. We will derive the likelihood for θ based on the data N for this problem, and define the PML of θ as the maximizer of that likelihood under the assumptions $\theta_1 \geq \theta_2 \geq \cdots, \sum \theta_i = 1$. However, typically, and with high probability, the PML $\hat{\theta}$ will not exist in the above model.

Therefore, besides the above described, *basic model*, we also consider an *extended model* which, in addition to the discrete probability part, also includes a continuum probability mass part. Then $\theta = (\theta_1, \theta_2, \dots)$, corresponding to the the discrete part of the distribution, only satisfies $\sum_{\alpha} \theta_{\alpha} \leq 1$, where the remaining probability mass $\theta_0 = 1 - \sum_{\alpha} \theta_{\alpha}$ belongs to the continuum part, the blob. We will derive the likelihood in this extended model and define the PML $\hat{\theta}$ as the maximizer under the assumptions $\theta_1 \geq \theta_2 \geq \cdots, \sum_{\alpha=1}^{\infty} \theta_{\alpha} \leq 1$. In Section 3, we state the existence of the PML $\hat{\theta}$ in the extended model, and give the proof of this result in [3], Supplement A. Uniqueness is not known.

Both in the basic or extended model one can give a *truncation level* $k = k_n$, and define $\widetilde{\phi} = (\theta_1, \dots, \theta_k)$ as well as $\phi_0 = 1 - \sum_{\alpha=1}^{k} \theta_{\alpha}$. Such a truncated model we call a *sieved model*. Analogous to the standard PML, one can write down a likelihood function for the sieved model and from this a PML, the so-called sieved PML. The introduction of the sieved PML (sPML) is novel and as discussed below is important for many applications.

The main theoretical results in the paper are almost sure consistency in an L^1-norm for the PML and sieved PML. In this connection, the Hardy–Littlewood–Polya monotone rearrangement algorithm [12] is interesting for two reasons. The first reason is that the algorithm is prominent in our proof of the consistency result, since a naive estimator of the probability mass function can be seen as a monotone rearrangement of the empirical probability mass function. In the proof, we need a certain contraction or nonexpansivity property of the algorithm cf. [2, 14]. Another result is the almost sure rate of convergence which is almost of the order $n^{-1/4}$ for both the standard and sieved PML, which should be compared with the rate for the naive estimator, for which Jankowski and Wellner [13] have obtained the rate $n^{-1/2}$, but then in distribution of norms, and furthermore for which we derive the almost sure rate in supnorm distance of order almost $n^{-1/2}$; cf. Section 3.

An important question is how to calculate the estimator. The main practical result is a stochastic approximation expectation maximisation (SA-EM) algorithm for the sieved estimator, where we use the EM algorithm to get a numeric approximation, treating the bijection χ as a latent variable; this is presented in [3], Supplement B. In this algorithm, in the M step, assuming given χ, we will use isotonic regression. We develop a modification of the standard PAVA algorithm for isotonic regression (cf. Robertson et al. [23]), to allow for lower bounds on the unknown frequencies, in [3], Supplement C. The paper is organized as follows: In Section 2, we introduce the model, the data that arise in this type of problem and the possible ways to estimate the probability mass function. In Section 3, we state the existence result for the PML. In Section 4, we discuss consistency of the

nonparametric maximal likelihood estimators: In Section 4.1, we will study an extended maximum likelihood estimator in the basic model, proving its consistency, and deriving rates for the consistency result. In Section 4.2, we derive similar consistency results for the sieved estimator. In Section 4.3, we discuss the consistency results that we obtained in the previous two subsections and compare them with the results for the naive estimator obtained by Jankowski and Wellner [13]. We conclude with a discussion in Section 5. In [3], Supplement A, we prove existence of the PML. In [3], Supplement B, we present the SA-EM algorithm for computing the PML. In [3], Supplement C, we derive the MLE of a decreasing multinomial probability mass function bounded below by a known constant.

2. The model, the data and the estimators.

2.1. *Introduction.* Imagine an area inhabited by a population of animals which can be classified by species. Which species live in the area (many of them previously unknown to science) is a priori unknown. Let $\mathcal{A}$ denote the set of all possible species potentially living in the area. For instance, if animals are identified by their genetic code, then the species' names α are equivalence classes of DNA sequences. The set of all possible DNA sequences is effectively uncountably infinite, and for present purposes so is the set of equivalence classes, each equivalence class defining one potential species.

Suppose that animals of species $\alpha \in \mathcal{A}$ form a fraction $\theta_\alpha \geq 0$ of the total population of animals. We assume that the probabilities θ_α are unknown. The *basic model* studied in this paper assumes that $\sum_{\alpha:\theta_\alpha>0} \theta_\alpha = 1$ but we shall also study an *extended model* in which it is allowed that (the discrete part of the distribution) $\sum_{\alpha:\theta_\alpha>0} \theta_\alpha < 1$. In either case, the set of species with positive probability is finite or at most countably infinite.

Imagine now an ecologist taking an i.i.d. random sample of n animals, one at a time. The jth animal in the sample belongs to species α with probability θ_α. For each animal in turn, the ecologist can only determine whether it belongs to the same species as an earlier animal in the sample, or whether it is the first representative in his sample of a new species. Suppose he labels the different species observed in the sample by their number in order of discovery. His data can then be represented as a string of n integers, where the jth integer equals r if and only if it belongs to the rth different species observed in the sample in order of discovery. For instance, for $n = 5$, the observed data could be the string 12231 meaning that the first, second and fourth animals in the sample belonged to new species; the third and the fifth were each occurrences of a previously observed species, namely the same as that of the second and first animal in the sample, respectively.

2.2. *Estimation in the extended model.* Since we treat the α as unknown, the parameter $(\theta_\alpha : \alpha \in \mathcal{A})$ is not identified. Since everything only depends on the ordered list of probabilities θ_α it is convenient to change notation and from now on

refer to species by their position in this ordering. If there is only a finite number of species of positive probability, then we will append to the list a countable number of possibly fictitious species each of probability zero. We redefine $\mathcal{A} = \mathbb{N} = \{1, 2, \ldots\}$ and redefine θ_α, where α is a positive integer, as the probability of the αth most frequent species in the population. We define the *deficit* $\theta_0 = 1 - \sum_{\alpha \geq 1} \theta_\alpha$. In the basic model, $\theta_0 = 0$, in the extended model $\theta_0 \geq 0$.

In the extended model, the deficit θ_0 equals the probability, when we observe just one animal, that it belongs to one of those species which individually each have zero probability. Each such species can only be observed at most once in a sample of n animals. The converse is not true: if an animal is observed only once in our sample, we do not know whether it belongs to a zero probability species or to a positive probability species.

We will discuss estimation in the extended model and in a truncated, or sieved, version of the extended model.

Let $\aleph$ be the total number of species of positive probability. If $\aleph < \infty$, we take $\theta_\alpha = 0$ for $\alpha > \aleph$. Thus, from now on $\mathcal{A} = \mathbb{N} = \{1, 2, \ldots\}$, and $\theta = (\theta_1, \theta_2, \ldots)$ where the θ_α, the probability of occurrence of an animal belonging to the αth most frequent species in the population, are nonnegative and nonincreasing and sum to 1.

Since our random sample of n animals is i.i.d., it can be further reduced, by sufficiency, to the *partition*, in the number-theoretic sense, of the integer n which it induces. This is a list $N = (N_1, N_2, \ldots)$ where $N_i \geq 0$ is the number of observed animals belonging to the ith most frequent species in the sample, $N_i \geq 0$, $N_1 \geq N_2 \geq \cdots$, and $\sum_i N_i = n$. The number K of different species of animals observed in the sample, is finite: for some $K \geq 0$, $N_K > 0$ and $N_i = 0$ for $i > K$. In the number-theoretic sense of the word, N (more precisely, the positive part of N, of length K) is a random *partition* of the number n. For instance, the string 12231 corresponds to the partition $N = (2, 2, 1)$ of the integer 5, meaning that two species were each observed twice and one species was observed once; $2 + 2 + 1 = 5$. It is convenient to append an infinite list of zero counts to N. In our example, we then write $N = (2, 2, 1, 0, 0, \ldots)$.

Both the data N and unknown parameter θ are represented by infinite lists of nonincreasing nonnegative numbers, summing to n and 1, respectively; the elements of N are moreover integers. However, there is no direct connection between the indices of the two lists. There exists a map χ from $\mathbb{N}$ (the species as ordered by the sample frequencies) to $\mathcal{A}$ (the species as ordered by population probabilities), defined by $\chi(i) = \alpha$ if and only if the ith most frequent species in the sample is the αth most frequent species in the population, with the tie-breaking rule $N_i = N_j$ implies $\chi(i) < \chi(j)$. The map χ is random, and the essential feature of our model is that χ is not observed.

Let us use the same symbol N to denote both the observed partition of sample size n thought of as a random sequence, as well as the possible sample values thereof. After reduction by sufficiency, the sample space is the set of all possible

partitions N of the sample size n. Write $P^{(n,\theta)}$ for the corresponding (discrete) probability measure on the sample space when the underlying parameter is θ. The basic model states that for any set A of partitions of n:

$$(1) \qquad P^{(n,\theta)}(A) = \sum_{(N_1,N_2,\ldots) \in A} \begin{pmatrix} n \\ N_1 \quad N_2 \quad \ldots \end{pmatrix} \sum_\chi \prod_i \theta_{\chi(i)}^{N_i}.$$

The likelihood function for θ based on the data N is therefore

$$(2) \qquad \mathrm{lik}(\theta) = \sum_\chi \prod_i \theta_{\chi(i)}^{N_i} = \sum_\chi \prod_\alpha \theta_\alpha^{N_{\chi^{-1}(\alpha)}}.$$

The maximum likelihood estimator (MLE) of θ is defined as

$$(3) \qquad \widehat{\theta} = \arg \max_{\theta:\theta_1 \geq \theta_2 \geq \cdots, \sum_{\alpha=1}^\infty \theta_\alpha = 1} \mathrm{lik}(\theta).$$

It is interesting to note that the likelihood (2) can be interpreted as a matrix permanent of the nonnegative matrix $M_{ij} := \theta_i^{N_j}$. This relation enables one to use several techniques of approximate inference to evaluate the likelihood [25, 26]. We will not pursue this idea further here. This is mainly because we are interested in the extended model, where a relation to matrix permanents is more involved.

Returning to the MLE, it is not clear that $\widehat{\theta}$ exists nor that it is unique. In fact, it is easy to exhibit observed data N for which it does not exist; for instance, with $n = 2$, the partition $N = (1, 1)$; see [3], Supplement A for the simple demonstration. For this reason, we study instead the extended model MLE. Define the extended model MLE or the *Pattern Maximum Likelihood estimator* (PML) as

$$(4) \qquad \widehat{\theta} = \arg \max_{\theta:\theta_1 \geq \theta_2 \geq \cdots, \sum_{\alpha=1}^\infty \theta_\alpha \leq 1} \sum_\chi \frac{n!}{N_0! \prod_{i \geq 1} N_i!} \theta_0^{N_0} \prod_{\alpha=1}^\infty \theta_\alpha^{N_{\chi^{-1}(\alpha)}},$$

with $N_0 = n - \sum_{\alpha=1}^\infty N_{\chi^{-1}(\alpha)}$ and $\theta_0 = 1 - \sum_{\alpha \geq 1} \theta_\alpha$. The mappings $\chi : \mathbb{N} \to \{0, 1, \ldots, \infty\}$ satisfy that for every $\alpha \geq 1$ there exists exactly one i such that $\chi(i) = \alpha$, with the tie-breaking rule $N_i = N_j$ implies $\chi(i) < \chi(j)$, and such that $\chi(i) = 0$ implies $N_i = 0$ or 1. Note that since the data ends in a block of 1's, with N_0 of them belonging to blob species, $\sum_{i \geq 0} N_i = n + N_0$. Furthermore $n!/N_0! \prod_{i:\chi(i) \neq 0} N_i! = n!/N_0! \prod_{i > 0} N_i!$, since $1! = 1$. According to Theorem 1 in [21], it is true in this *extended* model that a maximum likelihood estimator does exist; moreover they claim in Corollary 5 that the support of the PML (the number of indices for which $\hat{\theta}_\alpha$ is positive) is finite. We prove that the PML $\hat{\theta}$ exists in Section 3, although the uniqueness is not known. The probability measure corresponding to a possibly defective probability ϕ is given by, for any set A of partitions of n,

$$(5) \qquad P^{(n,\phi)}(A) = \sum_{(N_1,N_2,\ldots) \in A} \sum_\chi \frac{n!}{N_0! \prod_{i \geq 1} N_i!} \phi_0^{N_0} \prod_{\alpha=1}^\infty \phi_\alpha^{N_{\chi^{-1}(\alpha)}},$$

with $N_0 = n - \sum_{\alpha=1}^{\infty} N_{\chi^{-1}(\alpha)}$ and $\phi_0 = 1 - \sum_{\alpha \geq 1} \phi_\alpha$.

The underlying permutation of species generated by our finite sample of animals is not observed. Had it been observed, we would have access to full data counts $X = (X_\alpha : \alpha \in A)$. Here, $X_\alpha = N_{\chi^{-1}(\alpha)}$ is the number of occurrences of species α in the sample. This "underlying data" has the multinomial distribution with parameters n and θ.

For any summable list of nonnegative numbers $a = (a_1, a_2, \ldots)$, denote by $T(a)$ the *monotone rearrangement map* which rewrites the components of a in decreasing order. The relation between the actually observed N and the underlying data X is very simply $N = T(X)$.

To the underlying multinomial count vector X, we associate the empirical cumulative distribution function $F^{(n)}$ of the observed animals' true species label-numbers α, defined by $F^{(n)}(x) = n^{-1} \sum_{\alpha \leq x} X_\alpha$. Alongside this, we define the empirical probability mass function $f^{(n)}$, thought of as a vector or list rather than a function, $f_\alpha^{(n)} = X_\alpha/n = F^{(n)}(\alpha) - F^{(n)}(\alpha - 1)$. Finally, we define

$$\widehat{f}^{(n)} = N/n = T(f^{(n)})$$

the *naive estimator* of θ. The two ways we have expressed it, show that it is simultaneously the *ordered empirical* probability mass function of the *underlying* data, as well as being a *statistic* in the strict sense—a function of the actually observed data N.

The naive estimator $\widehat{f}^{(n)}$ of θ is a random element on our sample space of random partitions. Our main tool in proving L_1 consistency of the PML $\widehat{\theta}$ will be finding an observable event A, that is, a subspace of the set of all possible sample outcomes, which has large probability under $P^{(n,\theta)}$, where θ is the true value of the parameter, but small probability under $P^{(n,\phi)}$, for all ϕ outside of a small L_1 ball around θ. This event A will be defined in terms of $\widehat{f}^{(n)}$ and of the true parameter θ; in fact, it will be the event that $\widehat{f}^{(n)}$ lies within a certain small L_∞ ball around θ. Since this true value of θ is fixed, even if unknown to the statistician, there is no problem in using its value in the definition of the event A.

2.3. *Sieved estimation in the extended model.* In applications, maximization of the likelihood can be computationally very demanding. In the extended model, the parameter $\theta = (\theta_1, \theta_2, \ldots)$ satisfies $\sum_\alpha \theta_\alpha \leq 1$, and the total probability in the blob is $\theta_0 = 1 - \sum_{\alpha \geq 1} \theta_\alpha$. Whenever an animal is drawn from "the blob", it represents a new species in the sample, which is only observed exactly once. Thus, when $\theta_0 > 0$ and n is large, the observed partition N tends to terminate in a long sequence of components N_i all equal to 1, many if not most of them—in the long run, on average $\theta_0 n$ of them—corresponding to species in the blob.

A possibly clever strategy for the *basic* model would be to truncate the vector θ at some finite number of components. If, however, the true ordered probability mass function θ has a very slowly decreasing tail, truncation at too low a level

might badly spoil the estimate. This possibility can be made less harmful by not truncating the original model, but truncating the extended model. Thus, the parameter is taken to be $\widetilde{\theta} = (\theta_1, \ldots, \theta_k)$ where $k < \infty$ and $\sum_{\alpha=1}^{k} \theta_\alpha \leq 1$, and the probability deficit $\theta_0 = 1 - \sum_{\alpha=1}^{k} \theta_\alpha$ is supposed to be spread "infinitely thinly" over "continuously many" remaining species.

These considerations lead to the idea of a sieved maximum likelihood estimator which we denote the *sieved PML estimator* (sPML), in which we maximize the probability of the data over probability measures corresponding to a slightly different model from the true model, and indexed by a slightly different parameter: the model is both extended (to allow a blob) and truncated (θ has finite length).

For given true parameter θ of basic or of extended model, and given truncation level $k = k_n$, define $\widetilde{\theta} = (\theta_1, \ldots, \theta_k)$ and define $\theta_0 = 1 - \sum_{\alpha=1}^{k} \theta_\alpha$. In general, $\widetilde{\phi}$ will denote a possibly defective probability mass function on $\{1, \ldots, k\}$ where $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_k$, and $\phi_0 = 1 - \sum_{\alpha=1}^{k} \phi_\alpha$ will denote its deficit. Such parameters correspond to what we call the *sieved model*.

Imagine the sieved model to be true. For any $i \in \mathbb{N}$, the species corresponding to the observed count $N_i \geq 0$ is either one of the species $\alpha = 1, \ldots, k$, or it is one of the species lumped together in the blob. The latter can only be the case if $N_i = 1$ or 0. Different i can both correspond to species in the blob, but cannot correspond to the same species in $1 \leq \alpha \leq k$. We denote this mapping from $\mathbb{N}$ to $\{0, 1, \ldots, k\}$ by χ. It is a surjection on N with tie-breaking defined as above. Moreover, $\chi(i) = 0$ implies $N_i = 1$ or 0. Apart from this, it is arbitrary and not observed.

Again we can imagine the full data which we would have had, if we had observed χ. According to the sieved model, there is an underlying $X = (X_0, X_1, \ldots, X_k)$ which has the multinomial distribution with parameters n and $(\phi_0, \widetilde{\phi})$. To the "proper part" of X, that is to say, $(X_1, X_2, \ldots, X_k)$, corresponds a partition of $X_+ = \sum_{\alpha=1}^{k} X_\alpha$. Denote this partition by $N_+ = (N_1, N_2, \ldots, N_J)$. Thus, $J = \#\{1 \leq \alpha \leq k : X_\alpha > 0\}$ and $N_1 \geq N_2 \geq \cdots \geq N_J > 0$. Alongside these, X_+ animals of $J \leq k$ species from the set $\{1, \ldots, k\}$, we also observed X_0 animals each of different species, where each of those species separately has probability 0, but all such species together have probability ϕ_0. The observed data, finally, is the partition $N = (N_1, N_2, \ldots, N_J, 1, \ldots, 1)$ of n, in which we have appended exactly X_0 1's to the partition N_+ of X_+.

Note that a number of the N_i in the partition of X_+ can also equal 1. In the observed data N, we cannot see how its block of 1's should be split between species inside and outside the blob.

We can now write down the "sieved likelihood" and hence define the sPML estimator:

$$(6) \qquad \text{lik}(\widetilde{\phi}) = \sum_{\chi} \frac{n!}{N_0! \prod_{i \geq 1} N_i!} \phi_0^{N_0} \prod_{\alpha=1}^{k} \widetilde{\phi}_\alpha^{N_{\chi^{-1}(\alpha)}},$$

$$(7) \qquad \widehat{\phi} = \arg \max_{\widetilde{\phi}: \widetilde{\phi}_1 \geq \widetilde{\phi}_2 \geq \cdots \geq \widetilde{\phi}_k, \sum_{\alpha=1}^{k} \widetilde{\phi}_\alpha \leq 1} \text{lik}(\widetilde{\phi}),$$

with $N_0 = n - \sum_{\alpha=1}^k N_{\chi^{-1}(\alpha)}$ and $\phi_0 = 1 - \sum_{\alpha=1}^k \tilde{\phi}_\alpha$. The mappings $\chi : \mathbb{N} \to \{0, 1, \ldots, k\}$ in the sum in (6) have the properties that for every $1 \le \alpha \le k$ there exists exactly one i such that $\chi(i) = \alpha$, with tie-breaking defined by $N_i = N_j$ implies $\chi(i) < \chi(j)$, while $\chi(i) = 0$ implies $N_i = 0$ or 1. It follows that the number of i such that $N_i \ge 2$ cannot exceed k.

Our strategy will again be to find an event A such that A has large probability under the true parameter but small probability under all parameters some distance from the truth. We do have to carefully distinguish between two different "true" probability measures: the law of the data within the sieved model, under the sieved parameter $\tilde{\theta}$ corresponding to the truth, and the law of the data under the original, true model.

3. Existence of the pattern maximum likelihood estimator.

In this section, we state an existence result for the PML estimator over an (extended) parameter space of ordered probability mass distributions in which we allow for a continuous part, the blob. We show existence by showing that this parameter space is compact, in an appropriate metric, and that the likelihood is a continuous functional with respect to this metric.

Recall that the extended parameter space Θ consists of sequences $\theta = (\theta_\alpha : \alpha \in \mathcal{A})$ where $\mathcal{A} = \mathbb{N} = \{1, 2, \ldots\}$, and where $\theta_\alpha \ge 0$ for all α, and moreover $\theta_1 \ge \theta_2 \ge \cdots$ and $\sum_\alpha \theta_\alpha \le 1$.

We give Θ the topology of pointwise convergence. Thus, for $\theta^{(m)}, \theta \in \Theta$, $\theta^{(m)} \to \theta$ as $m \to \infty$ if and only if $\theta_\alpha^{(m)} \to \theta_\alpha$ for all α.

THEOREM 1. (i) *Under the topology of pointwise convergence, the parameter space Θ is compact.* (ii) *The functional $L : \Theta \mapsto \mathbb{R}_+$ defined by*

$$L(\theta) = \sum_\chi \frac{n!}{N_0! \prod_{i \ge 1} N_i!} \theta_0^{N_0} \prod_{\alpha=1}^\infty \theta_\alpha^{N_{\chi^{-1}(\alpha)}}$$

with $N_0 = n - \sum_{\alpha=1}^\infty N_{\chi^{-1}(\alpha)}$, is continuous.

Thus, the extended model pattern maximum likelihood estimator, defined in (4), *exists.*

4. Consistency results.

4.1. *Consistency for the PML estimator.* In this section, we prove the consistency of the PML estimator in the extended model defined in (4), based on a sample from the distribution P. From our result of the previous section, we know that there exists a PML. Uniqueness is not known; however, our results below hold for any PML, and in the sequel we let $\hat{\theta}$ denote *any* PML.

The idea of the proof is to first exhibit a sequence of events A_n for which the $P^{(n,\theta)}$-probability is large (converges to 1 as $n \to \infty$), and such that for all probabilities $P^{(n,\phi)}$ such that ϕ is an L_1-distance δ away from θ, the $P^{(n,\phi)}$-probability is small (goes to zero as $n \to \infty$). This is done in Lemma 1.

As a consequence, we show that the $P^{(n,\theta)}$-probability of $\{\frac{dP^{(n,\phi)}}{dP^{(n,\theta)}} > 1\}$ is small (goes to zero as $n \to \infty$), by intersecting with A_n, for all ϕ that are L_1-distance more than δ away form θ. On the other hand $\frac{dP^{(n,\hat{\theta})}}{dP^{(n,\theta)}} > 1$, if $\hat{\theta}$ is the ML estimator, for every ordered sample $(n_1, \ldots, n_k)$ with fixed $n = n_1 + \cdots + n_k$. Finally, we use an asymptotic formula for the number $p(n)$ of such $(n_1, \ldots, n_k)$, due to Ramanujan and Hardy, to make the argument uniform over every such sample, to show that $\hat{\theta}$ must be within L_1-distance of δ to θ with a large probability (that goes to one as $n \to \infty$), that is, that $\hat{\theta}$ is weakly consistent. This is the content of Theorem 2.

Using the bound established in Theorem 2, we obtain almost sure consistency of $\hat{\theta}$, in Corollary 1. Finally, in Theorem 2 and Corollary 2, we derive rates of the almost sure convergence of the L_1 norm over classes of probability mass functions with tail conditions.

Let θ be a fixed proper distribution. For $\delta > 0$ arbitrary define the class of (possibly defective) probability mass functions $\mathbb{Q}_{\theta,\delta} = \{\phi : \|\phi - \theta\|_1 \geq \delta\}$, where $\|\phi - \theta\|_1 = \sum_{i=1}^{\infty} |\phi_i - \theta_i|$. Note that ϕ is a possibly defective probability in the sense that $\sum_{i=1}^{\infty} \phi_i \leq 1$, and note that in this case we use (5) as the measure.

LEMMA 1. *Let $f^{(n)}$ be the empirical probability mass function based on a sample $x_1, \ldots, x_n$ from some fixed decreasing probability mass function θ, and $\hat{f}^{(n)} = T(f^{(n)})$. Then there is a finite $r = r(\delta, \theta)$ and $\varepsilon = \delta/(8r)$ such that*

$$P^{(n,\theta)}\Big(\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \theta_x| \leq \varepsilon\Big) \geq 1 - 2e^{-n\varepsilon^2/2},$$

$$\sup_{\phi \in \mathbb{Q}_{\theta,\delta}} P^{(n,\phi)}\Big(\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \theta_x| \leq \varepsilon\Big) \leq 2e^{-n\varepsilon^2/2}.$$

PROOF. Let θ be fixed and $\delta > 0$ fixed but arbitrary, and choose an arbitrary $\phi \in \mathbb{Q}_{\theta,\delta}$. Since θ sums to one, there is an $r = r(\theta, \delta)$ such that $\sum_{i=r+1}^{\infty} \theta_i \leq \delta/4$. Then

(8)
$$\sum_{i=1}^{r} |\theta_i - \phi_i| \geq \frac{\delta}{4}.$$

To show (8), note that either $\sum_{i=r+1}^{\infty} \phi_i$ is smaller or larger than $\delta/2$: (i) Assume first that $\sum_{i=r+1}^{\infty} \phi_i \leq \delta/2$. Then

$$\delta \leq \sum_{i=1}^{r} |\theta_i - \phi_i| + \sum_{i=r+1}^{\infty} |\theta_i - \phi_i|$$

$$\leq \sum_{i=1}^{r} |\theta_i - \phi_i| + \sum_{i=r+1}^{\infty} \theta_i + \sum_{i=r+1}^{\infty} \phi_i$$

$$\leq \sum_{i=1}^{r} |\theta_i - \phi_i| + \frac{\delta}{4} + \frac{\delta}{2},$$

which implies (8). (ii) Assume instead that $\sum_{i=r+1}^{\infty} \phi_i > \delta/2$, and write the assumptions as $\sum_{i=1}^{r} \theta_i > 1 - \delta/4$ and $\sum_{i=1}^{r} \phi_i = \sum_{i=1}^{\infty} \phi_i - \sum_{i=r+1}^{\infty} \phi_i \leq 1 - \delta/2$. Then

$$\sum_{i=1}^{r} |\theta_i - \phi_i| \geq \sum_{i=1}^{r} (\theta_i - \phi_i)$$

$$> 1 - \frac{\delta}{4} - 1 + \frac{\delta}{2}$$

$$= \frac{\delta}{4},$$

which again implies (8).

From (8) follows that for some $i \leq r$, we have

(9)
$$|\theta_i - \phi_i| \geq \frac{\delta}{4r} := 2\varepsilon = 2\varepsilon(\delta, \theta).$$

Note that r, and thus also ε depends only on θ, and not on ϕ.

Recall the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [5, 16]; for every $\varepsilon > 0$,

(10)
$$\mathbb{P}_\theta \left(\sup_{x \geq 0} |F^{(n)}(x) - F_\theta(x)| \geq \varepsilon \right) \leq 2e^{-2n\varepsilon^2},$$

where F_θ is the cumulative distribution function corresponding to θ, and $F^{(n)}$ is the empirical probability function based on i.i.d. data from F_θ. Since $\{\sup_{x \geq 0} |F^{(n)}(x) - F_\theta(x)| \geq \varepsilon\} \supset \{\sup_{x \geq 0} |f_x^{(n)} - \theta_x| \geq 2\varepsilon\} \supset \{\sup_{x \geq 1} |f_x^{(n)} - \theta_x| \geq 2\varepsilon\}$, with $f^{(n)}$ the empirical probability mass function corresponding to $F^{(n)}$, equation (10) implies

(11)
$$P^{(n,\theta)} \left(\sup_{x \geq 1} |f_x^{(n)} - \theta_x| \geq \varepsilon \right) = \mathbb{P}_\theta \left(\sup_{x \geq 1} |f_x^{(n)} - \theta_x| \geq \varepsilon \right)$$

$$\leq 2e^{-n\varepsilon^2/2}.$$

Let T be the monotone rearrangement map; cf. [14]. Then the map T is a contraction in the supnorm metric on $\mathbb{N}$, that is, if f, g are two functions $\mathbb{N} \to \mathbb{R}$ and $\|f\|_\infty = \sup_{k \geq 1} |f(k)|$ is the supnorm metric, then $\|T(f) - T(g)\|_\infty \leq \|f - g\|_\infty$, cf. [2] (see also [14] for a proof of the contraction property for L^p-norms). Noting

that $T(\theta) = \theta$ since θ is decreasing by assumption, and with $\hat{f}^{(n)} = T(f^{(n)})$, this implies that

$$\|\hat{f}^{(n)} - \theta\|_\infty \leq \|f^{(n)} - \theta\|_\infty,$$

so that $\{\|\hat{f}^{(n)} - \theta\|_\infty \geq \varepsilon\} \subset \{\|f^{(n)} - \theta\|_\infty \geq \varepsilon\}$, and thus by (11)

(12)
$$P^{(n,\theta)}\Big(\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \theta_x| \geq \varepsilon\Big) \leq P^{(n,\theta)}\Big(\sup_{x \geq 1} |\hat{f}_x^{(n)} - \theta_x| \geq \varepsilon\Big)$$
$$\leq 2e^{-n\varepsilon^2/2}.$$

For an analogue argument for a sample from the (possibly defective) distribution $\phi = (\phi_1, \phi_2, \ldots)$, we first append the mass point $\phi_0 = 1 - \sum_{x=1}^\infty \phi_x$ to this vector to obtain a corresponding (proper) distribution function F_ϕ. Using the corresponding cumulative empirical distribution $F^{(n)}$, and probability mass function $f^{(n)}$, and sorted such $\hat{f}^{(n)} = T(f^{(n)})$ we again have a contraction in the application of T, and going via the DKW inequality, we obtain [recall (5)]:

$$P^{(n,\phi)}\Big(\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \phi_x| \geq \varepsilon\Big) \leq 2e^{-n\varepsilon^2/2},$$

which is equivalent to

(13)
$$P^{(n,\phi)}\Big(\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \phi_x| < \varepsilon\Big) \geq 1 - 2e^{-n\varepsilon^2/2}.$$

Note that

(14)
$$\Big\{\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \phi_x| < \varepsilon\Big\} \cap \{\exists i \leq r : |\theta_i - \phi_i| > 2\varepsilon\}$$
$$\subset \{\exists i \leq r : |\hat{f}_i^{(n)} - \theta_i| > \varepsilon\} = \Big\{\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \theta_x| > \varepsilon\Big\}.$$

Since the second event in (14) is deterministic, for any $\phi \in \mathbb{Q}_{\theta,\delta}$, and with an ε small enough [see (9)], this together with equation (13) implies

$$P^{(n,\phi)}\Big(\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \theta_x| > \varepsilon\Big) \geq P^{(n,\phi)}\Big(\sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \phi_x| < \varepsilon\Big)$$
$$\geq 1 - 2e^{-n\varepsilon^2/2}.$$

Since $\phi \in \mathbb{Q}_{\theta,\delta}$ is arbitrary, the statement of the lemma follows. □

We next derive the almost sure consistency of (any) extended maximum likelihood estimator $\hat{\theta}$. Recall the definitions of $P^{(n,\theta)}$, $P^{(n,\phi)}$ for proper and possibly defective distributions θ and ϕ in (1) and (5), respectively.

THEOREM 2.   *Let $\hat{\theta} = \hat{\theta}^{(n)}$ be (any) extended maximum likelihood estimator. Then for any $\delta > 0$*

$$P^{(n,\theta)}\big(\|\hat{\theta} - \theta\|_1 > \delta\big) \leq \frac{1}{\sqrt{3n}} e^{\pi\sqrt{\frac{2n}{3}} - n\frac{\varepsilon^2}{2}}(1 + o(1)) \qquad as\ n \to \infty,$$

where $\varepsilon = \delta/(8r)$ and $r = r(\theta, \delta)$ such that $\sum_{i=r+1}^{\infty} \theta_i \leq \delta/4$.

PROOF.   Now let $\mathbb{Q}_{\theta,\delta}$ be as in the statement of Lemma 1. Then there is an r such that the conclusion of the lemma holds, that is, for each n there is a set

$$A = A_n = \Big\{ \sup_{1 \leq x \leq r} |\hat{f}_x^{(n)} - \theta_x| \leq \varepsilon \Big\}$$

such that

$$P^{(n,\theta)}(A_n) \geq 1 - 2e^{-n\varepsilon^2/2},$$

$$\sup_{\phi \in \mathbb{Q}_{\theta,\delta}} P^{(n,\phi)}(A_n) \leq 2e^{-n\varepsilon^2/2}.$$

For any $\phi \in \mathbb{Q}_{\phi,\delta}$, we can define the likelihood ratio $dP^{(n,\phi)}/dP^{(n,\theta)}$. Then for any $\phi \in \mathbb{Q}_{\phi,\delta}$,

$$\begin{aligned}
P^{(n,\theta)}\Big(A_n \cap \Big\{ \frac{dP^{(n,\phi)}}{dP^{(n,\theta)}} \geq 1 \Big\} \Big) &= \int_{A_n \cap \{\frac{dP^{(n,\phi)}}{dP^{(n,\theta)}} \geq 1\}} dP^{(n,\theta)} \\
&\leq \int_{A_n} \frac{dP^{(n,\phi)}}{dP^{(n,\theta)}} dP^{(n,\theta)} \\
&= P^{(n,\phi)}(A_n) \\
&\leq 2e^{-n\varepsilon^2/2},
\end{aligned}$$

which implies that

$$\begin{aligned}
P^{(n,\theta)}\Big(\frac{dP^{(n,\phi)}}{dP^{(n,\theta)}} \geq 1 \Big) &= P^{(n,\theta)}\Big(A_n \cap \Big\{ \frac{dP^{(n,\phi)}}{dP^{(n,\theta)}} \geq 1 \Big\} \Big) - P^{(n,\theta)}(A_n) \\
&\quad + P^{(n,\theta)}\Big(A_n \cup \Big\{ \frac{dP^{(n,\phi)}}{dP^{(n,\theta)}} \geq 1 \Big\} \Big) \\
&\leq 2e^{-n\varepsilon^2/2} - 1 + 2e^{-n\varepsilon^2/2} + 1 \\
&= 4e^{-n\varepsilon^2/2}.
\end{aligned}$$

If $\hat{\theta}$ is a PML estimator, then

$$\frac{dP^{(n,\hat{\theta})}}{dP^{(n,\theta)}} \geq 1.$$

For a given $n = n_1 + \cdots + n_k$ such that $n_1 \geq \cdots \geq n_k > 0$, (with k varying), there is a finite number $p(n)$ of possibilities for the value of $(n_1, \ldots, n_k)$. The number $p(n)$ is the partition function of n, for which we have the asymptotic formula:

$$p(n) = \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{\frac{2n}{3}}}(1 + o(1)),$$

as $n \to \infty$, cf. [22]. For each possibility of $(n_1, \ldots, n_k)$, there is a PML estimator (for each possibility we can choose one such) and we let $\mathcal{P}_n = \{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(p(n))}\}$ be the set of all such choices of PML estimators. Then

$$P^{(n,\theta)}(\hat{\theta} \in \mathbb{Q}_{\theta,\delta}) = \sum_{\phi \in \mathcal{P}_n \cap \mathbb{Q}_{\theta,\delta}} P^{(n,\theta)}(\hat{\theta} = \phi)$$

$$\leq \sum_{\phi \in \mathcal{P}_n \cap \mathbb{Q}_{\theta,\delta}} P^{(n,\theta)}\left(\frac{dP^{(n,\phi)}}{dP^{(n,\theta)}} \geq 1\right)$$

$$\leq p(n)4e^{-n\varepsilon^2/2},$$

which completes the proof. $\square$

That a $\hat{\theta}$ is consistent in probability is immediate from Theorem 2, and in fact we have almost sure consistency:

COROLLARY 1. *The sequence of maximum likelihood estimators $\hat{\theta}^{(n)}$ is strongly consistent in L_1-norm, that is,*

$$\lim_{n\to\infty} \|\hat{\theta}^{(n)} - \theta\|_1 \overset{a.s.}{\to} 0$$

as $n \to \infty$.

PROOF. This follows as a consequence of the bound in Theorem 2, by the characterization $X_n \overset{a.s.}{\to} 0 \Leftrightarrow \sum_{n=1}^{\infty} P(|X_n| > \delta) < \infty$ for all $\delta > 0$, since

$$\sum_{n=1}^{\infty} \frac{1}{\sqrt{3}n} e^{-\pi\sqrt{n}(\sqrt{n}\frac{\varepsilon^2}{2} - \sqrt{\frac{2}{3}})} < \infty.$$
$\square$

The above results are for a fixed distribution θ, and the rate depends, via ε on the distribution. The next theorem and corollary make the dependence explicit, and give a rate for the almost sure convergence as a function of the tail behaviour of the distribution.

THEOREM 3. *Let $\varepsilon_0 > 0$ be arbitrary and define*

$$\Theta_{\varepsilon_0} = \left\{\theta : \forall \delta > 0, \exists r \leq \delta/\varepsilon_0 \text{ such that } \sum_{i=r+1}^{\infty} \theta_i < \delta/4\right\}.$$

Then, if $\theta \in \Theta_{\varepsilon_0}$,

$$n^{\alpha} \|\hat{\theta}^{(n)} - \theta\| \overset{a.s.}{\to} 0$$

as $n \to \infty$, for any $\alpha < 1/4$.

PROOF. Let $\alpha > 0$ be an arbitrary constant, to be determined below. From Theorem 2, we get

$$(15) \qquad P^{(n,\theta)}\big(n^{\alpha} \|\hat{\theta}^{(n)} - \theta\|_1 > \delta\big) \leq \frac{1}{\sqrt{3}n} e^{-n^{1/2}\left(n^{1/2}\frac{\delta^2}{128 r^2 n^{2\alpha}} - \pi\sqrt{\frac{2}{3}}\right)}.$$

Since $\delta/r \geq \varepsilon_0 > 0$, the right-hand side of (15) converges to zero, and is summable, if

$$n^{-2\alpha + 1/2} \to \infty,$$

as $n \to \infty$, which is true if $\alpha < 1/4$. □

COROLLARY 2. *Let $\Theta_\kappa = \{\theta : \theta_x = l(x)x^{-\kappa}\}$, for $\kappa > 1$ fixed and with l some function slowly varying at infinity. Then if $\theta \in \Theta_\kappa$ the conclusion of Theorem 3 holds.*

PROOF. Assume that $\theta \in \Theta_\kappa$. Let $\varepsilon_0 > 0$ be fixed, and let $\delta > 0$ be fixed but arbitrary. Then for some r we should have $\sum_{i=r+1}^{\infty} \theta_i < \delta/4$, which is equivalent to

$$r^{-\kappa+1} l_1(r) \leq \frac{\delta}{4} \quad \Leftrightarrow \quad r \geq \left(\frac{\delta}{4}\right)^{1/(1-\kappa)} l_2(\delta),$$

when $\kappa > 1$, where l_1 and l_2 are functions which vary slowly at infinity and zero, respectively. It is possible to take r such that $(\frac{\delta}{4})^{1/(1-\kappa)} l_2(\delta) \leq r < \delta/\varepsilon_0$, thus $\theta \in \Theta_{\varepsilon_0}$. □

4.2. *Consistency for the sPML estimator.* Let $k = k_n$ be a positive integer (truncation level) such that $k_n \to \infty$ when $n \to \infty$, and define the sieve:

$$\widetilde{\Theta}_n = \bigg\{ \tilde{\phi} = (\phi_0, \phi_1, \ldots, \phi_k) \text{ where } \phi_0 = 1 - \sum_{\alpha=1}^{k} \phi_\alpha,$$

$$\text{and } \phi_i > \phi_{i+1}, i = 1, \ldots k - 1 \bigg\}.$$

Note that for each proper distribution $\phi \in \Theta_\kappa$ there is a corresponding sieved distribution $\tilde{\phi} \in \widetilde{\Theta}_n$ with $\phi_0 = \sum_{x=k_n+1}^{\infty} l(x)x^{-\kappa} \sim k_n^{-\kappa+1}$, if $\kappa > 1$.

Assume the random vector $X = (X_0, X_1, \ldots, X_k)$, underlying our observations, has a multinomial distribution with parameters n and $\tilde{\phi}$. Define $J = \#\{\alpha \geq 1 :$

$X_\alpha > 0\}$ and let $(N_1, N_2, \ldots, N_J)$ be a partition of $\sum_{\alpha=1}^{k} X_\alpha$, with $N_1 \geq N_2 \geq \cdots \geq N_J > 0$. Then the observed data is the partition $(N_1, N_2, \ldots, N_J, 1, \ldots, 1, 0, 0, \ldots)$ with $X_0 \geq 0$ (unknown) number of 1's appended after the J'th position. Let $I = \sup\{i : N_i \geq 2\}$. We observe I, the number of species observed at least twice, and we observe $(J - I) + X_0$, the number of species which is only observed once. (We do not observe $J - I$ or X_0.) Note that the number of different species that we have observed frequency counts for is $J + X_0 = \tilde{J}$, and that this number is known. We will let $k = k_n$ grow fast enough with n, so that always $\tilde{J} \leq k$.

Recall that $\chi : \{1, 2, \ldots, \tilde{J}\} \to \{0, 1, 2, \ldots, k\}$ is a (random) map taking the i'th most frequently observed species to its position in the truncated list of species ordered by population frequency, such that all species above the k'th most common are grouped together in a "zero category". We assume that for every α such that $1 \leq \alpha \leq k$ there is exactly one $1 \leq i \leq \tilde{J}$ such that $\chi(i) = \alpha$, with tie-breaking $N_i = N_j$ implies $\chi(i) < \chi(j)$. All other $i \in \{1, \ldots, \tilde{J}\}$ are mapped to the zero category. This means that χ is zero on its complement, so $\chi(\mathcal{I}^c) = 0$. Since $\tilde{J} \leq k$, χ need not be surjective. The number $|\mathcal{I}|$ of observed species that are mapped to an α in $\{1, \ldots, k\}$ is random, although we do know that $|\mathcal{I}| \leq k$.

Define the sieved maximum likelihood estimator:

$$(16) \qquad \hat{\theta}_{(s)}^{(n)} = \arg \max_{\tilde{\phi} \in \tilde{\Theta}_n} \sum_{\chi} \frac{n!}{N_0! \prod_{i \geq 1} N_i!} \phi_0^{N_0} \prod_{\alpha=1}^{k} \phi_\alpha^{N_{\chi^{-1}(\alpha)}},$$

with the sum running over all $\chi : \{1, 2, \ldots, \tilde{J}\} \to \{0, 1, \ldots, k\}$ such that for every α with $1 \leq \alpha \leq k$ there is exactly one $1 \leq i \leq \tilde{J}$ such that $\chi(i) = \alpha$, with tie-breaking rule $N_i = N_j$ implies $\chi(i) < \chi(j)$, and χ is surjective on a subset $\mathcal{I} \subset \{1, 2, \ldots, \tilde{J}\}$, that is, $\chi(\mathcal{I}) = \{1, \ldots, k\}$ and $\chi(\mathcal{I}^c) = 0$, and $N_0 = n - \sum_{\alpha=1}^{k} N_{\chi^{-1}(\alpha)}$.

If χ and $\mathcal{I}$ are arbitrary but fixed we define the "estimator" $f^{(n,\chi)}$ of a probability mass function on $\{0, 1, \ldots, |\mathcal{I}|\}$ by

$$(17) \qquad f^{(n,\chi)}(j) = \begin{cases} \displaystyle\sum_{i \in \mathcal{I}^c} \frac{N_{\chi(i)}}{n}, & \text{for } j = 0, \\[2mm] T\left(\dfrac{N_{\chi(i)}}{n} : i \in \mathcal{I}\right), & \text{for } j \in \{1, \ldots, |\mathcal{I}|\}. \end{cases}$$

This is not a proper estimator, since we can not calculate it only on the basis on our data $(N_1, N_2, \ldots, N_J, 1, \ldots, 1, 0, 0, \ldots)$: The map χ and, therefore, the set $\mathcal{I}$ can not be determined from the sample.

For a given χ, let r_χ be the restriction of a function g on $\{1, 2, \ldots\}$ to the set $\chi(\mathcal{I})$. Define the map T_χ on the set of functions g on $\{1, 2, \ldots\}$ as the concatenation of the map $g \to \sum_{\alpha \in \chi(\mathcal{I})^c} g_\alpha$, with the map composition of T with r_χ, so that

$$T_\chi(g) = \left(\sum_{\alpha \in \chi(\mathcal{I})^c} g_\alpha, T(r_\chi(g))\right).$$

Then

(18)
$$T_\chi : \{\text{pmf on } \{1, 2, \ldots\}\} \mapsto \{\text{pmf on } \{0, 1, \ldots, |\mathcal{I}|\},$$
$$\text{ordered on } \{1, \ldots, |\mathcal{I}|\}\}.$$

If $f^{(n)}$ is the empirical probability mass function, based on a sample $x_1, \ldots, x_n$ of ϕ (cf. Section 2), then

$$f^{(n, \chi)} = T_\chi(f^{(n)}).$$

Furthermore, for every χ, the map T_χ in (18) is a contraction, with the two spaces of probability mass functions equipped with the norms $\|\theta\| = \sup_{x \geq 1} |\theta_x|$ and $\|\theta\| = \sup_{0 \leq x \leq |\mathcal{I}|} |\theta_x|$, respectively. In particular,

(19)
$$\sup_{0 \leq x \leq |\mathcal{I}|} |T_\chi(f^{(n)})_x - T_\chi(\theta)_x| \leq \sup_{x \geq 1} |f_x^{(n)} - \theta_x|.$$

To show (19), note first that $T_\chi(\theta) = (\sum_{\alpha \in \chi(\mathcal{I})^c} \theta_\alpha, \theta(\chi(\mathcal{I})))$, since θ itself is sorted on $\chi(\mathcal{I})$ and, therefore, $T_\chi(\theta) = \theta$ on $\mathcal{I}$. Furthermore, $f^{(n)}$ is mapped to $(\sum_{\alpha \in \chi(\mathcal{I})^c} f_\alpha^{(n)}, T(f^{(n)}(\chi(\mathcal{I}))))$.
Therefore,

$$\sup_{0 \leq x \leq |\mathcal{I}|} |T_\chi(f^{(n)})_x - T_\chi(\theta)_x|$$

$$= \max\left(\left| \sum_{\alpha \in \chi(\mathcal{I})^c} f_\alpha^{(n)} - \sum_{\alpha \in \chi(\mathcal{I})^c} \theta_\alpha \right|, \sup_{1 \leq x \leq |\mathcal{I}|} |T(r_\chi(f^{(n)}))_x - T(r_\chi(\theta))_x| \right)$$

$$\leq \max\left(\left| \sum_{\alpha \in \chi(\mathcal{I})^c} f_\alpha^{(n)} - \sum_{\alpha \in \chi(\mathcal{I})^c} \theta_\alpha \right|, \sup_{x \in \chi(\mathcal{I})} |f_x^{(n)} - \theta_x| \right)$$

$$\leq \max\left(\sup_{x \in \chi(\mathcal{I})^c} |f_x^{(n)} - \theta_x|, \sup_{x \in \chi(\mathcal{I})} |f_x^{(n)} - \theta_x| \right)$$

$$= \sup_{x \geq 1} |f_x^{(n)} - \theta_x|,$$

where the first inequality follows since the restriction of T to any subset, and thus also to $\chi(\mathcal{I})$, *is* a contraction, and the second inequality by the triangle inequality and since the l^1 norm on $\chi(\mathcal{I})^c$ is bounded by the max-norm over $\chi(\mathcal{I})^c$. This shows that (19) holds.

Define next the estimator $\check{f}^{(n)}$ of a probability mass function on the set $\{0, 1, \ldots, I\}$, so on the blob together with the set of species observed at least twice, by

(20)
$$\check{f}^{(n)}(j) = \begin{cases} \sum_{i=I+1}^{k} \dfrac{N_i}{n}, & \text{for } j = 0, \\[2ex] \dfrac{N_j}{n}, & \text{for } j \in \{1, \ldots, I\}. \end{cases}$$

Note that this is a proper estimator. We extend this to an estimator on all of $\{0, \ldots, |\mathcal{I}|\}$ by defining $\check{f}^{(n)}(j) = 0$ for $I < j \leq |\mathcal{I}|$.

We now have the following lemma for the (extended) estimator $\check{f}^{(n)}$.

LEMMA 2. *Let f_n be the empirical probability mass function based on a sample $x_1, \ldots, x_n$ from a fixed decreasing probability mass function θ, and let $\check{f}^{(n)}$ be as defined in* (20). *For $\delta > 0$ arbitrary define the class of probability measures $\mathbb{Q}_{P,\delta} = \{Q : \|Q - P\|_1 \geq \delta\}$. Then there is a finite $r = r(\delta, P)$ and $\varepsilon = \delta/(8r)$ such that*

$$P^{(n,\theta)}\left(\sup_{1 \leq x \leq r} |\check{f}_x^{(n)} - \theta_x| \leq \varepsilon \right) \geq 1 - 2e^{-n(\varepsilon - \frac{1}{n})^2/2},$$

$$\sup_{\phi \in \mathbb{Q}_{\theta,\delta}} P^{(n,\phi)}\left(\sup_{1 \leq x \leq r} |\check{f}_x^{(n)} - \theta_x| \leq \varepsilon \right) \leq 2e^{-n(\varepsilon + \frac{1}{n})^2/2}.$$

PROOF. Let χ and I be the fixed random elements that correspond to the given sample. Recall that χ is unknown and I is known. From Lemma 1, there is an r such that the conclusion of that lemma holds.

We first claim that

$$\sup_{1 \leq x \leq |\mathcal{I}|} |f_x^{(n,\chi)} - \check{f}_x^{(n)}| \leq \frac{1}{n}.$$

To see this, note first that $f^{(n,\chi)}$ and $\check{f}^{(n)}$ are identical on the set of species $\{1, \ldots, I\}$ that are observed at least twice. Since $\check{f}^{(n)}$ is zero on $\{I + 1, \ldots, |\mathcal{I}|\}$ it is enough to show that $f^{(n,\chi)}(j) \leq 1/n$ for $j \in \{I + 1, \ldots, |\mathcal{I}|\}$. But this follows by the construction of $f^{(n,\chi)}$.

Therefore, with $\|f\| = \sup_{1 \leq x \leq k} |f(x)|$ and recalling that $|\mathcal{I}| \leq k$, we have $\|\check{f}^{(n)} - \theta\| \leq \frac{1}{n} + \|f^{(n,\chi)} - \theta\|$ so that

$$\{\|f^{(n,\chi)} - \theta\| \leq \varepsilon\} \subset \left\{ \|\check{f}^{(n)} - \theta\| \leq \varepsilon + \frac{1}{n} \right\},$$

and from Lemma 1, with n large enough that $1/n < \varepsilon$,

$$P^{(n,\theta)}\left(\sup_{1 \leq x \leq r} |\check{f}_x^{(n)} - \theta_x| \leq \varepsilon \right) \geq 1 - 2e^{-n(\varepsilon - \frac{1}{n})^2/2}.$$

Similarly,

$$\{\|\check{f}^{(n)} - \theta\| \leq \varepsilon\} \subset \left\{ \|f^{(n,\chi)} - \theta\| \leq \varepsilon + \frac{1}{n} \right\},$$

so that from Lemma 1

$$\sup_{\phi \in \mathbb{Q}_{\theta,\delta}} P^{(n,\phi)}\left(\sup_{1 \leq x \leq r} |\check{f}_x^{(n)} - \theta_x| \leq \varepsilon \right) \leq 2e^{-n(\varepsilon + \frac{1}{n})^2/2}. \qquad \square$$

We need to get a bound on the total variation distance between the two measures $P^{(n,\theta)}$ and $P^{(n,\tilde{\theta})}$ with θ a parameter and $\tilde{\theta}$ a sieved parameter. In order to get such a bound, we need to make a coupling of the two measures. In particular, the two random partitions $N, \tilde{N}$ of n will be defined on the same probability space.

Therefore, let $\theta = (\theta_1, \ldots, \theta_n)$ with $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{k-1} \leq \theta_k \leq \theta_{k+1} \leq \cdots \leq \theta_n$ be the ordered set of probabilities. Note that the cut-off point defining the sieve is $k = k_n$. The underlying full data is

$$(X_1, \ldots, X_n) \sim \text{Multi}(n, \theta),$$

where the X_i's can be zeros and they need not be ordered. Now let $X_0 = \sum_{i=k+1}^n X_i$ and define the new underlying data $\tilde{X} = (X_0, X_1, \ldots, X_k)$. Then

$$\tilde{X} \sim \text{Multi}(n, \check{\theta}),$$

where

$$\check{\theta} = \left(\sum_{i=k+1}^n \theta_i, \tilde{\theta} \right),$$

$$\tilde{\theta} = (\theta_1, \ldots, \theta_k).$$

Now N is the random partition of n, defined as the ordered $(X_1, \ldots, X_n)$, and $\tilde{N}$ is the random partition of n, defined by the ordered nonzero $X_1, \ldots, X_k$, to which we append a list of 1's of length X_0. Note that N and $\tilde{N}$ are defined on the same probability space. Next, for any set A of partitions on n we define the two measures $P^{(n,\theta)}, P^{(n,\tilde{\theta})}$ by

$$P^{(n,\theta)}(A) = \sum_{(N_1, N_2, \ldots) \in A} \begin{pmatrix} & n & \\ N_1 & N_2 & \cdots \end{pmatrix} \sum_\chi \prod_{i=1}^n \theta_{\chi(i)}^{N_i},$$

$$P^{(n,\tilde{\theta})}(A) = \sum_{(\tilde{N}_1, \tilde{N}_2, \ldots) \in A} \begin{pmatrix} & n & \\ \tilde{N}_1 & \tilde{N}_2 & \cdots \end{pmatrix} \sum_\chi \prod_i \theta_{\chi(i)}^{\tilde{N}_i},$$

in the case that θ is a proper distribution, and similarly if θ is a possibly defective distribution. Note that $P^{(n,\theta)}, P^{(n,\tilde{\theta})}$ have total mass one, and thus are probability measures. There is another measure, $\tilde{P}^{(n,\tilde{\theta})}$ say, not necessarily a probability measure and connected to $P^{(n,\tilde{\theta})}$, that is defined by distributing the sorted nonzero values of $X_1, \ldots, X_k$ to different θ_i's and the value X_0 to the blob θ_0. However, since we are only interested in when the measure $P^{(n,\theta)}$ differs from "the measure" generated by the partition $\tilde{N}$, it will not be of importance which of the two measures $P^{(n,\tilde{\theta})}, \tilde{P}^{(n,\tilde{\theta})}$ we use, and as a matter of fact using a measure with total mass one simplifies the reasoning somewhat, therefore, we will work with $P^{(n,\tilde{\theta})}$.

Now $P^{(n,\theta)}$ and $P^{(n,\tilde{\theta})}$ are the same if and only if all $X_{k+1}, X_{k+2}, \ldots, X_n$ are zero or one, and thus they differ on the set $\bigcup_{i=k+1}^n \{X_i \geq 2\}$. The probability, under

θ, of this is

$$P_\theta\left(\bigcup_{i=k+1}^{n}\{X_i \geq 2\}\right) \leq \sum_{i=k+1}^{n} P_\theta\{X_i \geq 2\}$$

$$\leq \sum_{i=k+1}^{n} \frac{E_\theta(X_i)}{2} = \frac{n}{2}\sum_{i=k+1}^{n}\theta_i,$$

by Markov's inequality.

THEOREM 4. *Let $\hat{\theta}_{(s)}^{(n)}$ be the sieved PML estimator defined in* (16). *Assume the sieve cut-off $k(n)$ satisfies $\sum_{i=k(n)+1}^{n}\theta_i \leq Ce^{-\beta n^{1/2+v}}(1+o(1))$, as $n \to \infty$, for some $v, \beta > 0$. Then for any $\delta > 0$,*

$$P^{(n,\theta)}(\|\hat{\theta}_{(s)}^{(n)} - \tilde{\theta}\|_1 > \delta)$$

$$\leq \frac{1}{2\sqrt{3}n}e^{\pi\sqrt{\frac{2n}{3}}}\left(e^{-n(\varepsilon+\frac{1}{n})^2/2} + e^{-n(\varepsilon-\frac{1}{n})^2/2} + Ce^{-\beta n^{1/2+v}}\right)(1+o(1))$$

as $n \to \infty$, where $\varepsilon = \delta/(8r)$ and $r = r(P,\delta)$ such that $\sum_{i=r+1}^{\infty}\theta_i \leq \delta/4$, and $\|\tilde{\theta} - \tilde{\phi}\|_1 = \sum_{i=1}^{k}|\tilde{\theta}_i - \tilde{\phi}_i|$.

PROOF. Lemma 2 implies that there is a set

$$A_n = \left\{\sup_{1\leq x\leq k_n}|\check{f}_x^{(n)} - \theta_x| \leq \varepsilon\right\}$$

such that

$$P^{(n,\theta)}(A_n) \geq 1 - 2e^{-n(\varepsilon-\frac{1}{n})^2/2},$$

$$\sup_{\phi\in\mathbb{Q}_{\theta,\delta}} P^{(n,\phi)}(A_n) \leq 2e^{-n(\varepsilon+\frac{1}{n})^2/2}.$$

Furthermore, under the assumption of the cut-off level $k(n)$ we have that

$$P^{(n,\tilde{\theta})}(A) - P^{(n,\theta)}(A) \leq e^{-\beta n^{1/2+v}}(1+o(1))$$

as $n \to \infty$, for any event A, and any sieved parameter $\tilde{\theta}$.

Let $\tilde{\theta}$ be a sieved parameter, derived from θ. For any ϕ, with corresponding sieved parameter $\tilde{\phi}$ we can define the likelihood ratio $dP^{(n,\tilde{\phi})}/dP^{(n,\tilde{\theta})}$. Let $\mathbb{Q}_{\tilde{\theta},\delta} = \{\tilde{\phi} : \|\tilde{\phi} - \tilde{\theta}\|_1 > \delta\}$. Then since $\{\|\theta - \phi\|_1 > \delta\} \supset \{\|\tilde{\theta} - \tilde{\phi}\|_1 > \delta\}$, we have that $\tilde{\phi} \in \mathbb{Q}_{\tilde{\theta},\delta} \Rightarrow \phi \in \mathbb{Q}_{\theta,\delta}$. Therefore, for any $\tilde{\phi} \in \mathbb{Q}_{\tilde{\theta},\delta}$, the corresponding $\phi \in \mathbb{Q}_{\theta,\delta}$, and

$$P^{(n,\theta)}\left(A_n \cap \left\{\frac{dP^{(n,\tilde{\phi})}}{dP^{(n,\tilde{\theta})}} \geq 1\right\}\right) - Ce^{-\beta n^{1/2+v}} \leq P^{(n,\tilde{\theta})}\left(A_n \cap \left\{\frac{dP^{(n,\tilde{\phi})}}{dP^{(n,\tilde{\theta})}} \geq 1\right\}\right)$$

$$= \int_{A_n\cap\{\frac{dP^{(n,\tilde{\phi})}}{dP^{(n,\tilde{\theta})}}\geq 1\}} dP^{(n,\tilde{\theta})}$$

$$\leq \int_{A_n} \frac{dP^{(n,\tilde{\phi})}}{dP^{(n,\tilde{\theta})}} \, dP^{(n,\tilde{\theta})}$$

$$= P^{(n,\tilde{\phi})}(A_n)$$

$$= P^{(n,\phi)}(A_n) + Ce^{-\beta n^{1/2+\nu}}$$

$$\leq 2e^{-n(\varepsilon+\frac{1}{n})^2/2} + Ce^{-\beta n^{1/2+\nu}},$$

which implies that

$$P^{(n,\theta)}\left(\frac{dP^{(n,\tilde{\phi})}}{dP^{(n,\tilde{\theta})}} \geq 1\right) = P^{(n,\theta)}\left(A_n \cap \left\{\frac{dP^{(n,\tilde{\phi})}}{dP^{(n,\tilde{\theta})}} \geq 1\right\}\right) - P^{(n,\theta)}(A_n)$$

$$+ P^{(n,\theta)}\left(A_n \cup \left\{\frac{dP^{(n,\tilde{\phi})}}{dP^{(n,\tilde{\theta})}} \geq 1\right\}\right)$$

$$\leq 2e^{-n(\varepsilon+\frac{1}{n})^2/2} + 2Ce^{-\beta n^{1/2+\nu}} - 1 + 2e^{-n(\varepsilon-\frac{1}{n})^2/2} + 1$$

$$= 2e^{-n(\varepsilon+\frac{1}{n})^2/2} + 2e^{-n(\varepsilon-\frac{1}{n})^2/2} + 2Ce^{-\beta n^{1/2+\nu}}.$$

If $\hat{\theta}_{(s)}^{(n)}$ is the sieved PML estimator then

$$\frac{dP^{(n,\hat{\theta}_{(s)}^{(n)})}}{dP^{(n,\tilde{\theta})}} \geq 1.$$

For a given $n = n_1 + \cdots + n_k$ such that $n_1 \geq \cdots \geq n_k > 0$, (with k varying), there is a finite number $p(n)$ of possibilities for the value of $(n_1, \ldots, n_k)$, for which the asymptotic formula

$$p(n) = \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{\frac{2n}{3}}}(1 + o(1)),$$

as $n \to \infty$ (cf. [22]) holds. For each possibility of $(n_1, \ldots, n_k)$, there is a sieved PML estimator and we let $\mathcal{P}_n = \{\hat{\theta}_{(s)}^{(n),(1)}, \ldots, \hat{\theta}_{(s)}^{(n),(p(n))}\}$ be the set of all possible sieved PML estimators. Then

$$P^{(n,\theta)}(\|\hat{\theta}_{(s)}^{(n)} - \tilde{\theta}\|_1 > \delta) = \sum_{\tilde{\phi}\in\mathcal{P}_n\cap\mathbb{Q}_{\tilde{\theta},\delta}} P^{(n,\theta)}(\hat{\theta}_{(s)}^{(n)} = \tilde{\phi})$$

$$\leq \sum_{\tilde{\phi}\in\mathcal{P}_n\cap\mathbb{Q}_{\tilde{\theta},\delta}} P^{(n,\theta)}\left(\frac{dP^{(n,\tilde{\phi})}}{dP^{(n,\tilde{\theta})}} \geq 1\right)$$

$$\leq 2p(n)\left(e^{-\frac{n}{2}(\varepsilon-\frac{1}{n})^2} + e^{-\frac{n}{2}(\varepsilon+\frac{1}{n})^2} + Ce^{-\beta n^{1/2+\nu}}\right).$$

This completes the proof. □

The sieved PML estimator is strongly consistent.

COROLLARY 3. *Under the assumption of Theorem* 4, *the sequence of sieved maximum likelihood estimators* $\hat{\theta}_{(s)}^{(n)}$ *is strongly consistent in* L_1-*norm, that is,*

$$\|\hat{\theta}_{(s)}^{(n)} - \tilde{\theta}\|_1 \overset{a.s.}{\to} 0$$

as $n \to \infty$.

PROOF. The proof follows from Theorem 4, analogously to Corollary 3. □

Note that if $\theta \in \Theta_\kappa$, so that $\theta_x = l(x)x^{-\kappa}$ with $l(x)$ a function slowly varying at infinity and $\kappa > 1$, then the condition on the cut-off point is

$$Ce^{-\beta n^{1/2+\nu}} \sim \sum_{i=k(n)+1}^{n} \theta_i \sim \sum_{i=k(n)+1}^{n} i^{-\kappa} = k(n)^{-\kappa} \sum_{i=1}^{n-k(n)} i^{-\kappa}$$

$$\sim k(n)^{-\kappa}(n-k(n))^{-\kappa+1}$$

$$\geq k(n)^{-\kappa}n^{-\kappa+1},$$

where the last inequality follows since $\kappa > 1$ and $k(n) < n$. There is no way that we can have the condition of Theorem 4 satisfied if we only assume $\theta \in \Theta_\kappa$.

THEOREM 5. *Let* $\Theta_{\nu,\beta} = \{\theta : \theta_x = o(x^{\nu-1/2}e^{-\beta x^{\nu+1/2}}) \text{ as } x \to \infty\}$ *for* $\nu > 0, \beta > 0$ *fixed. Then, if* $\theta \in \Theta_{\nu,\beta}$,

$$n^\alpha \|\hat{\theta}_{(s)}^{(n)} - \tilde{\theta}\| \overset{a.s.}{\to} 0$$

as $n \to \infty$, *with* $\alpha < 1/4$.

PROOF. Assume that $\theta \in \Theta_{\nu,\beta}$. Then the condition on exponentially decreasing tails in Theorem 4 is satisfied. Furthermore, the condition $\forall \delta > 0 \exists r < \infty$ such that $\sum_{x=r}^{\infty} \theta_x < \delta/4$, translates to

$$\delta/4 \geq e^{-\beta r^{1/2+\nu}} \quad \Leftrightarrow \quad r \geq \left(\frac{-\log\delta/4}{\beta}\right)^{2/(1+2\nu)}.$$

The dominant part of the exponent in the right-hand side of Theorem 4 is then, replacing δ with δ/n^α for an α to be chosen and with $\varepsilon = \delta/8r$ and $r \sim (-\log\delta)^{2/(1+2\nu)}$,

$$n^{1/2} - n\varepsilon^2 - 2\varepsilon - 1/n \sim n^{1/2} - \frac{n^{1-2\alpha}\delta^2}{(-\log\delta)^{4/(1+2\nu)}} - \frac{n^{-\alpha}\delta}{(-\log\delta)^{2/(1+2\nu)}}$$

$$= n^{1/2} - n^{1-2\alpha}c_1(\delta) - n^{-\alpha}c_2(\delta),$$

which converges to $-\infty$ as $n \to \infty$ if $1 - 2\alpha > 1/2$ and $\alpha > 0$, that is, if $0 < \alpha < 1/4$. Thus, the rate is n^α for any $\alpha < 1/4$. □

4.3. *Comparison to the naive estimator.* An alternative to the nonparametric maximum likelihood estimators, studied in the previous two subsections, is the naive estimator, consisting of estimating first the order relation from the data, and then given that estimate the population frequency by the observed population frequencies.

We can obtain stronger results for the naive estimator than for the nonparametric maximum likelihood estimators. In fact, we can state almost sure supnorm convergence of the naive estimator with an almost parametric rate.

LEMMA 3. *Let $\hat{f}^{(n)} = T(f^{(n)})$ be the naive estimator. Then for any $\varepsilon > 0$,*

$$P^{(n,\theta)}(\|\hat{f}^{(n)} - \theta\|_\infty > \varepsilon) \leq 2e^{-n\varepsilon^2/2}.$$

PROOF. We argue similarly to the proof of Lemma 1: Combining the Dvoretzky–Kiefer–Wolfowitz inequality,

$$\mathbb{P}_\theta\Big(\sup_x |F^{(n)}(x) - F_\theta(x))| \geq \varepsilon\Big) \leq 2e^{-2n\varepsilon^2},$$

with $\{\sup_x |F^{(n)}(x) - F_\theta(x)| \geq \varepsilon\} \supset \{\sup_x |f_x^{(n)} - \theta_x| \geq 2\varepsilon\}$, we get

$$\mathbb{P}_\theta\Big(\sup_x |f_x^{(n)} - \theta_x| \geq \varepsilon\Big) = P^{(n,\theta)}\Big(\sup_x |f_x^{(n)} - \theta_x| \geq \varepsilon\Big)$$

$$\leq 2e^{-n\varepsilon^2/2}.$$

From the contraction property $\|T(f) - T(g)\|_\infty \leq \|f - g\|_\infty$ of the monotone rearrangement map T and since $T(\theta) = \theta$, with $\hat{f}^{(n)} = T(f^{(n)})$, this implies that $\{\|\hat{f}^{(n)} - \theta\|_\infty \geq \varepsilon\} \subset \{\|f^{(n)} - \theta\|_\infty \geq \varepsilon\}$ and

$$P^{(n,\theta)}\Big(\sup_x |\hat{f}_x^{(n)} - \theta_x| \geq \varepsilon\Big) \leq 2e^{-n\varepsilon^2/2}. \qquad \square$$

Lemma 3 implies consistency in probability, with rate $\alpha(n) = n^{1/2}(\log n)^{-1/2}$, since then $e^{-n\varepsilon^2/2\alpha(n)^2} = e^{-\varepsilon^2 \log n/2} = n^{-\varepsilon^2/2}$, which goes to zero, for every ε. Almost sure consistency with rate $\alpha(n) = n^{1/2+\delta}$ holds, since $e^{-n\varepsilon^2/2\alpha(n)^2} = e^{-n^\delta \varepsilon^2/2}$ which is summable (in n).

Thus, we have the almost sure convergence and convergence in probability:

$$n^{1/2-\delta}\|\hat{f}^{(n)} - \theta\|_\infty \xrightarrow{\text{a.s.}} 0,$$

$$\frac{n^{1/2}}{\log n^{1/2}}\|\hat{f}^{(n)} - \theta\|_\infty \xrightarrow{P} 0,$$

for any $\delta > 0$, as $n \to \infty$,

For the sieved model, recall the definition (20) of the estimator $\check{f}^{(n)}$. Then similar to the proof of Lemma 2, we obtain the following result.

LEMMA 4. *Let f_n be the empirical probability mass function based on a sample $x_1, \ldots, x_n$ from a fixed decreasing probability mass function θ, and let $\check{f}^{(n)}$ be as defined in* (20). *Then, for any $\varepsilon > 0$,*

$$P^{(n,\theta)}\big(\big\| \check{f}^{(n)} - \theta \big\|_\infty > \varepsilon \big) \leq 2e^{-n(\varepsilon - \frac{1}{n})^2/2}.$$

As a consequence, this again give above rates in the two convergence modes.

5. Discussion. We discuss a nonparametric maximum likelihood estimator (PML) for a probability mass function with unknown labels, an estimator first introduced in the computer science literature by Orlitsky et al. [19] under the name of high profile estimator. In Section 2, we also introduced a sieved estimator which has a truncation level on the size of the probability vector. The existence of the PML estimator is proven in [3], Supplement A.

The possibility of extending the model to include a continuous probability mass was already mentioned in [19]; however, it was not pursued further there. The introduction of a sieved estimator on the extended model is new and as we discuss below is important for many practical applications.

In Section 4, we proved strong consistency of "the" (actually any) PML (Theorem 2 and Corollary 1) and sieved PML (Theorem 4 and Corollary 3). The consistency of the PML was already claimed in [20] without complete proof. The key ingredients to prove Theorems 2 and 4 are Lemmas 1 and 2, respectively. Both lemmas use a novel strategy in proving consistency of the MPL by finding an observable event A, which has large probability under $P^{(n,\theta)}$, where θ is the true value of the parameter, but small probability under $P^{(n,\phi)}$, for all ϕ outside of a small ball around θ. Besides strong consistency, we also determined the rate of convergence of the regular and sieved PML in Theorems 3 and 5, respectively, which in both cases is almost of the order $n^{-1/4}$. We conclude Section 4 by giving an comparison to the naive estimator by proving a result analogous to Lemmas 1 and 2 for the latter.

REMARK 1. The obtained almost sure rate of convergence for the PML is (almost) $n^{-1/4}$. It is not clear what the optimal almost sure rate is: From the results of [13], the rate of convergence for the naive estimator is $n^{-1/2}$; however, this is the distributional rate of the L_p norms. The best possible almost sure rate for this problem could be $n^{-1/2}$, and it could be slower. From our own results in Section 4.3, we get almost sure rates $n^{-1/2+\delta}$ for any $\delta > 0$ for the naive estimator, which is faster than the rates for our estimator, it is however not clear if this is the optimal rate. Concerning our estimator, either the rate we obtain is the right rate for the PML which would mean that the PML is not optimal. Or else, the approach we use for deriving the rates is not the strongest possible, and in fact the rate for the PML is faster than $n^{-1/4}$ and (perhaps) equal to the optimal.

One should also note that the standard approach to deriving best rates for estimators is to use more sophisticated methods, for instance localization techniques. Our method consists of giving maximal inequalities for each PML and combining the derived bounds with a bound on the number of such PML's. This is a crude method and it is perhaps even surprising that we obtain consistency and rates at all.

Another major result is the introduction of an algorithm to numerically compute the sieved PML. This is presented in [3], Supplement B where the computation is based on the stochastic approximation of an expectation maximisation algorithm (SA-EM). In [18], a Monte Carlo Hastings expectation maximisation algorithm (MH-EM) of the standard PML was given. Our main advancement over this work is that we introduced the algorithm for the sieved estimator, and that we improved the statistical part of the EM algorithm by using the stochastic approximation.

Using the sieved estimator instead of the extended standard estimator can be an advantage when there are many unknown species with correspondingly small probabilities in the populations. Such a situation appears, for example, in forensic DNA analysis.

We illustrate this advantage on a small data example: Consider the partition $6 = 3 + 1 + 1 + 1$, that is, one species was observed three times and three species were observed once. The solution to the estimation problem is intuitive and can be proven analytically [17]: One species, say 1, has probability $1/2$ and there is a continuous probability mass with a total probability $1/2$, that is, based on the data, when sampling a new element, one expects to obtain 1 again in half of the cases or to observe a new species in the other half of the cases. To derive this estimator numerically, one would have to use the extended model and the here presented algorithm. Using the algorithm for the standard model and a number of species of order of the sample size, a uniform distribution over all species apart from species 1, would give a too big probability to each element. Similar situations occur in real data problems, that is, situations in which one would like to choose the species size of order of the sample size, but still account for a large number of rare species which have a very small probability which is comparable in size among the rare species.

REMARK 2. For the SA-EM algorithm, we note that, for a given finite value of K we know that for a given data set a maximum likelihood estimate of θ does exist. For each smaller value of K, there will typically correspond another, necessarily different, maximum likelihood estimate. All these estimates, one for each value of K up to some maximum, correspond to fixed points of the EM algorithm when run with a larger still value of K. The SAEM algorithm therefore has many possible limits, corresponding to all values of K not larger than the value corresponding to *the* maximum likelihood estimate of K for the given data-set and also not larger than the value of K chosen in the implementation of the algorithm. These limits

lie on the boundary of the parameter space. Once the procedure has got rather close to the boundary of the parameter-space, it is very difficult to move away again, since the size of potential steps is continuously being made smaller through the weights γ. Another troublesome part of the boundary of the parameter space corresponds to a sequence of probabilities p_a which are all equal to one another. For large problems, once a long stretch of equal probabilities has arisen, this long segment is very resilient to change. Only very slowly can it get longer or shorter (at either end).

Therefore, in some cases unwanted results (i.e., local maxima of the optimisation problem) can be obtained when moving close to the boundary of the parameter space, that is, when components of the probability vector become zero. In those cases, the numerical estimation can be improved by explicitly putting a lower bound on the allowed components of the probability vector. This means that in the M step of the EM algorithm one should change the isotonic regression to an isotonic regression of a probability mass function with a lower bound. It turns out that this problem has not been addressed in the literature; see, however, Balabdaoui et al. [4] for the related problem in isotonic regression of a regression function, and also van Eeden [24] and [23], Theorem 2.1. We have given a full solution to the lower bounded isotonic regression of a probability mass function in [3], Supplement C.

SUPPLEMENTARY MATERIAL

Supplement to "Estimating a probability mass function with unknown labels" (DOI: 10.1214/17-AOS1542SUPP; .pdf). Supplement consisted of Supplement A: Existence of the PML; Supplement B: Computation of the PML, and Supplement C: An algorithm for estimating a decreasing multinomial probability with lower bound.

REFERENCES

[1] ACHARYA, J., ORLITSKY, A. and PAN, S. (2009). The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns. In *IIEEE International Symposium on Information Theory* 1135–1139.

[2] ANEVSKI, D. and FOUGÈRES, A.-L. (2007). Limit properties of the monotone rearrangement for density and regression function estimation. Lund University. Available at arXiv:0710.4617v1.

[3] ANEVSKI, D., GILL, R. D. and ZOHREN, S. (2017). Supplement to "Estimating a probability mass function with unknown labels." DOI:10.1214/17-AOS1542SUPP.

[4] BALABDAOUI, F., RUFIBACH, K. and SANTAMBROGIO, F. (2010). Least-squares estimation of two-ordered monotone regression curves. *J. Nonparametr. Stat.* **22** 1019–1037. MR2738880

[5] DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.* **27** 642–669. MR0083864

[6] EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

[7] ESTY, W. W. (1982). Confidence intervals for the coverage of low coverage samples. *Ann. Statist.* **10** 190–196. MR0642730

[8] ESTY, W. W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *Ann. Statist.* **11** 905–912. MR0707940

[9] FISHER, R. A., CORBET, A. S. and WILLIAMS, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**.

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

[11] GOOD, I. J. and TOULMIN, G. H. (1956). The population frequencies of species and the estimation of population parameters. *Biometrika* **43** 45–63.

[12] HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1952). *Inequalities*, 2nd ed. Cambridge University Press, Cambridge.

[13] JANKOWSKI, H. K. and WELLNER, J. A. (2009). Estimation of a discrete monotone distribution. *Electron. J. Stat.* **3** 1567–1605. MR2578839

[14] LIEB, E. H. and LOSS, M. (1997). *Analysis. Graduate Studies in Mathematics* **14**. Amer. Math. Soc., Providence, RI. MR1415616

[15] MAO, C. X. and LINDSAY, B. G. (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89** 669–681. MR1929171

[16] MASSART, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283.

[17] ORLITSKY, A. and PAN, S. (2009). The maximum likelihood probability of skewed patterns. In *IEEE International Symposium on Information Theory*.

[18] ORLITSKY, A., SAJAMA, S., SANTHANAM, N. P., VISWANATHAN, K. and ZHANG, J. (2004). Algorithms for modeling distributions over large alphabets. In *Information Theory*, 2004. *ISIT* 2004. *Proceedings. International Symposium on Information Theory* 304.

[19] ORLITSKY, A., SAJAMA, S., SANTHANAM, N. P., VISWANATHAN, K. and ZHANG, J. (2004). On modeling profiles instead of values. In *Proceeding UAI'04 Proceedings of the* 20*th Conference on Uncertainty in Artificial Intelligence* 426–435.

[20] ORLITSKY, A., SAJAMA, S., SANTHANAM, N. P., VISWANATHAN, K. and ZHANG, J. (2005). Convergence of profile based estimators. In *Information Theory*, 2005. *ISIT* 2005. *Proceedings. International Symposium on Information Theory* 1843–1847.

[21] ORLITSKY, A., SANTHANAM, N. P., VISWANATHAN, K. and ZHANG, J. (2004). On modeling profiles instead of values. In *Proceedings of the Twentieth Conference Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*-04) 426–435. AUAI Press, Arlington, VA.

[22] RAMANUJAN, S. and HARDY, G. H. (1918). Asymptotic formulae in combinatorial analysis. *Proc. Lond. Math. Soc.* **17** 75–115.

[23] ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

[24] VAN EEDEN, C. (1957). Maximum likelihood estimation of partially or completely ordered parameters. ii. *Indag. Math.* **60** 201–211.

[25] VONTOBEL, P. O. (2012). The Bethe permanent of a non-negative matrix. *IEEE Trans. Inform. Theory* **59** 1866–1901.

[26] VONTOBEL, P. O. (2014). The Bethe and Sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the Valiant–Valiant estimate. In *Proceedings of Information Theory and Applications Workshop (ITA), 9–14 Feb*.

[27] ZHANG, C.-H. and ZHANG, Z. (2009). Asymptotic normality of a nonparametric estimator of sample coverage. *Ann. Statist.* **37** 2582–2595. MR2543704

D. ANEVSKI
CENTRE FOR MATHEMATICAL
  SCIENCES
LUND UNIVERSITY
BOX 118
221 00 LUND
SWEDEN
E-MAIL: dragi@maths.lth.se

R. D. GILL
MATHEMATICAL INSTITUTE
LEIDEN UNIVERSITY
NIELS BOHRWEG 1
2333 CA LEIDEN
THE NETHERLANDS
E-MAIL: gill@math.leidenuniv.nl

S. ZOHREN
DEPARTMENT OF MATERIALS
UNIVERSITY OF OXFORD
PARKS ROAD
OX1 3PH, OXFORD
UNITED KINGDOM
E-MAIL: stefan.zohren@materials.ox.ac.uk