THEORETICAL AND NUMERICAL ASPECTS OF THE PATTERN MAXIMUM LIKELIHOOD ESTIMATOR, WITH A VIEW TOWARDS SYMMETRIC FUNCTIONALS ESTIMATION

PAULINA BENTHEM CIANO

Bachelor's thesis
2022:K12

LUND UNIVERSITY

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Abstract

Suppose that an infinite population is partitioned into different species. Given a random sample from the population of species, we are interested in estimating the species richness, which is the number of different species inhabiting a given area. The species richness is a symmetric functional of the probability mass function. A suitable model for estimating the probability mass function is via the pattern maximum likelihood. When the pattern maximum likelihood cannot be found analytically, a sieved version of the pattern maximum likelihood can be used to find a numerical solution to the likelihood problem. The sieved pattern maximum likelihood estimator is consistent and can be calculated numerically using the SAEM algorithm.

Acknowledgements

I am grateful to my supervisor Dragi Anevski for supporting me not only during the process of this thesis but also during all the courses he taught me. It has been a great pleasure working with you and learning from you.

Contents

1 Introduction

How many words did Shakespeare know, cf. [7]? How many rare species of Malayan butterflies could Corbet find if he went back to British Malaya for two more years, cf. [8]? These are examples of the unseen species problem, which deals with estimating the number of species that have not been observed in samples. The problem was first studied in the early 1940s by Fisher in [8]. Since then, it has been studied extensively from different perspectives and using different methods, as reviewed in [4].

This thesis presents a detailed study of the so-called profile maximum likelihood estimator for estimating the unknown population frequency distribution for species in a natural habitat. The estimator was first introduced by Orlitsky et al. in [10], and a more detailed study was done by Anevski et al. in [3]. We follow the exposition of [3], and in particular, describe the Monte Carlo techniques used for the computation of the estimator. These techniques use an expectation-maximisation (EM) algorithm approach cf. [9], which we make a careful exposition of, as well as the stochastic approximation expectation-maximisation (SAEM) algorithm cf. [5], which we also discuss in detail. Finally, the research question of interest for us, and the question underlying the study of the species population distribution that we undertake, is the estimation of the species richness. The species richness is a function of the species population distribution, and it is a symmetric functional, a concept that we also discuss briefly.

Given a random sample from an infinite population of species in a specific area, it is interesting to determine the species richness, i.e. the number of species that have a positive probability of being observed. The species richness $S(p)$ is an example of a symmetric functional of the probability mass function p. A species richness estimator can be found using the plug-in approach. If $\hat{p}$ is an estimate of the probability mass function, then $S(\hat{p})$ is the estimate of the species richness.

Given a sample, the naive estimator of the relative frequencies of the species is given by the vector of relative frequencies of the observed species. For example, consider a sample of 7 observations. If two different species are observed three times and another species is observed once, the naive estimator of the relative frequencies is (3/7, 3/7, 1/7) for the observed species.

For the naive estimator the species-richness estimate equals the number of species observed. However, if the species have small probabilities of being observed, it is possible to see species that have not been observed before when taking a new sample. The evident problem with the naive estimator is that it assigns zero probability to unobserved species.

The thesis is organised as follows:

In section 2, we define and give examples of symmetric functionals of the probability mass function that are interesting for population estimation, ecology, biology, and information theory, among others. These functionals can be estimated by the profile maximum likelihood (PML) plug-in estimator.

In section 3, we introduce different models to estimate the probability mass function for problems where the observed data comes from a population with unknown relative frequencies. The observed data can be seen as an ordering of the true data. The true data can be seen as an observation coming from a multinomial distribution with unknown parameters of probabilities. The basic model uses a bijection between the observed and true data. However, the maximum likelihood for this model does not always exist. The extended model, or PML model, was first introduced in [10]. The extended model introduces a continuous part, consisting of uncountably many species, each with zero probability of being observed. Each species from the continuous part can be observed at most once in a sample. The pattern maximum likelihood estimator can be shown to exist. The PML estimator can be calculated analytically for small models. When it is not possible, one can use the sieved version of the PML (sPML) estimator. The sPML introduces a truncation level in the vector of unknown probabilities.

In section 4, we present estimation methods for the sieved version of the pattern maximum likelihood estimator. The EM algorithm is a method to maximise the likelihood of the observed data. However, this method requires knowing the density of the unobserved mappings ψ. The SAEM algorithm is a modification of the EM algorithm that can be applied to finding the sPML estimator.

Finally, in section 5, we present a conclusion and some open problems.

2 Symmetric functionals of the probability mass function

Given a random sample from an infinite population of species in a specific area, it is interesting to determine the species richness, that is, the number of species that have positive probabilities of being observed. The species richness can be seen as a symmetric functional of the probability mass function.

Let $\Delta = \{(p_1, p_2, ...);\ p_i \geq 0,\ \forall i \in \{1, 2, ...\}$ and $\sum_{i=1}^{\infty} p_i = 1\}$ be the set of probability mass functions. A functional $f : \Delta \to \mathbb{R}$ of the probability mass function is symmetric if it is invariant under label permutations, i.e.

$$f(p_1, p_2, ...) = f(p_{\phi(1)}, p_{\phi(2)}, ...)$$

for all bijections ϕ from $\{1, 2, ...\}$ to $\{1, 2, ...\}$.

Some examples of symmetric functionals of the probability mass function are

> **Support size**: $S(p) = \sum_{i=1}^{\infty} \mathbb{1}_{\{p_i > 0\}}$, the number of elements with positive probability.
>
> **Support coverage**: $S_m(p) = \sum_{i=1}^{\infty} (1 - (1 - p_i)^m)$, the expected number of elements observed in m samples.
>
> **Shannon entropy**: $H(p) = -\sum_{i=1}^{\infty} p_i \log p_i$, the measure of the degree of indeterminacy of a random variable.

Let p be a list of probabilities and let $f(p)$ be a symmetric functional. Given $\hat{p}$, any estimator of p, one can estimate $f(p)$ with the plug-in approach, so that $f(\hat{p})$ is the estimator of $f(p)$. An estimator of the probability that will be discussed in the sequel, the pattern maximum likelihood (PML) plug-in estimator, is shown to be competitive for estimating any symmetric functional in [1].

The species richness equals the support size of the vector of relative frequencies for the species. The vector of relative frequencies $(\theta_1, \theta_2, ...)$ is unknown. In the next section, we look at different models for the unknown vector. If the PML estimator $\hat{\theta}$ can be found analytically, the species richness can be estimated by

$$S(\hat{\theta}) = \sum_{\alpha=1}^{\infty} \mathbb{1}_{\{\hat{\theta}_\alpha > 0\}}.$$

An interesting question that was not investigated further is: if we know that an estimator of the probability mass function has some asymptotic properties, e.g. being consistent, whether or not the plug-in estimator also has such properties.

3 The problem

The naive estimator for the probability mass function assigns zero probability to unobserved species. This is problematic when the relative frequencies are small, since it is possible to observe species that were not observed in previous samples. To also be able to calculate the relative frequencies of unobserved species, the observed data can be seen as an ordering of the true data coming from a multinomial distribution. Three different models are presented in this section that make use of a mapping χ that relates the observed data with the true data.

3.1 Basic model

Let the random sample consist of T observed individuals from U different species. Denote by $N_1, ..., N_U$ the set of absolute frequencies in decreasing order so that N_i is the number of individuals observed from the ith most frequently observed species in the sample. Since T is the number of observed individuals, $N_1, ..., N_U$ sums up to T and the set $N_1, ..., N_U$ can be seen as a partition of the positive integer T. Consider as an example $T = 7 = 3+1+2+1$, so that one species is observed three times, another species is observed twice and two species are observed once. The partition of the integer 7 is $(3, 2, 1, 1)$. Let $N = (N_1, N_2, ...)$ be the list of absolute frequencies in decreasing order to which a list of zeros is appended representing unobserved species, so that $N_i > 0$ for $i = 1, ..., U$ and $N_i = 0$ for $i > U$.

Let $\mathcal{A}$ be the set of all possible species, and $\aleph$ be the number of species in the population that have positive probability. Denote by $\theta_1, ..., \theta_\aleph$ the set of unknown probabilities in decreasing order so that θ_α is the probability of finding the αth most common species. Let $\theta = (\theta_1, \theta_2, ...)$ be the list of unknown probabilities in decreasing order to which a list of zeros is appended representing non-existent species, so that $\theta_\alpha > 0$ for $\alpha = 1, ..., \aleph$ and $\theta_\alpha = 0$ for $\alpha > \aleph$. The infinite list of probabilities θ is required to sum up to 1 in the basic model.

The ith most observed species in the sample does not have to be the ith most common species in the population. Therefore, the data N can be seen as an ordering of the true data $(X_1, X_2, ...)$, where X_α is the number of times the αth most common species was observed. Let χ be the bijection from $\mathbb{N}$, the order of the observed species, to $\mathcal{A}$, the true order of the species in the population, so that $\chi(i) = \alpha$ if and only if the ith most frequent species in the sample is the αth most frequent species in the population. The bijection χ is not observed.

The true data $(X_1, X_2, ...)$ can be seen as an observation from a multinomial distribution $\text{Multi}(T, \theta_1, \theta_2, ...)$. Thus, the list of absolute frequencies $(N_1, N_2, ...) \in \text{Multi}(T, \theta_{\chi^{-1}(1)}, \theta_{\chi^{-1}(2)}, ...)$, and therefore $(N_1, N_2, ...)$ is a sufficient statistic for θ. The sample space is the set of all possible partitions N of the integer T, where T is the sample size. We can define a discrete probability measure on the sample space with underlying parameter θ as

$$P^{(T,\theta)}(A) = \sum_{(N_1,N_2,...)\in A} \binom{T}{N_1\ N_2\ \ldots} \sum_{\chi} \prod_{i} \theta_{\chi(i)}^{N_i},$$

and the likelihood for θ based on the sample N as

$$L(\theta) \propto \sum_{\chi} \prod_{i} \theta_{\chi(i)}^{N_i} = \sum_{\chi} \prod_{\alpha} \theta_{\alpha}^{N_{\chi^{-1}(\alpha)}}.$$

A maximum likelihood estimator (MLE) for θ can be introduced as

$$\hat{\theta} = \underset{\theta:\theta_1\geq\theta_2\geq\cdots,\sum_{i=1}^{\infty}\theta_\alpha=1}{\text{argmax}} L(\theta).$$

The maximum likelihood estimator for the basic model does not always exist. An example of this can be found in [3]. Therefore, the extended model MLE was studied, also called the pattern maximum likelihood (PML) estimator. The PML estimator exists, cf. [3].

3.2 Extended model

In the extended model, the set of unknown probabilities in decreasing order $\theta = (\theta_1, \theta_2, ...)$ is not required to sum up to one, allowing θ to be possibly defective. We introduce a blob species state 0, with population frequency θ_0, and require that $\theta_0 = 1 - \sum_{\alpha=1}^{\infty} \theta_\alpha$. The blob corresponds to the collection of species that each have zero probability of being observed, and an arbitrary species in the collection of the blob is observed with probability θ_0. Note, in particular, that each species in the collection of blob species can only be observed once in a sample. For that reason, the singletons in our sample could belong to either blob species or species with positive probability. Therefore, the mapping χ satisfies that if $\chi(i) = 0$, then $N_i = 0$ or 1. However, the mapping χ satisfies that for species with positive probability, i.e. $\alpha \geq 1$, there exists exactly one i so that $\chi(i) = \alpha$.

A possibly defective probability measure with underlying parameter θ may be defined as

$$P^{(T,\theta)}(A) = \sum_{(N_1,N_2,...)\in A} \sum_{\chi} \frac{T!}{N_0! \prod_{i\geq 1} N_i!} \theta_0^{N_0} \prod_{\alpha=1}^{\infty} \theta_\alpha^{N_{\chi^{-1}(\alpha)}},$$

where $N_0 = T - \sum_{\alpha=1}^{\infty} N_{\chi^{-1}(\alpha)}$. The measure is defective in the sense that $(\theta_1, \theta_2, ...)$ does not necessarily have total mass 1.

Then the extended model MLE or PML estimator is defined as

$$\hat{\theta} = \underset{\theta:\theta_1\geq\theta_2\geq\cdots,\sum_{i=1}^{\infty}\theta_\alpha\leq 1}{\operatorname{argmax}} \sum_{\chi} \frac{T!}{N_0! \prod_{i\geq 1} N_i!} \theta_0^{N_0} \prod_{\alpha=1}^{\infty} \theta_\alpha^{N_{\chi^{-1}(\alpha)}}.$$

The PML estimator exists and can be shown to be almost surely consistent in L^1-norm, cf. [3].

3.3 Sieved model

The PML estimator can be calculated analytically for small models, cf. [2]. For larger models, maximising the likelihood can be very computationally expensive since one needs to optimise over an infinite size vector. In order to simplify the computations, a modification of the PML was studied in [3]. The sieved PML (sPML) differs from the PML model in that the probability vector θ is truncated. Given a truncation level K, let θ be the finite vector of decreasing probabilities $\theta = (\theta_1, ..., \theta_K)$ and let the blob have population frequency $\theta_0 = 1 - \sum_{\alpha=1}^{K} \theta_\alpha$. Similarly as for the PML model, the mapping χ satisfies that if $\chi(i) = 0$, then $N_i = 0$ or 1, and that for $\alpha \in \{1, ..., K\}$ there exists exactly one i so that $\chi(i) = \alpha$.

When the sample size T is large and θ_0 is positive, the observed data tends to end in a long list of ones, where the singletons can either belong to species with positive probability or species in the blob. For this model, the underlying data $X = (X_0, X_1, ..., X_K)$ can be seen as an observation from a multinomial distribution $\mathrm{Multi}(T, \theta_0, \theta_1, ..., \theta_K)$. Denote the partition of the integer $X_+ = \sum_{\alpha=1}^{K} X_\alpha$ by $N_+ = (N_1, ..., N_J)$, where J is the number of species observed that have positive probability $X_\alpha > 0$, and note that J is unobserved. The X_0 animals correspond to animals observed from different blob species, and thus each species has zero probability of being observed and an animal from the set of blob species is observed at most once with probability θ_0. The observed data is the ordered partition $N = (N_1, ..., N_J, 1, ..., 1)$, i.e. the ordered partition N_+ to which we append a list of X_0 ones.

We can introduce a sieved likelihood as

$$L(\theta) = \sum_{\chi} \frac{T!}{N_0! \prod_{i \geq 1} N_i!} \theta_0^{N_0} \prod_{\alpha=1}^{K} \theta_\alpha^{N_{\chi^{-1}(\alpha)}},$$

where $N_0 = T - \sum_{\alpha=1}^{K} N_{\chi^{-1}(\alpha)}$.

Then the sPML estimator is defined as

$$\hat{\theta} = \operatorname*{argmax}_{\theta : \theta_1 \geq \cdots \geq \theta_K, \sum_{i=1}^{K} \theta_\alpha \leq 1} L(\theta).$$

The sPML estimator exists and can be shown to be almost surely consistent in L^1-norm, cf. [3].

4 Estimation of the sieved model

The sPML model depends on the unobserved mapping ψ, $\psi = \chi^{-1}$. In section 4.1 the EM algorithm is presented. This algorithm can be used to calculate an MLE of observed data where the model depends on unobserved, missing or latent variables. In section 4.1.1 a compact representation of the data is given. This representation is used to find an expression for the complete data likelihood. In section 4.1.2 the EM algorithm is formulated to obtain the sPML estimator. However, in this case the density of the latent variable Ψ is unknown and the expression for the E step in the EM algorithm therefore cannot be expressed in closed form.

To address this, in section 4.2 the SAEM algorithm is presented. The SAEM algorithm is a modification of the EM algorithm that replaces the E step with a simulation step and a stochastic approximation step. In the simulation step of the SAEM algorithm, realisations of the missing data are used. In section 3, the Metropolis-Hastings (MH) algorithm is presented. This algorithm is a Markov Chain Monte Carlo method to obtain realisations of densities when direct sampling is tedious. In section 4.2.2 the MH algorithm is applied to obtain a sample from Ψ. In section 4.2.3 the SAEM algorithm is formulated to obtain the sPML estimator.

4.1 The expectation-maximisation (EM) algorithm

The expectation-maximisation (EM) algorithm is an iterative method to find the maximum likelihood estimator of the observed data with respect to some unknown parameter $\theta \in \Theta$, where the model depends on missing or latent variables, cf. [9].

Consider the observed data y generated by the random variable Y, as well as the missing or unobserved data from the random variable Ψ. The complete data if observed, would be generated by the random variable $X = (Y, \Psi)$, where Ψ can be seen as being removed from X by the application of some mapping $Y = T(X)$. The random variable Y can be seen as a marginalisation of X. Let the density with respect to the measure μ of X be denoted by

$f_{X|\theta}(x)$. The density of Y is then given by

$$f_{Y|\theta}(y) = \int_{\{x:T(x)=y\}} f_{X|\theta}(x) d\mu(x).$$

Maximising the likelihood of the observed data $L(\theta|y) = f_{Y|\theta}(y)$ over $\theta \in \Theta$ is often more involved than maximising the likelihood of the complete data $L(\theta|x) = f_{X|\theta}(x)$ over $\theta \in \Theta$. Since the complete data is not observed, maximising the likelihood of the complete data $L(\theta|x)$ is not possible. The EM algorithm uses an approach where one iteratively maximises the expectation of the likelihood of the complete data $L(\theta|x) = f_{X|\theta}(x)$ given the observed data y and the current maximiser $\theta^{(t-1)}$.

The EM algorithm maximises the likelihood of the observed data $L(\theta|y)$ with respect to the unknown parameter θ. The algorithm is initialised at some parameter value $\theta^{(0)}$ and for iterations $t = 1, 2, ...$ alternates between an expectation (E) step and a maximisation (M) step until convergence.

The E and M steps are characterised by:

- E step: Compute $\mathcal{Q}_t(\theta)$, the expectation of the log-likelihood of the complete data X, given the observed data y and the estimated maximiser $\theta^{(t-1)}$ at iteration $t-1$

$$\begin{aligned}\mathcal{Q}_t(\theta) &= \mathrm{E}\big(\log L(\theta|X)\big|y, \theta^{(t-1)}\big) \\ &= \mathrm{E}\big(\log f_{X|\theta}(x)\big|y, \theta^{(t-1)}\big) \\ &= \int \log(f_{X|\theta}(x)) f_{\Psi|\theta^{(t-1)}}(\psi) d\mu(\psi),\end{aligned}$$

  where $f_{\Psi|\theta^{(t-1)}}(\psi)$ is the density of Ψ with respect to the measure μ. Given y and $\theta^{(t-1)}$ the only random part of X is Ψ.

- M step: Update $\theta^{(t-1)}$ by

$$\theta^{(t)} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{Q}_t(\theta).$$

Increasing $\mathcal{Q}_t(\theta)$ forces an increase in the observed data likelihood, cf. [6]. Convergence, and other properties of the EM algorithm, can be found in [11].

4.1.1 Compact representation of the data

As seen in section 3.1, the observed sample can be reduced by sufficiency to the partition N of the integer T, where T is the sample size and N is a decreasing sequence of positive integers corresponding to the absolute frequencies for each of the observed species.

The partition N can be represented more compactly by two sequences n and r of equal length J. These are $n = (n_1, ..., n_J)$, $n_1 < \cdots < n_J$ and $r = (r_1, ..., r_J)$, where n_j are the distinct absolute frequencies for different species, i.e. the distinct numbers appearing in the partition N, ordered strictly increasing, and r_j are the number of times n_j appears in the partition N. Consider again the example $T = 7 = 3 + 1 + 2 + 1$. The partition of the integer 7 is $N = (3, 2, 1, 1)$, and the sequences are $n = (1, 2, 3)$ and $r = (2, 1, 1)$ since the distinct numbers appearing in N ordered increasingly are $(1, 2, 3)$, and number 1 appears twice in N and numbers 2 and 3 appear once in N. The sequences n and r are a sufficient representation of the observed data, since knowing n and r is equivalent to knowing N.

Assume that the sample contains singletons and non-singletons, i.e. $n_1 = 1$ and $J \geq 2$. Also, assume, as in the sieved model, that the number K of species with positive probability is finite, so that the unknown vector of probabilities is $\theta = (\theta_1, ..., \theta_K)$, $\theta_1 \geq ... \geq \theta_K > 0$, and the blob consists of uncountably many species that each have zero probability of being observed, but a species of the blob can be observed at most once with probability $\theta_0 = 1 - \sum_{\alpha=1}^{K} \theta_\alpha$.

The missing data, i.e. how the observed data relates to the order in nature, can be represented by a function $\psi : \{1, ..., K\} \longrightarrow \{0, 1, ..., J\}$ with the constraints

C1: $\sum_{\alpha=1}^{\aleph} \mathbb{1}\{\psi(\alpha) = j\} = r_j$ for each $j > 1$,

C2: $\sum_{\alpha=1}^{\aleph} \mathbb{1}\{\psi(\alpha) = 1\} \leq r_1$.

The species observed more than once are species with positive probability. Then, for $j > 1$, C1 follows because r_j, i.e. the number of species observed j times, equals the number of species with positive probability observed j times. However, for $j = 1$, C2 follows since r_1, i.e. the number of species observed only once, could be from species with positive probability, as well as from species in the blob. The sequences n and r together with the function ψ constitute a sufficient statistic for θ based on the complete data.

To find a simple representation of the complete data, let f be the vector of distinct relative frequencies for the observed species, so that $f_j = n_j/T$ for $j \in \{1, ..., J\}$, where $T = \sum_{j=1}^{J} r_j n_j$ is the sample size. The vector f is also a sufficient statistic for the observed data since f is equivalent to N. Let $g = (g_0, g_1, ..., g_K)$ be the vector of relative frequencies for the underlying population species in our sample, and note that this is unobserved. The vector g is a sufficient statistic of the complete data, and it is uniquely determined by the vector f and the missing map ψ. Given f and ψ, g is given by

$$\begin{cases} g_\alpha = f_j & \text{if } \psi(\alpha) = j \geq 1, \\ g_\alpha = 0 & \text{if } \psi(\alpha) = 0, \\ g_0 = n_0/T, \end{cases}$$

for $\alpha \in \{0, 1, ..., K\}$, where $n_0 = r_1 - \sum_{\alpha=1}^{K} \mathbb{1}\{\psi(\alpha) = 1\}$ is the number of blob species observed.

The likelihood of the complete data is

$$\begin{aligned} L(\theta|g) &= \frac{T!}{n_0! \prod_{1 \leq \alpha \leq K : \psi(\alpha) \geq 1} n_{\psi(\alpha)}!} \theta_0^{n_0} \prod_{1 \leq \alpha \leq K : \psi(\alpha) \geq 1} \theta_\alpha^{n_{\psi(\alpha)}} \\ &\propto \frac{1}{n_0!} \theta_0^{n_0} \prod_{1 \leq \alpha \leq K : \psi(\alpha) \geq 1} \theta_\alpha^{n_{\psi(\alpha)}}, \qquad (1) \end{aligned}$$

where (1) holds since n_1 is equal to one and $\prod_{1 \leq \alpha \leq K : \psi(\alpha) \geq 2} n_{\psi(\alpha)}! = \prod_{2 \leq j \leq J} (n_j!)^{r_j}$, so the product in the likelihood is constant, i.e. does not depend on θ.

The log-likelihood of the complete data is

$$\log L(\theta|g) = -\sum_{i=1}^{n_0} \log i + n_0 \log \theta_0 + \sum_{1 \leq \alpha \leq K : \psi(\alpha) \geq 1} n_{\psi(\alpha)} \log \theta_\alpha + C, \qquad (2)$$

where C is a constant, i.e. it does not depend on θ.

The likelihood of the observed data can be calculated by the sum of the likelihoods of the complete data, given in (1), over all mappings ψ allowed by the constraints C1 and C2. Therefore, calculating the sieved pattern maximum likelihood estimator of the observed data in closed form seems a formidable problem.

4.1.2 Estimation of the sPML estimator via the EM algorithm

We next apply the EM algorithm to obtain the maximum likelihood of the observed data with respect to $\theta \in \Theta$ for the sPML model. The EM algorithm iteratively maximises the expectation of the likelihood of the complete data $L(\theta|g)$ given the observed data f and the current maximiser $\theta^{(t-1)}$.

Given the observed data f, the sieved pattern maximum likelihood can be approximated using the EM algorithm. The EM algorithm maximises the likelihood of the observed data $L(\theta|f)$ with respect to the unknown probabilities $\theta = (\theta_0, \theta_1, ..., \theta_K)$ in Θ, where

$$\Theta = \left\{ \theta;\ \theta_1 \geq \cdots \geq \theta_K \geq 0,\ \theta_0 \geq 0,\ \sum_{i=0}^{K} \theta_\alpha = 1 \right\}.$$

The algorithm is initialised at some $\theta^{(0)}$, and for iterations $t = 1, 2, ...$ alternates between an expectation (E) step and a maximisation (M) step until convergence.

The E and M steps are given by:

- E step: Compute $\mathcal{Q}_t(\theta)$, the expectation of the log-likelihood of the complete data g, given the observed data f

$$\begin{aligned} \mathcal{Q}_t(\theta) &= \mathrm{E}\big(\log L(\theta|g) \big| f, \theta^{(t-1)} \big) \\ &= \sum_{\psi:\text{C1,C2 hold}} \log L(\theta|g) \cdot f_{\Psi|\theta^{(t-1)}}(\psi), \end{aligned}$$

  where $\theta^{(t-1)}$ denotes the estimated maximiser at iteration $t-1$ and $f_{\Psi|\theta^{(t-1)}}(\psi)$ is the probability mass function of Ψ. Given f and $\theta^{(t-1)}$ the only random part of the complete data g is Ψ. The sum is over all mappings ψ allowed by the constraints C1 and C2.

- M step: Update $\theta^{(t-1)}$ by

$$\theta^{(t)} = \underset{\theta:\theta_1 \geq \cdots \geq \theta_K \geq 0, \theta_0 \geq 0, \sum_{i=0}^{K} \theta_\alpha = 1}{\operatorname{argmax}} \mathcal{Q}_t(\theta).$$

The expectation of the complete data $\mathcal{Q}_t(\theta)$ cannot be expressed in closed form since the density of Ψ is unknown. However, it can be approximated via the stochastic approximation of EM algorithm.

4.2 The stochastic approximation expectation-maximisation (SAEM) algorithm

The E step in the EM algorithm consists of finding $\mathcal{Q}_t(\theta)$, the expected log-likelihood of the complete data given the observed data. In some cases, this expectation cannot be computed analytically. Instead, one can replace the E step with a simulation (S) step and a stochastic approximation (AE) step, yielding the stochastic approximation of EM (SAEM) algorithm, cf. [5].

The SAEM algorithm calculates the maximum likelihood estimator of the observed data y with respect to the unknown parameter $\theta \in \Theta$. The algorithm is initialised at some $\theta^{(0)}$ and $\mathcal{Q}_0(\theta)$ and for iterations $t = 1, 2, ...$ alternates between a simulation (S) step, a stochastic approximation (AE) and a maximisation (M) step until convergence.

- S step: Generate $m(t)$ realisations of the missing data $\psi_t(i)$, $i = 1, ..., m(t)$ under the density $f_{\Psi|\theta^{(t-1)}}(\psi)$, where $\theta^{(t-1)}$ denotes the estimated maximiser at iteration $t-1$.
- AE step: Set

$$\mathcal{Q}_t(\theta) = \mathcal{Q}_{t-1}(\theta) + \delta_t \left(\frac{1}{m(t)} \sum_{i=1}^{m(t)} \log L(\theta | y, \psi_t(i)) - \mathcal{Q}_{t-1}(\theta) \right),$$

  where $\{\delta_t\}_{t \geq 1}$ is a positive sequence of step size and y is the observed data.

- M step: Update $\theta^{(t-1)}$ by

$$\theta^{(t)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathcal{Q}_t(\theta).$$

The convergence of the SAEM algorithm depends on the step size δ_t and the number of iterations $m(t)$ used in the stochastic approximation. It is recommended to decrease δ_t or increase $m(t)$ as the parameter approximation approaches a stationary point. If the maximisation step is computationally faster than the simulation step, the number of simulations $m(t)$ of the missing data can be set to 1 for all iterations. In the SAEM algorithm all simulated values of the missing data contribute to the evaluation of $\mathcal{Q}_t(\theta)$, with a forgetting factor δ_t.

More details on the SAEM algorithm in our setting is given in section 4.2.3.

4.2.1 The Metropolis-Hastings (MH) algorithm

The S step of the SAEM algorithm consists of generating realisations of the missing data under the density $f_{\Psi|\theta^{(t-1)}}(\psi)$ of Ψ. However, this is not always possible. The Metropolis-Hastings algorithm can be used to obtain a sample from the density of Ψ.

When it is hard to generate realisations of a density f or f is high dimensional, such as in our case where f is the density for ψ, one can sample from the density f by constructing a Markov chain having f as the stationary distribution. Note that the constructed samples will be dependent.

Markov Chain Monte Carlo (MCMC) methods can be used to sample from the distribution f by constructing a Markov chain $\{X_t\}_{t\geq 0}$ with a unique stationary distribution that equals the target distribution f. A realisation of the chain X_t for sufficiently large t will have approximately the distribution f.

The Metropolis-Hastings (MH) algorithm is an MCMC method to construct Markov chains. Assume that it is possible to simulate from a transition kernel $r(z|x)$ referred to as the proposal kernel. The MH algorithm is initialised at some X_0 and for iterations $t = 1, 2, ...$ alternates between the following steps.

- Generate the candidate X^* from the proposal kernel $r(X^*|X_{t-1})$.
- Set the acceptance probability

$$A(X_{t-1}, X^*) = \min\left(1, \frac{f(X^*)r(X_{t-1}|X^*)}{f(X_{t-1})r(X^*|X_{t-1})}\right). \tag{3}$$

- Draw $U \in \mathrm{Un}(0,1)$.
- Set

$$X_t = \begin{cases} X^* & \text{if } U \leq A(X_{t-1}, X^*), \\ X_{t-1} & \text{otherwise,} \end{cases}$$

  so that the candidate X^* is accepted with probability $A(X_{t-1}, X^*)$.

The chain X_t constructed via the MH algorithm is Markov since X_t only depends on the previous iteration X_{t-1}. One should check that the chain X_t generated has a unique stationary distribution.

For symmetric kernel proposals, it holds that $r(z|x) = r(x|z)$, for all x and z. In this case, the acceptance probability in (3) reduces to

$$A(x, z) = \min\left(1, \frac{f(z)}{f(x)}\right). \tag{4}$$

4.2.2 Draw samples from Ψ via the MH algorithm

The Metropolis-Hastings (MH) algorithm can be used to draw new samples of the missing data Ψ under the probability mass function $f_{\Psi|\theta^{(t-1)}}(\psi)$ of Ψ . To generate new candidates ψ^* of Ψ, we define a random walk proposal on the set of all mappings ψ allowed by the constraints C1 and C2. The two possible kinds of moves in the random walk are exchange moves or blob moves.

> **Exchange moves**: Exchange the values $\psi(\alpha_i)$ and $\psi(\alpha_j)$ of two different non-blob species α_i and α_j, so that $\psi(\alpha_i) \neq 0$, $\psi(\alpha_j) \neq 0$ and $\psi(\alpha_i) \neq \psi(\alpha_j)$.

> **Blob moves**: Increase or decrease the number of blob species n_0 by one by choosing a species from the set of blob species and singletons $\{\alpha;\ \psi(\alpha) = 0 \text{ or } \psi(\alpha) = 1\}$, and exchange the value of $\psi(\alpha) = 0$ to $\psi(\alpha) = 1$ or vice-versa.

These moves are not always possible.

Lemma If $J \geq 3$, then an exchange move is always possible. If the number of singletons r_1 is greater than 0 and the number of species with positive probability K is greater than the number of non-singletons $\sum_{j=2}^{J} r_j$, then a blob move is always possible.

This lemma is proved in [3].

Note that we use the likelihood of the complete data instead of the density of Ψ in the acceptance probability. The density of Ψ is unknown. Given the observed data f and the candidate ψ^* of Ψ, we can look at the ratio of the complete data likelihoods. The log-likelihood of the complete data is given in equation (2).

Exchange moves

To perform an exchange move, pick uniformly at random a pair of distinct non-blob species (α_i, α_j), so that $\psi(\alpha_i) \neq 0$, $\psi(\alpha_j) \neq 0$ and $\psi(\alpha_i) \neq \psi(\alpha_j)$. The proposal is symmetric since the pair (α_i, α_j) is chosen uniformly at random and the number of candidate pairs before and after the move remains equal.

The logarithm of the complete-data likelihood after the move to before the move equals

$$\begin{aligned}\log\left(\frac{L(\theta|f,\psi^*)}{L(\theta|f,\psi)}\right) &= n_{\psi(\alpha_j)}\log\theta_{\alpha_i} + n_{\psi(\alpha_i)}\log\theta_{\alpha_j} - (n_{\psi(\alpha_i)}\log\theta_{\alpha_i} + n_{\psi(\alpha_j)}\log\theta_{\alpha_j}) \\ &= (n_{\psi(\alpha_i)} - n_{\psi(\alpha_j)})(\log\theta_{\alpha_j} - \log\theta_{\alpha_i}).\end{aligned}$$

Since the proposal is symmetric, the acceptance probability is given by

$$A(\psi,\psi^*) = \min\left(1, \frac{L(\theta|f,\psi^*)}{L(\theta|f,\psi)}\right),$$

cf. equation (4).

The MH algorithm is initialised at ψ_0 and at each iteration t performs the steps:

- Given ψ_{t-1}, perform an exchange move to obtain the candidate move ψ^*.
- Calculate $A(\psi_{t-1}, \psi^*)$.
- Draw $U \in \mathrm{Un}(0,1)$.
- If $U \leq A(\psi_{t-1}, \psi^*)$, then the move is accepted and $\psi_t = \psi^*$. Otherwise, the move is rejected and $\psi_t = \psi_{t-1}$.

Blob moves

The number of species in $\{\alpha;\ \psi(\alpha) = 0 \text{ or } \psi(\alpha) = 1\}$ is the number of species with positive probability K minus the number of non-singleton species M. The exception is when the number of observed species L is less than the number of species with positive probability K, and n_0 the number of species in the blob is 0, then the number of species in $\{\alpha;\ \psi(\alpha) = 0 \text{ or } \psi(\alpha) = 1\}$ is the number of singletons S which is smaller than $K - M$.

Choose a species from $\{\alpha;\ \psi(\alpha) = 0 \text{ or } \psi(\alpha) = 1\}$. The proposed move is to change $\psi(\alpha) = 0$ to $\psi(\alpha) = 1$ or vice-versa.

If a species is chosen so that $\psi(\alpha) = 1$, then n_0 is increased by 1 and a non-blob species is removed. The logarithm of the ratio before and after the move is

$$\log\left(\frac{L(\theta|f,\psi^*)}{L(\theta|f,\psi)}\right) = -\log(n_0 + 1) + \log\theta_0 - \log\theta_\alpha.$$

This proposal is symmetric and the acceptance probability is

$$A(\psi,\psi^*) = \min\left(1, \frac{L(\theta|f,\psi^*)}{L(\theta|f,\psi)}\right),$$

cf. equation (4), except when $L < K$ and n_0 is equal to 0, then the proposal is not symmetrical and the acceptance probability is

$$A(\psi,\psi^*) = \min\left(1, \frac{L(\theta|f,\psi^*)\frac{1}{K-M}}{L(\theta|f,\psi)\frac{1}{S}}\right).$$

If a species is chosen so that $\psi(\alpha) = 0$, then n_0 is reduced by 1 and a non-blob species is added. The logarithm of the ratio before and after the move is

$$\log\left(\frac{L(\theta|f,\psi^*)}{L(\theta|f,\psi)}\right) = \log\theta_\alpha + \log n_0 - \log\theta_0.$$

This proposal is symmetric and the acceptance probability is

$$A(\psi,\psi^*) = \min\left(1, \frac{L(\theta|f,\psi^*)}{L(\theta|f,\psi)}\right),$$

cf. equation (4), except when $L < K$ and n_0 is equal to 1, then the proposal is not symmetrical and the acceptance probability is

$$A(\psi,\psi^*) = \min\left(1, \frac{L(\theta|f,\psi^*)\frac{1}{S}}{L(\theta|f,\psi)\frac{1}{K-M}}\right).$$

For any of these cases with corresponding acceptance probability $A(\psi,\psi^*)$, the MH algorithm is initialised at ψ_0 and at each iteration t performs the steps:

- Given ψ_{t-1}, perform a blob move to obtain the candidate move ψ^*.
- Calculate $A(\psi_{t-1},\psi^*)$.
- Draw $U \in \text{Un}(0,1)$.
- If $U \leq A(\psi_{t-1},\psi^*)$, then the move is accepted and $\psi_t = \psi^*$. Otherwise, the move is rejected and $\psi_t = \psi_{t-1}$.

Given the current state ψ, all other possible states can be accessed by performing several exchange or blob moves. Since all states communicate, the Markov chain $\{\psi_t\}_{t\geq 0}$ is irreducible. The state space of Ψ is finite for the sieved model. All states of an irreducible Markov chain with finite state space are positive recurrent. Consequently, the Markov chain obtained by the MH algorithm has a unique stationary distribution.

4.2.3 Estimation of the sPML estimator via SAEM algorithm

We now give a slightly more detailed description of the use of the SAEM algorithm for estimating the sPML estimator.

The SAEM algorithm calculates the maximum likelihood estimator of the observed data f with respect to the unknown parameter $\theta \in \Theta$. The algorithm is initialised at some $\theta^{(0)}$ and ψ_0, and for iterations $t = 1, 2, ...$ alternates between a simulation (S) step, a stochastic approximation (AE) and a maximisation (M) step until convergence.

The SAEM algorithm consists of the steps:

- S step: Draw ψ_t by doing one iteration of the MH algorithm. Given ψ_t and the observed data f, determine g_t, cf. section 4.1.1.
- AE step: Set
$$\mathcal{Q}_t(\theta) = \mathcal{Q}_{t-1}(\theta) + \delta_t\big(\log L(\theta|g_t) - \mathcal{Q}_{t-1}(\theta)\big),$$
where $\{\delta_t\}_{t\geq 1}$ is a decreasing positive sequence so that $\delta_1 = 1$, $\sum_{t=1}^{\infty} \delta_t = \infty$ and $\sum_{t=1}^{\infty} \delta_t^2 < \infty$. The suggested step size is $\delta_t = 1/t^{2/3}$ in [3].
- M step: Update $\theta^{(t-1)}$ by
$$\theta^{(t)} = \underset{\theta:\theta_1\geq\cdots\geq\theta_K\geq 0,\theta_0\geq 0,\sum_{i=0}^{K}\theta_\alpha=1}{\operatorname{argmax}} \mathcal{Q}_t(\theta).$$

An implementation of the SAEM algorithm for estimating the sPML estimator can be found in [3].

The maximiser $\theta^{(t)}$ in the M step of the SAEM algorithm can be found using a modification of the pool adjacent violators algorithm (PAVA) introduced in [3]. This modification consists of a bounded isotonic regression method in which the last positive element of the estimator $\theta^{(t)} = (\theta_0^{(t)}, \theta_1^{(t)}, \ldots, \theta_{\tilde{k}}^{(t)}, 0, 0, ...)$, namely, $\theta_{\tilde{k}}^{(t)}$ is assumed to be greater than a constant c, where $\tilde{k} \in (0, 1/c)$ to ensure that the vector $\theta^{(t)}$ is a list of probabilities. This modification was introduced for numerical reasons. When running the isotonic regression algorithm, the elements in the tail of the estimator $\theta^{(t)}$ get smaller in each iteration but not exactly equal to zero, so it does not converge. To ensure that the algorithm converges to the maximum likelihood estimator, one can modify the isotonic regression algorithm to a lower bounded isotonic regression algorithm. The bounded isotonic regression algorithm is proven to converge to the global maximum in [3].

5 Conclusions, discussion and open problems

The main goal of this thesis was to study symmetric functionals, in particular, the study of the species richness, which is an example of a symmetric functional of the probability mass function. To examine the performance of the sieved pattern maximum likelihood (sPML) plug-in estimator, one needs to understand the algorithm for obtaining the sPML estimator. The initial goal was to implement the sPML algorithm and make simulation studies. However, since the estimation problem and the algorithm are highly involved, in the sense that it consists of many components, most of the study has been dedicated to describing the algorithm in a detailed manner, so that it is accessible to the reader.

This thesis has studied two different methods for estimating the probability mass function. The first method is the pattern maximum likelihood (PML) estimator, which exists but sometimes cannot be computed analytically. When the PML estimator can be found, the species richness is estimated by the plug-in approach. The second method is the sieved version of the PML estimator. This estimator exists and can be estimated by the stochastic approximation expectation-maximisation (SAEM) algorithm. The properties of symmetric functional estimators for this method have not been studied in-depth in this thesis.

Further studies with an empirical approach could use the implementation given in [3] to study the performance of the sPML estimator or implement the algorithm to obtain the sPML estimator and then examine its performance. One could apply the algorithm on real data or simulated data, where the species richness and species population frequency are known, and use the plug-in sPML approach to obtain species richness estimates, consistency results and confidence intervals.

A more theoretical approach for continuing this project would be studying the properties of the symmetric functionals of sPML, in particular, how the properties of the sPML estimator translate to the properties of the plug-in estimator. For example, it would be of interest to study whether the consistency of the sPML implies consistency of the plug-in estimator.

References

[1] J. Acharya, H. Das, A. Orlitsky, and A. T. Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. *Proceedings of the 34th International Conference on Machine Learning*, 70:11 – 21, 2017.

[2] J. Acharya, A. Orlitsky, and S. Pan. The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns. *2009 IEEE International Symposium on Information Theory*, pages 1135–1139, 2009.

[3] D. Anevski, R. D. Gill, and S. Zohren. Estimating a probability mass function with unknown labels. *The Annals of Statistics*, 45(6):2708 – 2735, 2017.

[4] J. Bunge and M. Fitzpatrick. Estimating the number of species: A review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[5] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94 – 128, 1999.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[7] B. Efron and R. Thisted. Estimating the number of unsen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[8] R. A. Fisher, A. S. Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12(1):42–58, 1943.

[9] G. H. Givens and J. A. Hoeting. *Computational statistics*. John Wiley & Sons, 2 edition, 2012.

[10] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang. On modeling profiles instead of values. In *UAI*, 2004.

[11] C. F. J. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95 – 103, 1983.

Bachelor’s Theses in Mathematical Sciences 2022:K12
ISSN 1654-6229

LUNFMS-4065-2022

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/