A Model for Species Accumulation based on the Multinomial Distribution

Miguel Nakamura
Centro de Investigación en Matemáticas
Apartado Postal 402
Guanajuato, Gto. 36000, Mexico

Felipe Peraza
Escuela de Ciencias Físico-Matemáticas
Universidad Autónoma de Sinaloa
Apartado Postal 1872
Ciudad Universitaria, Culiacán, Sin. 80030, Mexico

Abstract

A *species accumulation curve* is a plot of the number of species found in a given region as a function of some measure of effort used to find them. In ecological studies, there are important goals achieved when the relationship between these two variables is understood. In this paper, we construct a model—which turns out to involve a multinomial distribution—based on the premise that the form of the accumulation curve depends on the joint behavior of species' *capture probabilities.* Maximum likelihood and profile likelihood methods are explored for the estimation of parameters. The expectation-maximization algorithm is described for dealing with incomplete collecting lists. Two examples which illustrate the techniques developed are considered, and comparisons are made with estimates based on other approaches.

Key words: Number of species; EM algorithm; profile likelihood.

Running head: Species accumulation curves.

1 Introduction

Although the problem addressed in this paper finds its motivation in biodiversity studies, it is possible that the setting may be applicable in other sciences as well. In a general context, suppose a population of objects is classified into an unknown number of classes and that one can observe and classify a sequence of individual objects drawn at random. Interest lies in modelling the relationship between the total number of classes identified and the effort applied. In ecological studies, there are important goals achieved when a theoretical relationship is found between these two variables. Examples are found in the planning of future expeditions, quantitative comparisons between two or more collection efforts, and in biodiversity and conservation studies when interest lies in total number of species present.

In this ecological setting, a *species accumulation curve* or "collector's curve" is a plot of the number of species found in a given region, $\mathrm{S}(n)$, as a function of some measure n of the effort used to find them. The simplest measure of effort is the total number of individuals examined, but in ecological studies typical units of measure are quadrant areas, sampled biomass, hours spent on observation, number of traps per day, and others (see Coldwell & Coddington, 1994).

Numerous references exist in the literature regarding estimation of the number of different types of objects in a population (see Bunge & Fitzpatrick, 1993). For example, biologists and ecologists are interested in estimating the total number of species in a population of plants or animals (Engen, 1976; Soberón & Llorente, 1993; Coldwell & Coddington, 1994). A linguist may be interested in estimating the size of the vocabulary of an author (Efron & Thisted, 1976), or a programmer in the number of bugs in new software, among other examples. But most of these works are either based on a capture-recapture sampling scheme or on n_x, the number of objects which appear exactly x times, and x_s, the number of individuals which belong to class s. Computing these stated frequencies requires the record of the whole collection process, which is not necessarily available in ecological studies. Furthermore, the techniques do not always take into account the time at which each new identified class appears or is discovered, so an explicit relationship between effort and capture is not addressed.

Soberón & Llorente (1993) (henceforth denoted SL) relate the concepts of effort and number of accumulated species by analytical methods. Their model is based on pure birth processes, where parametric transition probabilities (called *collecting functions*) are postulated on the basis of biological interpretation. The expected value of the ensuing stochastic process, $\mathrm{A}(n)$, is fit to the observed accumulation curve, using regression techniques to estimate the parameters of the process. Because comparisons will be made in a later section, we here state the three mean functions which SL consider. Their *exponential model* is $\mathrm{A}(n) = (a/b)\,[1 - \exp(-bn)]$. The *Clench* model (Clench, 1979), which is based on the Michelis-Menten equation, $\mathrm{A}\,(n) = an/\,(1 + bn)$ is also a special case of the SL approach. Here $\mathrm{A}(n) \to a/b$ as $n \to \infty$, so a/b is interpreted as the total number of species present. The *logarithmic* model is $\mathrm{A}\,(n) = \ln\,(1 + zan)\,/\,z$. In this case, $\mathrm{A}\,(n)$ diverges as $n \to \infty$.

Coldwell & Coddington (1994) discuss objectives in the modelling of species accumulation curves, and state that one of the main goals is often to predict the expected increase in species richness per additional unit of effort. The shape of the accumulation depends on the relative abundance of the sampled species. Some authors have adopted this fact implicitly (see Engen, 1978, for example). Although the notion of accumulation curve is not used (because attention is focused on estimating total number of species), the methods are based on abundance specified through explicit forms or probability distributions.

In this paper, we first propose a model using a different approach. We will postulate that the form of the accumulation curve depends on the *capture probabilities* of each species (that is, the probability that a given species is observed on a given trial). These probabilities depend, amongst other things, on the relative abundance present and the sampling method used. Additionally, the model will take into account the discrete nature of units of effort, the possibility of multiple occurrences of new species at a single unit of time, and the fact that capture may depend on random factors.

A few working hypotheses will be required in order to obtain analytical results. We will assume that the collection processes are independent in time; that each species is captured independently from all others, each with a certain probability; and that capturabilities are a realization of a random variable. This idea of random capturabilities is not new. However, no distribution is here justifiable on empirical means because no observed frequencies are available. We resort to a $\{\text{Beta}(\alpha,\, \beta),\, \alpha > 0,\, \beta > 0\}$ family on grounds of richness and because it yields analytical tractability. The beta family has previously been used by some authors (Seber 1982; Otis *et al.* 1978) for modelling capture probabilities.

In Section 2 we establish that the vector of new species captured in the first n times has a multinomial distribution, with parameters which are functions of $(k,\, \alpha,\, \beta)$, where k is the total number of species in the area and α, β are the parameters of the beta distribution. Maximum likelihood and profile likelihood methods are explored for the estimation of parameters.

An additional issue considered here is when collecting lists are incomplete. In many cases, a complete record of the collection process is not available. This may be considered a case of censoring, since there exist time lapses where observation was performed but accumulation detail has been lost, due for instance to a compounding effect (historical reporting of lists where units are months instead of days, for example). For estimation in these cases, we will use the EM algorithm (Dempster *et al.*, 1977).

To finalize, in Section 4 we analyze two collecting lists which illustrate the techniques developed. Results will be discussed and comparisons will be made with estimates based on other approaches.

2 Model formulation

2.1 Basic model

Let k denote the total number of different species present in an area. Its value is fixed, but unknown. The collection process consists in a sequence of trials indexed by n, $n = 1, 2, \ldots$, which may conveniently be referred to as a *time* index and represents the amount of effort . The probability of observing a species in unit time is a function of various factors, such as abundance, diversity, detectability, field worker's experience, and the method of observation. For modelling purposes, the totality of these factors may be considered as random, due to weather conditions or specific characteristics of the particular land area selected for sampling. For this reason, it is reasonable to consider that the probabilities that each species is caught or observed in a time unit are a realization of a random vector $\underset{\sim}{P} = (P_1,\, P_2,\, \ldots,\, P_k)$, having density supported on $[0,1]^k$.

For $i = 1,\, 2,\, ...,\, k$, in any time unit let

$$p_i = \text{Probability of capturing the species labeled } i,$$

where $(p_1,\, \ldots,\, p_k)$ is a single realization of $\underset{\sim}{P}$.

We will adopt the following assumptions as a starting point:

$$\begin{array}{ll} \text{(a)} & \text{Collection processes are independent in time.} \\ \text{(b)} & \text{At each time } n\text{, each species is observed independently} \\ & \text{from all others, with the same probabilities } p_1,\, p_2,\, \ldots,\, p_k. \\ \text{(c)} & P_1,\, P_2,\, \ldots,\, P_k \text{ are independent and identically distributed.} \end{array} \qquad (1)$$

Assumption (1b) is the most critical, because it indicates that capture probabilities remain constant in time. It is well-known that abundance in many species is related directly to periodicity in yearly weather, so this assumption is definitely not applicable in all cases. It will be valid to the extent in which the sampling period is not affected by the seasons or by the biological age of specimens. The role of assumption (1c) is to model capture probabilities as a group. We are not interested in the probability of capturing the i-th species of the population, p_i, but in their joint behavior, since it is this which determines the shape of the accumulation curve. Capture probabilities may be related to the notion of *abundance* (see Engen, 1978), through a mechanism which can be as simple as proportionality, or a far more complex process. The point here is that we are interested only in the final capture probabilities, disregarding any relationships which may be present with abundance. The Beta(α, β) distribution will be considered here because of its flexibility and analytical convenience. Seber (1982) and Otis *et al.* (1978) also mention the use of the Beta(α, β) distribution for modelling capture probabilities.

The fact of finding or not finding the i-th species at time n, will be described through a Bernoulli random variable. We use the notation $Y_i^{(n)}$ for such random variable, that is

$$Y_i^{(n)} \sim \mathrm{Ber}(p_i), \text{ for } n = 1, 2, ...; \, i = 1, 2, ..., k.$$

We also use the notation

$$X_n = \text{number of new species collected at time } n.$$

Variables X_n are then described in terms of $Y_i^{(j)}$ as follows:

$$\begin{aligned} X_0 &= 0, \\ X_1 &= \sum_{i=1}^{k} Y_i^{(1)}, \\ &\vdots \\ X_n &= \sum_{i=1}^{k} Y_i^{(n)}(1 - Y_i^{(n-1)}) \cdots (1 - Y_i^{(1)}). \end{aligned} \qquad (2)$$

In these terms, the *accumulation curve* is the plot of $\mathrm{S}(n)$ *vs.* n, where $\mathrm{S}(0) = 0$, and $\mathrm{S}(n) = \mathrm{S}(n-1) + X_n$, for $n = 1, 2, \ldots$. The random variable X_n is the size of the jump in the collection process at time n.

Under assumptions (1), the joint distribution of $(X_1, X_2, \ldots, X_n)$ is multinomial,

$$\mathrm{MN}_n \left(k,\, \mathrm{E}(P_1),\, \mathrm{E}\left(P_1\left(1 - P_1\right)\right),\, \ldots,\, \mathrm{E}(P_1\left(1 - P_1\right)^{n-1})\right).$$

That is, its probability density is

$$
\begin{aligned}
&\Pr\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) \\
= \quad & \frac{k!}{x_1!\, x_2! \cdots x_n!(k - \sum_{i=1}^n x_i)!} \left[\mathrm{E}(1-P_1)^n\right]^{k-\sum_{i=1}^n x_i} \prod_{i=1}^n \mathrm{E}^{x_i}\left[P_1(1-P_1)^{i-1}\right], \qquad (3)
\end{aligned}
$$

where $0 \leq x_i \leq k$ for all i and satisfy $0 \leq \sum_{i=1}^n x_i \leq k$. The proof is sketched in Appendix A. It is interesting to note that the collection process described is equivalent to a multinomial experiment. This interpretation considers each species as performing a sequence of k independent trials, each with $n+1$ mutually exclusive outcomes A_0, A_1, …, A_n—where A_i is the event "species is observed for the first time at time i" for $i \geq 1$ and A_0 is the event "species is not yet observed in n time units"—such that the probability of A_i occurring in each independent trial is $\mathrm{E}\left[P_1\left(1-P_1\right)^{i-1}\right]$ for $i \geq 1$.

Assuming that the capture probabilities are a realization of a random vector having independent and identically distributed Beta(α, β) entries —a situation which will hereafter be referred to as the "beta case"—, the joint density of $(X_1, X_2, \ldots, X_n)$ is reduced to

$$
\begin{aligned}
&\Pr\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) \qquad (4) \\
= \quad & \frac{k!}{x_1!\, x_2! \cdots x_n!(k - \sum_{i=1}^n x_i)!} \left[\frac{\mathrm{b}\left(\alpha, \beta+n\right)}{\mathrm{b}\left(\alpha, \beta\right)}\right]^{k-\sum_{i=1}^n x_i} \prod_{i=1}^n \left[\frac{\mathrm{b}\left(\alpha+1, \beta+i-1\right)}{\mathrm{b}\left(\alpha, \beta\right)}\right]^{x_i},
\end{aligned}
$$

where $\mathrm{b}\left(\cdot, \cdot\right)$ is the beta function. This density function is the basis for inferential procedures which will be addressed in Section 3. A few useful consequences of these results in the beta case follow.

For fixed n, the value of the accumulation curve $\mathrm{S}(n)$ has a binomial distribution,

$$
\mathrm{S}(n) \sim \mathrm{Bin}\left(k,\, 1 - \frac{\mathrm{b}\left(\alpha, \beta+n\right)}{\mathrm{b}\left(\alpha, \beta\right)}\right),
$$

and thus

$$
\begin{aligned}
\mathrm{E}\left(\mathrm{S}(n)\right) &= k\left(1 - \frac{\mathrm{b}\left(\alpha, \beta+n\right)}{\mathrm{b}\left(\alpha, \beta\right)}\right), \text{ and} \\
\mathrm{V}\left(\mathrm{S}(n)\right) &= k\left(\frac{\mathrm{b}\left(\alpha, \beta+n\right)}{\mathrm{b}\left(\alpha, \beta\right)}\right)\left(1 - \frac{\mathrm{b}\left(\alpha, \beta+n\right)}{\mathrm{b}\left(\alpha, \beta\right)}\right).
\end{aligned} \qquad (5)
$$

The value of $\mathrm{E}\left(\mathrm{S}(n)\right)$ as a function of n is comparable in meaning to the values of $\mathrm{A}(n)$ which are produced by the SL approach.

Let s be the number of species observed until time N. If $Z_M = \sum_{i=N+1}^{N+M} X_i$ is the number of species which would be collected in time units $N+1$ to $N+M$, then

$$
Z_M \mid N_1, \ldots, N_N \sim \mathrm{Bin}\left(k-s,\, 1 - \frac{\mathrm{b}(\alpha,\, \beta+N+M-1)}{\mathrm{b}(\alpha,\, \beta+N)}\right).
$$

A plot of

$$
\mathrm{E}\left[Z_M \mid N_1, \ldots, N_N\right] = (k-s)\left(1 - \frac{\mathrm{b}(\alpha,\, \beta+N+M-1)}{\mathrm{b}(\alpha,\, \beta+N)}\right) \qquad (6)
$$

versus M would be a complementary device useful in deciding if a given value of M is rewarding in terms of collection effort.

3 Inference

We now establish inference procedures for the stated model. Data is $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, where the vector $(X_1, X_2, \ldots, X_n)$ has density given by (4). Since this density is given explicitly, we will consider estimation via maximum likelihood, and because interest lies mainly in the parameter k, we also turn to the notion of profile likelihood. When censored data is present, the EM algorithm is a plausible way of obtaining parameter estimates.

All of these methods require optimization of non-linear functions in dimension three, at most. In our experience, the method of conjugate gradients (Press *et al.,* 1988) was adequate, although careful grid selections had to be considered in searching for starting points. Derivatives of the log-likelihood are useful for this purpose and are stated in Appendix B.

3.1 Likelihood and profile likelihood

Let $x_1, x_2, \ldots, x_n$ be a realization of the random variables $X_1, X_2, \ldots, X_n$ whose density is given by (4), and let $s = \sum x_i$. The log-likelihood (disregarding a constant additive term) is given by

$$\begin{aligned} \mathrm{l}\,(k, \alpha, \beta) \quad &= \quad \log\,(k!) - \log\,((k-s)!) - (k-s)\log\left[\frac{\mathrm{b}\,(\alpha, \beta+n)}{\mathrm{b}\,(\alpha, \beta)}\right] \\ &\quad + \sum_{i=1}^{n} x_i \log\left[\frac{\mathrm{b}\,(\alpha+1, \beta+i-1)}{\mathrm{b}\,(\alpha, \beta)}\right]. \end{aligned} \tag{7}$$

Maximization of this quantity provides point estimates of parameter values and functions of them. In particular, the quantities described towards the end of Section 2.1, which are of special interest in the biological context, may be estimated (by a substitution principle). However, precision and confidence intervals are not as straightforward to obtain, because the problem is highly non-regular. In Monte Carlo experimentation, we have found conclusive evidence that the non-regularity in this setting is such that normal approximations are not really useful for inferential procedures, and bootstrap intervals are preferred instead.

Since α and β may be considered nuisance parameters, and k the parameter of interest, *profile likelihood* seems a plausible tool to use for inference (Hinkley *et al.,* 1991; Barndorff-Nielsen & Cox, 1994). We here only state material necessary for constructing confidence intervals for k. The profile likelihood is

$$\mathrm{L}_P\,(k) = \mathrm{L}\left(k, \hat{\alpha}_k, \hat{\beta}_k\right) = \sup_{\alpha, \beta} \mathrm{L}(k, \alpha, \beta),$$

where $\mathrm{L}(k, \alpha, \beta)$ is the likelihood and $\hat{\alpha}_k$, $\hat{\beta}_k$ are the maximum likelihood estimators of α and β for fixed k. Inference on k is based on treating $\mathrm{L}_P\,(k)$ as if it were a legitimate likelihood. The log-profile likelihood is

$$\mathrm{l}_P\,(k) = \sup_{\alpha, \beta} \mathrm{l}(k, \alpha, \beta)$$

where $\mathrm{l}(k, \alpha, \beta)$ is defined by (7). An approximate $(1-a) \times 100\%$ confidence interval for k based on profile likelihood is

$$\mathrm{R}_P = \left\{ k : -2 \log \frac{\mathrm{L}_P \left(k, \hat{\alpha}_k, \hat{\beta}_k\right)}{\mathrm{L}_P \left(k, \hat{\alpha}_{\hat{k}}, \hat{\beta}_{\hat{k}}\right)} \leq \chi^2_{1,1-a} \right\}, \tag{8}$$

(see Barndorff-Nielsen & Cox, 1994). Correct coverage probabilities rely on the fact that the profile likelihood ratio statistic,

$$\mathrm{r}_P\left(k\right) = -2 \log \frac{\mathrm{L}_P \left(k, \hat{\alpha}_k, \hat{\beta}_k\right)}{\mathrm{L}_P \left(k, \hat{\alpha}_{\hat{k}}, \hat{\beta}_{\hat{k}}\right)}$$

has an asymptotic χ^2_1 distribution. We have performed a few simulations (reported in Section 4.1) to examine the distribution of $\mathrm{r}_P\left(k\right)$. The conclusion is that the distribution of $\mathrm{r}_P\left(k\right)$ is very similar to a χ^2_1 distribution, and more so when α, β are large for fixed k and for large k for fixed (α, β). We believe these results suggest that the χ^2_1 approximation is reasonably good for practical applications. Future research may consider the possibility of modifying the profile likelihood, as in the conditioning argument of Cox & Reid (1987), the modification of Barndorff-Nielsen (1986), or the adjustment of McCullagh & Tibshirani (1990).

3.2 Bootstrapping methods

An alternative technique for estimating the covariance matrix and other measures of precision of estimators is the bootstrap (Efron, 1982). Here, we present the implementation of the method for building confidence intervals for functions of the parameters in (4).

Let $\boldsymbol{\theta} = (k, \alpha, \beta)$ and let $\mathrm{F}\left(\boldsymbol{\theta}\right)$ denote the multinomial distribution (4); the data vector $\mathbf{y} = (x_1, x_2, \ldots, x_n)$ is a sample from $\mathrm{F}\left(\boldsymbol{\theta}\right)$. Suppose interest is in $\mathrm{g}(\boldsymbol{\theta})$ for some real-valued function g. Based on $\mathbf{y}$, one computes $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimator, and interest lies in evaluating the precision of $\mathrm{g}(\hat{\boldsymbol{\theta}})$ as an estimator of the true value of $\mathrm{g}(\boldsymbol{\theta})$. Computation of $\hat{\boldsymbol{\theta}}$ is achieved numerically, and bootstrap intervals are found as follows (Efron & Tibshirani, 1986):

(i) Find $\hat{\boldsymbol{\theta}}$ using data $\mathbf{y}$;

(ii) Simulate B independent vectors $\mathbf{y}^*\left(1\right), \mathbf{y}^*\left(2\right), \ldots, \mathbf{y}^*\left(B\right)$ from the multinomial distribution $\mathrm{F}\left(\hat{\boldsymbol{\theta}}\right)$, and for each of them compute $\mathrm{g}(\hat{\boldsymbol{\theta}}^*\left(b\right))$, the maximum likelihood estimator of $\mathrm{g}(\boldsymbol{\theta})$ based on $\mathbf{y}^*\left(b\right)$, $b = 1, 2, \ldots, B$.

(iii) Having produced $\mathrm{g}(\hat{\boldsymbol{\theta}}^*\left(1\right))$, $\mathrm{g}(\hat{\boldsymbol{\theta}}^*\left(2\right))$, $\ldots$, $\mathrm{g}(\hat{\boldsymbol{\theta}}^*\left(B\right))$, define $\hat{\mathrm{G}}$, the empirical distribution function of $\{\mathrm{g}(\hat{\boldsymbol{\theta}}^*(b)),\ b = 1, 2, 3, \ldots, B\}$, as

$$\hat{\mathrm{G}}\left(s\right) = \frac{\#\left(\mathrm{g}(\hat{\boldsymbol{\theta}}^*(b)) \leq s\right)}{B}.$$

This last empirical distribution is regarded as an approximation to $\Pr\{\mathrm{g}(\hat{\boldsymbol{\theta}}) \leq s\}$, and motivates the following approximate $(1-2a) \times 100\%$ confidence interval for $\mathrm{g}(\boldsymbol{\theta})$ (Efron, 1982, Sec. 10.4):

$$\left[\hat{\mathrm{G}}^{-1}(a)\,,\ \hat{\mathrm{G}}^{-1}(1-a)\right],$$

where a, $0 \le a \le 0.5$, is fixed.

Under model (4), the simulation in step (ii) above consists in generating $p_1^*, p_2^*, \ldots, p_{\hat{k}}^*$ as a random sample from a Beta $\left(\hat{\alpha},\ \hat{\beta}\right)$ distribution, and then simulating the process (2). Equivalently, B independent draws of vectors $\mathbf{y}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ may be made from the

$$\mathrm{MN}_n\left(\hat{k},\ \frac{\mathrm{b}(\hat{\alpha}+1,\hat{\beta})}{\mathrm{b}(\hat{\alpha},\hat{\beta})},\ \frac{\mathrm{b}(\hat{\alpha}+1,\hat{\beta}+1)}{\mathrm{b}(\hat{\alpha},\hat{\beta})},\ldots,\ \frac{\mathrm{b}(\hat{\alpha}+1,\hat{\beta}+n-1)}{\mathrm{b}(\hat{\alpha},\hat{\beta})}\right)$$

distribution.

3.3 Censored observations: EM algorithm

When censored or missing values occur, likelihoods may be maximized using the EM algorithm, formalized by Dempster, Laird, & Rubin (1977). Using their notation, let $\mathbf{x} = (\mathbf{y},\, \mathbf{z})$ denote observed data $\mathbf{y}$, augmented with missing data $\mathbf{z}$. Let $\mathrm{f}\,(\mathbf{x};\boldsymbol{\theta})$ and $\mathrm{g}\,(\mathbf{y};\boldsymbol{\theta})$ be families of densities indexed by the parameter $\boldsymbol{\theta}$. The relationship between f and g is given by $\mathrm{g}\,(\mathbf{y};\boldsymbol{\theta}) = \int_{\mathcal{X}(\mathbf{y})} \mathrm{f}\,(\mathbf{x};\boldsymbol{\theta})\, d\mathbf{x}$, where $\mathcal{X}\,(\mathbf{y})$ is the set of all possible values for missing data given a fixed value of $\mathbf{y}$. The aim of the EM algorithm is to maximize $\mathrm{g}\,(\mathbf{y};\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, but using $\mathrm{f}\,(\mathbf{x};\boldsymbol{\theta})$ through a different type of relationship.

Define the function

$$\mathrm{Q}\,(\boldsymbol{\theta}',\boldsymbol{\theta}) = \mathrm{E}\left[\log \mathrm{f}\,(\mathbf{x};\boldsymbol{\theta}') \mid \mathbf{y},\boldsymbol{\theta}\right].$$

The iteration $\boldsymbol{\theta}^{(p)} \to \boldsymbol{\theta}^{(p+1)}$, which is called an EM step, is defined as follows:

E Step: Compute $\mathrm{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(p)})$.

M Step: Choose $\boldsymbol{\theta}^{(p+1)}$ as the value of $\boldsymbol{\theta}$ which maximizes $\mathrm{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{(p)})$.

The heuristic idea behind this is that we would like to choose $\boldsymbol{\theta}^*$ to maximize $\log \mathrm{f}\,(\mathbf{x};\boldsymbol{\theta})$. Since we cannot compute $\log \mathrm{f}\,(\mathbf{x};\boldsymbol{\theta})$ because we do not observe all of $\mathbf{x}$, we instead maximize its expected value given $\mathbf{y}$ and the current value of $\boldsymbol{\theta}$. Iteration produces convergence to the maximum likelihood estimate.

3.3.1 Implementation for the beta case

Let $(m_1, \mathrm{s}(m_1))$, $(m_2, \mathrm{s}\,(m_2))\,,\ \ldots,\ (m_r, \mathrm{s}\,(m_r))$ be the observed accumulation curve. The iteration prescribed by the EM algorithm is described as follows. A brief justification of these steps is included in Appendix C.

E Step: First define the quantities $x_1, \ldots, x_n$ as follows.

For $i = 1, 2, \ldots r$, do:

If $m_i = m_{i-1} + 1$, set

$$x_{m_i} = \mathrm{s}(m_i) - \mathrm{s}(m_{i-1}).$$

Otherwise, for $j = m_{i-1} + 1,\ m_{i-1} + 2,\ \ldots,\ m_i$, set

$$x_j = [\mathrm{s}\,(m_i) - \mathrm{s}\,(m_{i-1})]\,\frac{e_j}{\displaystyle\sum_{m_{i-1}<j\le m_i} e_j},$$

where

$$e_j = \frac{\mathrm{b}\,(\alpha^p + 1, \beta^p + j - 1)}{\mathrm{b}\,(\alpha^p, \beta^p)}.$$

Then define

$$\begin{aligned} & \mathrm{Q}^*\left((k,\alpha,\beta)\,,\,(k^p,\alpha^p,\beta^p)\right) \\ = \; & \log(k!) - \log\left((k - s\,(m_r))!\right) - (k - \mathrm{s}\,(m_r))\log\left[\frac{\mathrm{b}\,(\alpha,\beta+m_r)}{\mathrm{b}\,(\alpha,\beta)}\right] \\ & + \sum_{i=1}^{m_r} x_i \log\left[\frac{\mathrm{b}\,(\alpha+1,\beta+i-1)}{\mathrm{b}\,(\alpha,\beta)}\right]. \end{aligned}$$

M Step: Choose $(k^{p+1}, \alpha^{p+1}, \beta^{p+1})$ as the value of (k, α, β) which maximizes

$$\mathrm{Q}^*\left((k,\ \alpha,\ \beta)\,,\,(k^p,\ \alpha^p,\ \beta^p)\right).$$

4 Simulations and examples

This final section is intended to show distributional properties of the profile likelihood ratio statistic via simulation, and two examples previously studied by SL. The first example regards a non-censored collecting list of bats, and the second is a censored list of butterfly data which will therefore be analyzed using the EM algorithm.

4.1 Distribution of r_P and confidence intervals for k

This Monte-Carlo experiment shows the distribution of r_P and the empirical coverage probability of confidence intervals based on this statistic. Consider the eight cases obtained from combinations of $n = 100$, 150; $k = 100$, 150; and $(\alpha,\ \beta) = (.5,\ .8)\,,\ (5,\ 250)\,.$ Each of 1000 replications consists in simulating an accumulation curve, computing r_P, and finding the confidence region R_P defined by (8) with $a = 0.05$. The empirical coverage probability is then the proportion of times R_P contains k.

Histograms for values of r_P prove to be quite close to superimposed plots of χ^2_1 densities. The empirical distributions are fairly close to a χ^2_1 for all configurations, with Beta $(5, 250)$ giving the best agreements. There is a slight tendency to heavier tails than expected, but this effect is not serious. Table 1 sheds information in terms of sample means. The mean of r_P is closer to 1 (the mean of the χ^2_1 distribution) for $(\alpha, \beta) = (5,\ 250)$ for the same value of k. For these same cases, coverage is closer to 95%, which is the intended level.

Table 1: Summary results of 1000 simulations of the profile likelihood ratio statistic under eight parameter configurations.

Parameters	n	mean	Coverage Probability
$\alpha = 0.5 \quad \beta = 0.8$	100	1.2429	91.4
$k = 100$	150	1.2491	91.9
$\alpha = 0.5 \quad \beta = 0.8$	100	1.1348	93.6
$k = 150$	150	1.1805	93.4
$\alpha = 5 \quad \beta = 250$	100	1.0771	94.7
$k = 100$	150	1.0074	94.6
$\alpha = 5 \quad \beta = 250$	100	1.085	93.6
$k = 150$	150	1.1387	94.6

4.2 Examples

4.2.1 Chajul bat data

This example considers the list of bat species at Chajul Biological Station in southeastern Mexico. The list was analyzed previously by SL; these authors report 50 observed species after setting nets on 49 nights. The unit of effort in this example is *sampled night.*

We use maximum likelihood for parameter estimation. Estimated total number of species in the area is 54, while the estimated values of the beta parameters describing capture probabilities are $\hat{\alpha} = 24.00$ and $\hat{\beta} = 441.49$. Figure 1 shows the Beta$(\hat{\alpha}, \hat{\beta})$ distribution. It is almost symmetric and concentrated about 0.06, which suggests that all bat species roughly share a common capture probability and are fairly prone to be captured. This is, there do not exist species which are either extremely rare or extremely hard to capture.

Figure 2 shows the profile likelihood function, $l_p(k)$, and the 95% confidence interval for the total number of species, k. The 95% interval is $(50, 68]$. Because of empirical evidence and since estimates are suggesting large values of α and β, the actual coverage probability of this interval is prone to be near 95%.

It is interesting to compare these results with those found by SL. Their "Clench" and "exponential" models produce estimates 85 and 56 for k, respectively. Using biological arguments, they favor the exponential model, so the estimate of k which we find using profile likelihood agrees with the previous analysis, with the added feature of confidence interval being stated.

Figure 3 shows the estimated expected value (5) and the conditional expectation (6) for $M \geq 50$, using maximum likelihood parameter estimates. In this example, both of these are very close to each other, the reason being that the estimated value of the accumulation curve at time 49 is very close to the actual value observed. The following statements refer to estimated expected value and estimated conditional expected value (in that order), and show the type of answers which can be obtained: at time 79, estimated values are 52.86 and 53.00, that is, one expects to add 3 new species to the list after 30 additional units of effort.

At time 99, estimated values are 53.53 and 53.59, that is, 50 additional units are required to achieve (in the sense of an average) the estimated total number of species present.

4.2.2 Pakitza butterfly data

SL analyzed butterfly data obtained in a set of 200 person-hours (in September 1989) at Pakitza Biological Station, Peru. Considering the Clench, exponential and logarithmic models (see Section 1), produces indistinguishable fits over the range of observed data, but extrapolations are strikingly different (see Figure 4). SL argue in favor of the logarithmic fit because of the biological complexity and the large extent of the area.

In this example, the complete history of the collection process is not available. There are 14 points in time (person-hours) where the height of the collecting curve is known. This data set provides an example of a censored list, in the sense of Section 3.3. The parameter estimates using the EM algorithm (Section 3.3) in the multinomial model are $\hat{k} = 981$, $\hat{\alpha} = .68$ and $\hat{\beta} = 74.99$. The estimated total number of species is thus 981, compared with the asymptotic values of 890 and 611 obtained in the Clench and exponential models, respectively.

The value 981 would seem, at first sight, to be very extreme when compared to either of the estimates obtained by the regression approach. In this respect, there exists an interesting posterior observation: after continuation for 565 person-hours, the species accumulation count stood at 979. This value is quite close to the estimate obtained by our model, and exceeds the values obtained by the Clench and exponential models. The logarithmic model lacks an asymptote, which makes it hard to interpret over the extrapolation range. Figure 4 shows the three models fitted by SL, as well as the estimated expected accumulation curve in our model computed with estimated parameter values.

In Figure 5 we plot the Beta$(\hat{\alpha}, \hat{\beta})$ density function. The shape of this distribution, in contrast with the previous example regarding bats, suggests that the Pakitza ecosystem has more complex characteristics, that is, that it is a region with much greater biodiversity and geographical extension. Here, many species are hard to be sighted, and few species have a high probability of being observed. This produces an accumulation curve which grows quite rapidly in the beginning, but later requires very large amounts of effort to achieve further significant increase.

Figure 6 shows plots of estimated expected accumulation and conditional expectation. In this situation, these functions are notoriously different.

We observe that the estimated expected value of accumulation is 574.86 at the next time unit, while the estimated conditional is 568.82, that is, 0.82 species on the next sighting attempt. If the collection process were continued for 150 additional time units, we find expectation at 718.71, while the conditional expectation is 724.12 species. At time 565, at which future sampling reports 979 species, the model yields 751.97 as an estimate of the mean accumulation curve. This value seems too low. Although we do not have further information regarding the subsequent study, a possible reason for this is that the original study used a different sampling procedure. This suggests one possible extension of the model, of immediate application: that capture probabilities vary in time (in this case, it is possible that capture probabilities changed because of a change in collection crews and/or methods).

5 Conclusion

The model presented in this paper constitutes an alternative for solving the problem of modelling the relationship between effort and number of accumulated species. Taking into account the discrete nature of units of effort and considering that only capture probabilities are directly relevant, make the model more realistic and easily interpreted. The availability of a formula for conditional expected values of the collection process given the observed curve, is an advantage. This seems an adequate concept, in addition to the fact that the relationship between effort and capture is modelled.

On the other hand, it is important to evaluate precision of inferential procedures. Quantities such as estimators of conditional expectations, which are usually not considered, are easy to compute under this approach. Because censored collecting curves often appear, the solution in this case is also an important advantage. Results in a number of real and simulated data sets are very encouraging.

A disadvantage is that the bootstrap and EM methods here described require fairly sophisticated numerical methods which are not readily available in commercial software. The approach of SL, in contrast, uses simple least-squares procedures which are commonly available.

One interesting problem which remains open is the exploration of corrections to the profile likelihood procedures used for inference on the parameter k. However, at a modelling stage, a more compelling issue is the generalization to capture probabilities which are more general. The model we have stated in this paper is applicable in the event that capture probabilities remain approximately constant over the span of the collection, and in a large number of applications, this situation is not necessarily true.

Correlations between entries in the vector P give rise to intractable expressions, but perhaps the main reason for not considering them here is that postulating a correlation function seems to be a difficult task. On the other hand, it is certainly natural to conceive that capture probabilities may vary in time. Possible reasons for this are seasonal variations in abundance and diversity due to weather and immigration, or the field worker gathering experience or changing a method of capture. A possible generalization of the present model is to consider that probabilities are realizations of random variables with time-dependent densities. This suggests to model non-stationary capture distributions through a family indexed by time, perhaps with the inclusion of additional parameters. A brief argument would show that the resulting distribution is again multinomial.

Acknowledgments

Miguel Nakamura's work was partially supported by CONACYT grant number 4224E9405. David Sprott suggested the shorter and more transparent proof of the main multinomial result. We also thank Jorge Soberón for relevant discussion at the modeling and interpretation stages.

A Proof of main result

Let $Z_{ij} = 1$ if species i is observed for the first time on trial j, $1 \le j \le n$, and $Z_{ij} = 0$ otherwise. Let $Z_{i0} = 1$ if species i is not yet caught by the n-th trial, $1 \le i \le k$ and 0 otherwise. For $1 \le i \le k$ and $1 \le j \le n$ we define the conditional probabilities

$$\begin{aligned}\pi_{ij} &= \Pr(Z_{ij} = 1 \mid P_i = p_i) = p_i(1-p_i)^{j-1}, \text{ and}\\ \pi_{i0} &= \Pr(Z_{i0} = 1 \mid P_i = p_i) = (1-p_i)^n.\end{aligned}$$

Let P have density g. The marginal probabilities are $\Pr(Z_{ij} = 1) = \int_{p=0}^{1} \pi_{ij}\, \mathrm{g}(p)\, dp = \mathrm{E}(\pi_{ij})$, for $1 \le i \le k$ and $0 \le j \le n$. Because for fixed i only one of the variables $Z_{i0}, Z_{i1}, \ldots, Z_{in}$ takes on the value 1 while the rest are all 0, it can be easily checked that the joint probability generating function (pgf) of this set of random variables is

$$\varphi_{Z_{ij}}(s_0, \ldots, s_n) = \mathrm{E}(\pi_{i0})s_0 + \cdots + \mathrm{E}(\pi_{in})s_n.$$

The numbers of new species caught on each trial are $X_j = \sum_{i=1}^{k} Z_{ij}$, $1 \le j \le n$. Assuming independence of species, the joint pgf of these variables is

$$\varphi_{X_i}(s_0, \ldots, s_n) = \varphi_{Z_{ij}}^{k}(s_0, \ldots, s_n) = [\mathrm{E}(\pi_{i0})s_0 + \cdots + \mathrm{E}(\pi_{in})s_n]^k,$$

which is identified as the pgf of the multinomial distribution with index k and $n+1$ classes having probability parameters $\mathrm{E}(\pi_{ij})$. When each P_i has the same distribution, these are $\mathrm{E}(\pi_{ij}) = \mathrm{E}[P(1-P)^{j-1}]$, for $j \ge 1$ and $\mathrm{E}(\pi_{i0}) = \mathrm{E}[(1-p)^n]$.

B Expressions for derivatives

The gradient of the log-likelihood (7) is

$$\begin{aligned}\frac{\partial \mathrm{l}\,(k,\alpha,\beta)}{\partial k} &= \sum_{i=0}^{s-1} \frac{1}{k-i} + \ln\Gamma(\alpha+\beta) + \ln\Gamma(\beta+n) - \ln\Gamma(\beta) - \ln\Gamma(\alpha+\beta+n),\\ \frac{\partial \mathrm{l}\,(k,\alpha,\beta)}{\partial \alpha} &= -(k-s)\sum_{j=0}^{n-1} \frac{1}{j+(\alpha+\beta)} - \sum_{i=1}^{n} x_i \sum_{j=0}^{i-1} \frac{1}{j+(\alpha+\beta)} + \frac{s}{\alpha},\\ \frac{\partial \mathrm{l}\,(k,\alpha,\beta)}{\partial \beta} &= (k-s)\sum_{j=0}^{n}\left(\frac{1}{j+\beta} - \frac{1}{j+(\alpha+\beta)}\right) + \sum_{i=2}^{n} x_i \sum_{j=0}^{i-2} \frac{1}{j+\beta}\\ &\quad - \sum_{i=1}^{n} x_i \sum_{j=0}^{i-1} \frac{1}{j+(\alpha+\beta)}.\end{aligned}$$

C Steps in the EM algorithm

The E step is based on a decomposition of the multinomial distribution. Let $(m_1, \mathrm{s}(m_1))$, $(m_2, \mathrm{s}(m_2))$, $\ldots$, $(m_r, \mathrm{s}(m_r))$ be the observed accumulation curve and let $x_1, x_2, \ldots, x_n$ constitute "complete data". We define the following notation:

For $i = 1, 2, \ldots, r$ and $m_{i-1} < j \leq m_i$, let $x_{ij} = x$ and

$$p_{ij} = \frac{\mathrm{b}\,(\alpha + 1, \beta + j - 1)}{\mathrm{b}\,(\alpha, \beta)}.$$

In addition, $x_{i+} = \sum_j x_{ij} = s(m_i) - s\,(m_{i-1})\,;\; x_{++} = \sum_i x_{i+};\; p_{i+} = \sum_j p_{ij}$; and $p_{++} = \sum_i p_{i+}$.

With this notation, the likelihood for complete data is a product of multinomial densities:

$$\begin{aligned}
& \frac{k!}{(k - x_{++})!\prod_{i,j} x_{ij}!} p_{++}^{k - x_{++}} \prod_{i,j} p_{ij}^{x_{ij}} \\
= & \prod_i \left[\frac{x_{ij}!}{\prod_j x_{ij}!} \prod_j \left(\frac{p_{ij}}{p_{i+}} \right)^{x_{ij}} \right] \times \frac{k!}{(k - x_{++})!\prod_i x_{i+}!} p_{++}^{k - x_{++}} \prod_i p_{i+}^{x_{i+}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
Q\,(\boldsymbol{\theta}', \boldsymbol{\theta}) &= \mathrm{E}\left[\log f\,(\mathbf{x} \mid \boldsymbol{\theta}') \mid \mathbf{y}, \boldsymbol{\theta}\right] \\
&= -\sum_i \sum_j \mathrm{E}\left[\ln\,(x_{ij}!) \mid \mathbf{y}, \boldsymbol{\theta}\right] + \sum_i \sum_j \ln\,(p_{ij})\,\mathrm{E}\left[x_{ij} \mid \mathbf{y}, \boldsymbol{\theta}\right] + \ln k! \\
&\quad - \mathrm{E}\left[\ln (k - x_{++})! \mid \mathbf{y}, \boldsymbol{\theta}\right] + \ln\,(p_{++})\,\mathrm{E}\left[(k - x_{++}) \mid \mathbf{y}, \boldsymbol{\theta}\right].
\end{aligned}$$

Because x_{i+} and x_{++} are actually observed, we have

$$\begin{aligned}
\mathrm{Q}\,(\boldsymbol{\theta}', \boldsymbol{\theta}) &= \ln k! - \ln (k - x_{++})! + \ln\,(p_{++})\,(k - x_{++}) \\
&\quad + \sum_i \sum_j \ln\,(p_{ij})\,\mathrm{E}\left[x_{ij} \mid \mathbf{y}, \boldsymbol{\theta}\right] - \sum_i \sum_j \mathrm{E}\left[\ln\,(x_{ij}!) \mid \mathbf{y}, \boldsymbol{\theta}\right].
\end{aligned}$$

On the other hand, x_{ij} has the Bin $(x_{i+},\, p_{ij}/p_{i+})$ distribution. Therefore,

$$x_{ij}^* := \mathrm{E}\left[x_{ij} \mid \mathbf{y}, \boldsymbol{\theta}\right] = \begin{cases} x_{ij}, \text{ if } x_{ij} \text{ was observed} \\ x_{i+} \dfrac{p_{ij}}{p_{i+}}, \text{ otherwise} \end{cases}.$$

Since the term $\sum_i \sum_j \mathrm{E}[\ln\,(x_{ij}!) \mid \mathbf{y}, \boldsymbol{\theta}]$ is only a function of observations and current parameter values, it can be ignored in the M step of the algorithm. In other words, to maximize $\mathrm{Q}\,(\boldsymbol{\theta}', \boldsymbol{\theta})$, is equivalent to maximizing

$$\mathrm{Q}^*\,(\boldsymbol{\theta}', \boldsymbol{\theta}) = \ln k! - \ln (k - x_{++})! + (k - x_{++}) \ln\,(p_{++}) + \sum_i \sum_j \ln\,(p_{ij})\, x_{ij}^*.$$

It is interesting to note that this $\mathrm{Q}^*\,(\boldsymbol{\theta}', \boldsymbol{\theta})$ is very much related to equation (4), which is the log-likelihood of the complete model. Here, non-observed values were replaced by expected values.

References

[1] Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics.* Monographs on Statistics and Applied Probability 52. Chapman & Hall.

[2] Bunge, J. and Fitzpatrick, M. (1993). Estimating the Number of Species: A Review. *Journal of the American Statistical Association,* **88**, 364–373.

[3] Coldwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity trough extrapolation. *Phil. Trans. R. Soc. Lond.* B. Special volume on "The quantification and estimation or organismal biodiversity", (In press). D. L. Hawkswort, ed.

[4] Cox, D. R. and Reid, N. (1992). A note on then difference between profile and modified profile likelihood. *Biometrika*, **79**, 2, 408–411.

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

[6] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics. Philadelphia, Pennsylvania.

[7] Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science,* **1**, 1, 54–77.

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

[9] Engen, S. (1978). *Stochastic Abundance Models.* Monographs on Applied Probability and Statistics. Chapman and Hall.

[10] Hinkley, D. V., Reid, N. and Snell, E. J. (1991). *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS.* Chapman and Hall.

[11] McCullagh, P. and Tibshirani, R. (1990). A Simple Method for the Adjustment of Profile Likelihoods. *Journal of the Royal Statistical Society,* B **52**, 2, 325–344.

[12] Otis, D. L., Burnham, K. P., White, G. C. and Anderson, D.R. (1978). *Wildlife Monographs: Statistical Inference from Capture Data on Closed Animal Populations.* The Wildlife Society, Inc.

[13] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1988). *Numerical Recipes in C.* Cambridge University Press.

[14] Seber, G. A. F. (1982). *The Estimation of Animal Abundance and related parameters (Second Edition).* Charles Griffin & Company Limited.

[15] Soberón, J. and Llorente, J. (1993). The Use of Species Accumulation Functions for the Prediction of Species Richness. *Conservation Biology,* **7**, 3, 480–488.