CatchAll

Kendrick Li

December 2018

1 Introduction

CatchAll is a procedure for computing species richness C (total number of species) from a family of models and obtaining the associated standard error. The given data are $\{(i, f(i)), i \leq 1\}$, where $f(i)$ is the number of sample classes of size i. The current implementation of CatchAll chooses the richness estimate from 12 existing richness estimators. Model selection for parametric models is based on AIC, χ^2 and other criteria.

2 Notation

We denote the unknown number of total classes by C and the number of observed classes by c. We denote the number of unobserved classes by C_0. We use $X_i, i = 1, ..., c$ to denote the count of individuals in class i and $n \equiv \sum_{i=1}^{\infty} X_i$. Note that $c = \sum_{i=1}^{\infty} f(i)$ and $n = \sum_{i=1}^{\infty} if(i)$. Normally the estimation is based on frequency counts with incidence $i \leq \tau$ where τ is the cut-off. We denote $c_\tau = \sum_{i=1}^{\tau} f(i)$ and $n_\tau = \sum_{i=1}^{\tau} if(i)$.
The standard error is computed based on an asymptotic approximation:

$$Var(\hat{C}) \approx \sum_{i \leq 1} \sum_{j \leq 1} \frac{\partial \hat{C}}{\partial f_i} \frac{\partial \hat{C}}{\partial f_j} \hat{cov}(f_i, f_j)$$

where

$$\hat{cov}(f_i, f_j) = \begin{cases} f_i(1 - \frac{f_i}{\hat{C}}), & \text{if } i = j \\ -\frac{f_i f_j}{\hat{C}}, & \text{if } i \neq j \end{cases}$$

3 Richness estimators

3.1 Poisson Model

With zero-inflated Poisson model, the counts of the taxa have pmf

$$P(X_i = x) = \frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})}.$$

The score function of observed counts is given by

$$\dot{l}(\lambda) = \frac{\sum if(i)}{\lambda} - \frac{c_\tau}{1 - e^{-\lambda}}$$

The MLE $\hat{\lambda}$ solves the equation

$$\frac{1 - e^{-\lambda}}{\lambda} - \frac{c_\tau}{\sum if(i)} = 0$$

The MLE of C under poisson model is thus given by

$$\hat{C} = c_\tau/(1 - \exp(-\hat{\lambda})) + c - c_\tau = c_\tau/(\exp(\hat{\lambda}) - 1) + c$$

and the approximate variance of $\hat{C}$ is given by

$$\hat{var}(\hat{C}) = \frac{c_\tau/(1 - \exp(-\hat{\lambda}))}{\exp(\hat{\lambda}) - 1 - \hat{\lambda}}.$$

The model fit is evaluated by goodness-of-fit statistic and AIC. The goodness of fit statistic is given by

$$W = \sum_{k=1}^{\tau} \frac{(O_k - E_k)^2}{E_k}$$

where O_k is the observed frequency count with frequency k and E_k is the expected frequency count. We have

$$O_k = f_k$$

and

$$E_k = \frac{c_\tau}{1 - \exp(-\hat{\lambda})} \times \frac{\hat{\lambda}^k}{k!} \exp(-\hat{\lambda}) = \frac{c_\tau \hat{\lambda}^k}{(\exp(\hat{\lambda}) - 1)k!}.$$

Starting from the lowest cell frequency $k = 1$, cells are binned one by one until the expected count is greater than or equal to five. If the last binned cell has an expected count less than five, then the last two cells are binned.

3.2 Geometric Model

The Geometric Model assumes the species abundances follow a geometric (exponential-mixed Poisson) distribution:

$$P(X = j) = \frac{1}{1 + \theta} \left(\frac{\theta}{1 + \theta} \right)^j,$$

for $j = 0, 1, 2, ...$ Thus the observed species abundances follow a zero-inflated geometric distribution with success probability θ:

$$P(X = j | X > 0) = \frac{1}{\theta} \left(\frac{\theta}{1 + \theta} \right)^j,$$

for $j = 1, ..., \tau$. The MLE of θ is

$$\hat{\theta} = \frac{n_\tau}{c_\tau} - 1.$$

Then

$$\begin{aligned}
\hat{C} &= \frac{c_\tau}{1 - P(X = 0|\hat{\theta})} + c - c_\tau \\
&= \frac{c_\tau}{1 - \frac{1}{1+\hat{\theta}}} \\
&= \frac{c_\tau}{1 - \frac{1}{1+n_\tau/c_\tau - 1}} + c - c_\tau \\
&= \frac{n_\tau c_\tau}{n_\tau - c_\tau} + c - c_\tau
\end{aligned}$$

The fitted values of the frequency counts are given by

$$E_k = c_\tau P(X = k|X > 0)|_{\theta=\hat{\theta}} = \frac{c_\tau}{\hat{\theta}} \left(\frac{\hat{\theta}}{1 + \hat{\theta}} \right)^j,$$

for $j = 1, 2, .., \tau$. The standard error of $\hat{C}$ is

$$\hat{se}(\hat{C}) = \frac{\hat{C} - (c - c_\tau)}{\sqrt{n_\tau - c_\tau}}.$$

3.3 Two-component Geometric Model

3.4 Three-component Geometric Model

3.5 Weighted Linear Regression Model

3.6 Log-transformed Weighted Linear Regression Model

3.7 Kemp-type estimator

3.8 Good-Turing estimator

Good-Turing estimator assumes equal abundance of each species. The estimator of richness is given by

$$\hat{C} = \frac{c_-(\tau)}{1 - f_1/n_\tau} + c_+(\tau)$$

3.9 Chao1 estimator

Chao1 estimator is generally considered as a lower bound for C.

$$\hat{C} = \begin{cases} c + f_1^2/(2f_2), & \text{if } f_2 > 0 \\ c + f_1(f_1 - 1)/2, & \text{if } f_2 = 0 \end{cases}$$

3.10 Abundance-Based Coverage Estimator (ACE)

3.11 Abundance-Based Coverage Estimator 1 (ACE1)

3.12 Chao-Bunge estimator

3.13 Asymmetric Confidence Interval

Let $\hat{f}_0$ be the estimated number of unobserved species and $\hat{\sigma} = \hat{se}(\hat{f}_0) = \hat{se}(\hat{C})$. By Chao 1987, a 95% asymmetric confidence interval is given by

$$\left[c + \hat{f}_0/\eta, c + \hat{f}_0\eta\right],$$

where

$$\eta = \exp\{1.96[\log(1 + \hat{\sigma}^2/\hat{f}_0^2)]^{1/2}\}$$

4 Model selection

5 Proposal

1. Complete poisson model, geometric model, wlrm model and coverage based estimators, kemp-type estimators. Output include chi-sq and AIC.
2. Write function for chi-sq test

6 Future steps

1. Adjust standard error: After model selection, the standard error may no longer be valid. How can we adjust the standard error?
2. Is it possible to perform model selection with independently drawn samples from the same environment?
3. Include estimators from SpadeR package, breakaway, kemp-type estimators?
4. NPMLE methods unstable (discussed in Barger 2008)
5. model selection by varying τ?
6. In Barger 2008, negative binomial model θ moves to boundary of the parameter space. Some regularization methods to avoid boundary problem?
7. Is the finite mixture model biased?
8. Hypothesis testing for exponential mixture?
9. adjusted standard error or regularization method for kemp-type method

References