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 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.
The standard error is computed based on an asymptotic approximation:

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

where

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

3 Richness estimators

3.1 Poisson Model

3.2 Geometric Model

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

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