ESTIMATING THE NUMBER OF SPECIES WITH CATCHALL

JOHN BUNGE

Department of Statistical Science, 1198 Comstock Hall, Cornell University, Ithaca, NY 14853 , USA
E-mail: jab18@cornell.edu

In many situations we are faced with the need to estimate the number of classes in a population from observed count data: this arises not only in biology, where we are interested in the number of taxa such as species, but also in many other fields such as public health, criminal justice, software engineering, etc. This problem has a rich history in theoretical statistics, dating back at least to 1943, and many approaches have been proposed and studied. However, to date only one approach has been implemented in readily available software, namely a relatively simple nonparametric method which, while straightforward to program, is not flexible and can be prone to information loss. Here we present CatchAll, a new, platform-independent, user-friendly, computationally optimized software package which calculates a powerful and flexible suite of parametric models (based on current statistical research) in addition to all existing nonparametric procedures. We briefly describe the software and its mathematical underpinnings (which are treated in depth elsewhere), and we work through an applied example from microbial ecology in detail.

Keywords: species richness; finite mixture model; abundance.

1. Introduction

In many applied settings we encounter the need to estimate the number of classes in a population based on observed sample count data. For example, in biology it is common to collect a sample of organisms and sort them into taxa – we will use the term "species" for these taxa, recognizing that this may not be exact in some cases – count the number of representatives of each species in the sample, and from this data estimate the total number of species, both seen and unseen, in the underlying population. This is called the "species richness." There are many examples from other fields as well, such as veterinary medicine, where one may wish to estimate the number of farms with animals having a certain disease, or software engineering, where interest is in the number of potential types of errors in a complex software program.[1] Statisticians have been interested in this problem since the time of R. A. Fisher,[2] and many approaches have been studied theoretically and tested empirically, ranging from frequentist to Bayesian and from parametric to nonparametric.[3] However, to date only one class of statistical methods has been implemented in readily available software, namely the (frequentist) coverage-based nonparametric estimators of Chao and colleagues.[4] (This is not the only possible class of nonparametric estimators; see Section 3 below.) These are provided in, e.g., SPADE[5] and EstimateS,[6] and in some broader-use bioinformatics packages such as mothur[7] and QIIME;[8] see Section 2 for details.

The coverage-based nonparametric estimators are mathematically simple and computationally straightforward, and these estimators, known as Good-Turing, Chao1, the Abundance-Based Coverage Estimator ACE and its variants, and Chao-Bunge, can be accurate in some situations. (The associated standard errors are more complex computationally; see Section 2.2 below.) However, it is known that these estimators are typically downwardly-biased in

high-diversity situations[9] such as arise in modern high-throughput DNA sequencing studies, for instance. Furthermore, they are sensitive to inclusion/exclusion of outliers, i.e., species that appear with high abundance in the sample, so it is standard practice to truncate species abundance counts at 10 when using these estimators, that is, to ignore species with sample counts higher than 10, adding the number of such species to the estimate *ex post facto*. In addition, these estimators do not admit goodness-of-fit testing or other diagnostic assessments, and it is not clear how to graph or visualize the results.

In contrast, recent statistical research has elucidated a class of parametric *finite mixture models*[9] which are accurate in high-diversity populations (when the model is correct), are relatively insensitive to outliers, and permit a broad array of diagnostic and goodness-of-fit assessments, both quantitative and graphical. The basic idea is to "mix" several component parametric models together (i.e., to form a convex combination of them) so that one component fits the rare species and another the abundant ones (possibly using one or more additional components for improved fit to the sample count data). Estimators based on these models are not simple to compute, though, requiring multidimensional numerical search routines to obtain maximum likelihood estimates of the parameters (based on the expectation-maximization or EM algorithm), and model-selection procedures which are partly statistical and partly heuristic. In addition, computation of standard errors is quite involved, requiring numerical computation of inverse Fisher information matrices that can involve thousands of lines of code.

We originally explored the use of these models in biological applications by building a proof-of-concept system on a cluster in Cornell's Center for Advanced Computing using Maple,[10] but while functional the system was very slow, sometimes taking a week to complete an analysis. We analyzed several hundred datasets using this system (many from microbial ecology), and based on this experience we re-engineered our algorithms and rebuilt the system using a combination of C# and C. The result is CatchAll, a freely downloadable, user-friendly, platform-independent (Windows/Macintosh/Unix, single-processor/cluster, GUI/batch) software program which computes the full suite of finite-mixture models and all known nonparametric coverage-based estimates. CatchAll then compares all of these results; selects the best in each category and the "best-of-the-best"; and returns recommended estimates to the user along with associated standard errors, confidence intervals, and goodness-of-fit assessments. For the GUI version there is also an Excel-based module which produces publication-quality graphics displaying the the fit of the parametric models to the data, and the comparative performance of the various estimators. CatchAll usually computes a complete analysis in a minute or two on a single-processor machine.

Our purpose here is not to enter into the mathematical, statistical and computational details of CatchAll, which are discussed elsewhere[9],[11] but rather to describe a complete case study resulting in an estimate of total species richness in a particular setting. In Section 2 we discuss the data and its analysis, and Section 3 we draw conclusions and mention some future directions for expansion of CatchAll. In the Appendix we give a brief outline of our most important algorithm, which computes maximum likelihood estimates for the parametric models.

2. Analysis of a microbial diversity dataset

The International Census of Marine Microbes (ICoMM) is a large-scale research project on microbial diversity, intended "to (1) catalogue all known diversity of single-cell organisms inclusive of the Bacteria, Archaea, Protista and associated viruses, (2) to explore and discover unknown microbial diversity, and (3) to place that knowledge into appropriate ecological and evolutionary contexts."[12] Part of the ICoMM activity consists of taking samples of marine microbial organisms for (among other purposes) diversity evaluation. Essentially, a sample of water is taken and microbial 16S rRNA sequences are extracted. These sequences are then clustered into "operational taxonomic units" or OTUs; in our example below two sequences are assigned to the same OTU if they share 97% sequence identity, but the 97% value is conventional rather than theoretically based and can be varied at the discretion of the investigator. The OTU frequencies are then counted: some OTUs contain only one member sequence (the "singletons"), others two, others three, and so on. Finally we reorganize this information as "frequency count" data, consisting of the number of OTUs having one member or element; the number having two; the number having three, and so on. We note that each stage of this process, from sampling to sequence alignment and comparison to clustering, is nontrivial and subject to variation across labs and differing interpretation of results,[13] but for our purposes here we will assume that the frequency-count data is obtained in an unambiguous and closed-ended manner.

2.1. *Example dataset*

The sample data analyzed below was collected on January 7, 2005, as part of the ICoMM subproject "Application of the 454 technology to active-but-rare biosphere in the oceans: large-scale basin-wide comparison in the Pacific Ocean," by Koji Hamasaki and Akito Taniguchi of The University of Tokyo; for full details on this sub-project see the ICoMM Microbial Oceanographic Biogeographic Information System MICROBIS.[14] The complete frequency-count data is shown in Table 1. There were 19854 sequences grouped into 3018 OTUs, with (in particular) 2013 singleton OTUs; the maximally abundant sample OTU contained 1784 sequences. Figure 1 shows the data with the best fitted parametric curve (we explain this in Section 2.2). We retain the original scale (rather than, say, a log-log scale) to show the steep descent from the left, followed by the long slow decay to the right (in fact the plot is truncated at a maximum frequency of 254, while the actual data extends to 1784). This shape is typical of high-diversity data, which is often encountered in microbial diversity studies.

2.2. *CatchAll analysis of example data*

The basic idea of parametric species richness estimation is to fit a curve to the frequency count data and to project this curve upwards and to the left, to an abscissa of zero, so as to obtain an estimate of f_0. The estimate of the total number of species, unobserved + observed, is then $f_0 + f_1 + f_2 + ...$ The curve is a mixed-Poisson distribution based theoretically on an underlying species abundance distribution. It is fitted to the data via maximum likelihood, and the same procedure also yields standard errors, fitted values, goodness-of-fit statistics, etc.[9] CatchAll fits an ordered suite or family of five parametric curves: the (ordinary or unmixed)

Table 1. ICoMM frequency count dataset "ABR 0005 2005 01 07." i = frequency; f_i = # of sample OTUs with frequency i.

i	f_i	i	f_i	i	f_i	i	f_i
1	2013	23	1	56	1	165	1
2	416	25	3	57	1	173	1
3	173	27	2	59	1	191	1
4	85	28	1	71	1	195	1
5	63	29	3	73	1	201	1
6	43	30	1	76	1	202	1
7	39	31	3	80	1	208	1
8	18	32	1	84	1	223	1
9	24	33	1	85	1	225	1
10	8	34	1	93	1	233	1
11	17	35	1	94	1	254	1
12	8	36	1	114	1	319	2
13	6	38	2	119	1	328	1
14	3	40	1	122	1	548	1
15	4	42	1	123	2	560	1
16	6	43	1	131	1	675	1
17	9	46	1	148	1	1036	1
18	2	48	1	150	1	1361	1
20	6	53	2	154	1	1526	1
22	4	54	1	155	1	1784	1

Poisson, which unrealistically stipulates equal species abundances and is useful mainly as a lower-bound benchmark for the true richness; the (single) geometric; and mixtures of two, three, and four geometrics. (For technical reasons these are called mixtures of exponentials in the output and display.) The Poisson is mathematically the zero-order model in this scheme, followed by first- (single geometric), second- (mixture of two geometrics), third- and fourth-order mixture models. The idea, as noted above, is to mix (form a convex combination of) several component sub-models, one component fitting the steep decline of the frequency count data on the left, another fitting the shallow decline on the right, and possibly others fitting intermediate parts of the data.

At this point a second issue arises. Any parametric curve is defined by a finite number of parameters (1, 1, 3, 5 and 7 in our models of order 0, 1, 2, 3, and 4, respectively), and consequently has finite flexibility. In most datasets it is not possible for any parametric curve to fit the entire extent of the data from f_1 to $f_{\max}$ (where $f_{\max}$ is the number of species, i.e., the frequency count, at the largest sample frequency). It is therefore standard practice to truncate the data on the right at some frequency, which we call τ; the statistical analysis is then based on $f_1, f_2, \ldots, f_\tau$, and the number of species with frequencies greater than tau (i.e., $f_{\tau+1} + f_{\tau+2} + \ldots + f_{\max}$) is added to the estimate *ex post facto*. As noted above, in the coverage-based nonparametric methods τ is fixed at 10 (we return to this issue below). The parametric methods are more flexible, and we wish to base the statistical analysis on as much of the frequency count data as possible, that is, to use the largest possible τ for which we can still obtain a good fit of the model. CatchAll therefore fits every model at every value of

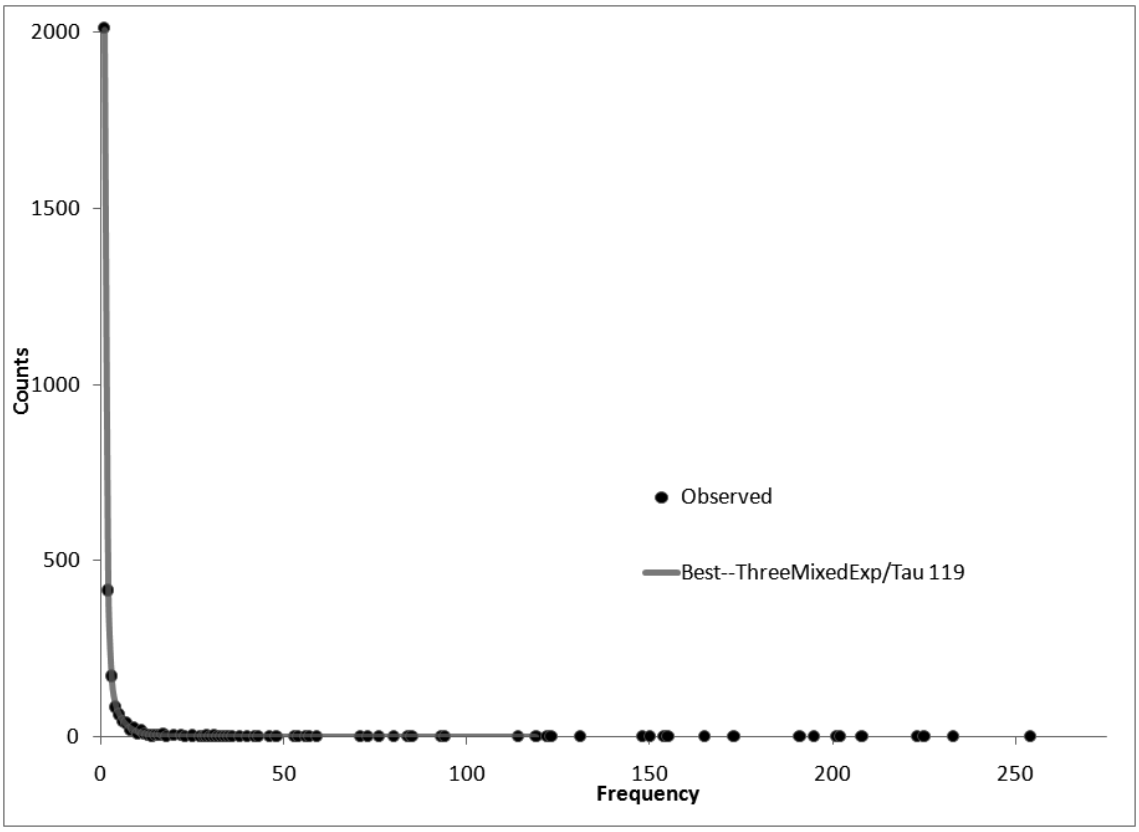

Fig. 1. Frequency count distribution of sample ICoMM data, with best fitted parametric curve.

τ and compares the results. Essentially we select the best-fitting parametric model at each fixed τ using the small-sample-size-adjusted Akaike Information Criterion, AICc, and then select across τ's using the p-values from the following two Pearson χ^2 statistics. "GOF0" is the p-value of a Pearson χ^2 goodness-of-fit test based on the "raw" or unadjusted frequency counts, and "GOF5" is the p-value of the χ^2 test after concatenating adjacent frequencies to obtain a minimum cell count of 5. We use both because the p-value of the χ^2 test is based on an asymptotic (large-sample) approximation to the distribution of the test statistic, and the aforementioned concatenation of adjacent cells is standard practice to obtain sufficiently large cell counts in sparse tables such as we typically have in our data (Table 1). Thus GOF0 should be regarded as a diagnostic "divergence statistic" signalling divergence from the null hypothesis (which states that the model is correct), while GOF5 represents the p-value from a legitimate statistical hypothesis test of model fit. In either case smaller p-values represent evidence against the fit of the model, and larger p-values represent evidence in favor.

The final model-selection algorithm, in outline form, is as follows.

Model selection algorithm

(1) (Statistical.) Eliminate model*τ combinations for which GOF5 < 0.01.
(2) (Statistical.) For each τ, select the model with minimum AICc (Akaike Information Criterion, corrected when necessary for small sample sizes).
(3) (Heuristic.) Eliminate model*τ combinations for which SE $>$ estimate/2.

Table 2. CatchAll analysis summary for ICoMM dataset. "Selection" = status of model, "Model" = order of parametric model or designation of nonparametric method, τ = upper frequency cutoff, "Est." = estimated total species richness, "SE" = standard error of estimate, "LCB" = lower 95% confidence bound, "UCB" = upper 95% confidence bound, "GOF0" = unadjusted χ^2 p-value, "GOF5" = adjusted χ^2 p-value.

Selection	Model	τ	Est.	SE	LCB	UCB	GOF0	GOF5
Best	3	119	15369	1322	13037	18243	0.0102	0.3199
2a	3	71	16032	1615	13231	19600	0.0704	0.0932
2b	4	254	16245	1833	13111	20352	0.0004	0.0785
2c	2	13	16604	1607	13802	20134	0.0004	0.0105
NP1	Chao1	2	7888	330	7283	8580		
NP2	ACE1	10	13519	777	12104	15156		
Parm τmax	3	1784	13476	810	12005	15188		0.0000
NP τmax	ACE1	1784	12227106	4012967	6532741	22887348		

(4) (Heuristic.) Then:

- Best model: Select the largest τ for which GOF0 ≥ 0.01.
- Model 2a: Select the τ with maximum GOF0.
- Model 2b: Select the largest τ.
- Model 2c: Select τ as close as possible but ≤ 10.

(5) (Heuristic.) If all model*τ combinations are eliminated, relax the restrictions in (3) and (4) and iterate.

We then report the "best-of-the-best" parametric model, along with three competing models 2a–2c, which are unordered in terms of preference.

The results of the ICoMM data analysis are shown in Table 2. The best analysis overall is given in the first row of the table. The fitted model is a mixture of three geometric components, one fitting the steep decline of the data on the left, one fitting the middle, and one fitting the shallow decline to the right. This is the curve shown in Figure 1, although the three components are of course not visible separately. The estimated total number of species is 15369 (i.e., the estimate of f_0 is 12351 so that $15369 = 12351 + 3018$). The standard error associated with the estimate of species richness is 1322. We do not form the "Wald" or Gaussian 95% confidence interval consisting of the estimate $\pm 1.96 * \text{SE}$; rather, we use an asymmetric confidence interval based on a lognormal approximation due to Chao,[15] which is more realistic in this context. (In the parametric modeling setting the Chao interval is an approximation to the profile likelihood confidence interval, which though optimal is more complicated computationally and will appear in a later version of CatchAll.) The last two columns display the goodness-of-fit statistics GOF0 and GOF5. Both p-values exceed 0.01, indicating good fit of the model to the data.

Note that $\tau = 119$ for the best selected model, which, while still some distance from the maximum frequency of 1784, represents the use of the first 53 of the 80 frequencies existing in the data (66%). Thus the fitted curve in Figure 1 extends through 119 on the horizontal axis. The competing models 2a, 2b, and 2c (these are unordered in terms of desirability) represent various good but suboptimal compromises vis-à-vis goodness-of-fit, large τ, and other factors.

Model 2a again has 3 components and better GOF0 (than the best model), but smaller τ; 2b has four components and higher τ but GOF0$<$ 0.01, and 2c has low τ and low GOF0. Nevertheless their estimates, SEs and confidence intervals do not differ too much from those of the best model.

CatchAll also computes all known coverage-based nonparametric richness estimates, including that of Chao and Bunge.[4] The best of these are reported in the results table. Table 2 first shows "NP1" (meaning the first reported, not best selected) nonparametric analysis which is the "Chao1" statistic, a simple lower bound estimator with $\tau = 2$. This is useful as a cross-check or benchmark. Next, as "NP2" (the second reported nonparametric analysis), we report either ACE or its high-diversity variant ACE1, selected according to a criterion based on the coefficient of variation of the frequency count data.[4] These both have τ fixed at 10 (as noted above), as recommended in the original statistical research. In this connection we note that we have also re-engineered the standard error computation algorithms for the coverage-based nonparametric methods, yielding improved precision relative to existing software for these methods. For the ICoMM example data we see that ACE1 returns both an estimate and an SE that are lower than those of the parametric models, in accordance with the downward bias in high-diversity situations mentioned above.

We also report the best parametric and the best nonparametric analysis with τ fixed at the maximum frequency in the data, i.e., using the entire frequency count dataset. Table 2 shows that, while the parametric analysis at maximum τ differs little from the best selected parametric analysis (reflecting the relative insensitivity to outliers referred to above), the coverage-based nonparametric analysis "drifts off to infinity" along with its SE, when the large outlying frequencies are included.

To display the behavior of the various models and estimators as the larger frequencies are added to the data, the GUI version of CatchAll provides a bubble plot, shown in Figure 2. The figure displays the increase of the nonparametric estimates and their SEs as a function of increasing τ, compared to the relatively stable behavior of two of the parametric estimates (3rd- and 4th-order mixtures) as functions of τ. The bubble areas are proportional to 1/2 the associated standard error at each point. (Figure 2 has been reduced for simplicity to show only part of the τ-range and only two of the five parametric models.) It is clear that, while the estimators agree reasonably well at $\tau = 10$ (as was seen in Table 2), the nonparametric coverage-based estimates (and their error terms) diverge to infinity as τ increases, whereas the parametric estimates decrease slightly. Thus the coverage-based nonparametric methods can produce non-overlapping, hence contradictory, confidence intervals from the same dataset, depending on which large "outlying" frequencies are included in the analysis. The cause of this behavior has not yet been mathematically ascertained (although it is universally observed), and is a topic for further theoretical research.

The final selected analysis, referring again to the first row of Table 2, is the 3rd-order model at $\tau = 119$, and is indicated by an arrow in Figure 2.

Several existing and widely-used programs also compute (some of) the coverage-based nonparametric estimates. The Chao-Bunge estimator[4] is produced by SPADE;[5] ACE/ACE1 are produced by SPADE, EstimateS[6] and mothur;[7] and Chao1 is produced by SPADE, Es-

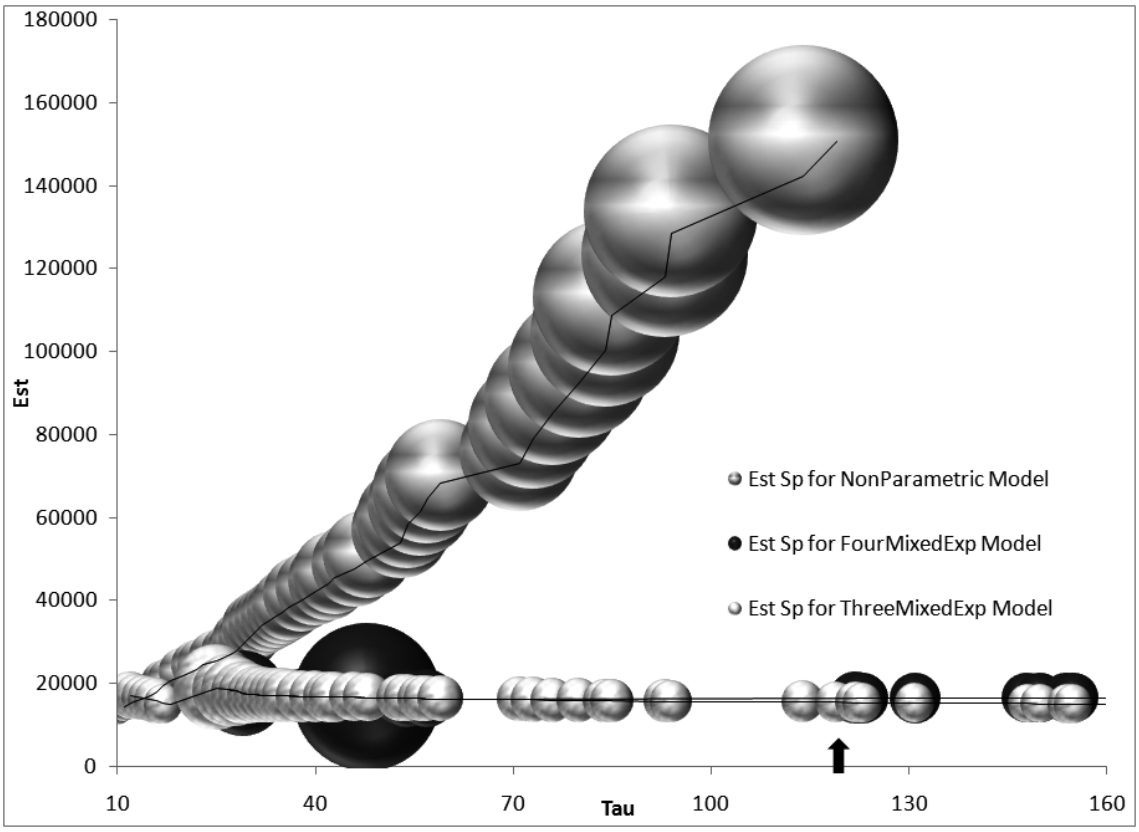

Fig. 2. Total species richness estimates as a function of τ in ICoMM data. "Est Sp for NonParametric Model" = total richness estimated using ACE or ACE1, "Est Sp for FourMixedExp Model" = total richness estimated using 4th-order parametric mixture model, "Est Sp for ThreeMixedExp Model" = total richness estimated using 3rd-order parametric mixture model.

timateS, mothur and QIIME[8] (accompanied in all cases by standard errors and confidence bounds). Thus the user will essentially find replicates of these nonparametric estimates in the cited programs and in CatchAll. There are two notable differences, however. First, CatchAll is unique in computing these nonparametric estimates at *every* value of τ, so as to reveal their behavior as more frequency counts are included in the data. Second, the standard error and confidence interval computations in CatchAll are based on new algorithmic representations of the underlying mathematics, and are considerably more accurate and precise than the usual algorithms used for this purpose.

3. Conclusions and future directions

We have presented CatchAll, a software program for parametric and nonparametric statistical estimation of total species richness, along with visualization and comparison of competing analyses. CatchAll is user-friendly — it requires no input other than the input data file specification from the user, that is, no options need to be set; it is freely downloadable, from http://www.northeastern.edu/catchall/; it is platform-independent and will run under Windows or the Macintosh operating system or Unix, on single- or multiple-processor machines, in GUI or in batch mode; and it is fast, completing most analyses in a minute or two on a

modestly-specified machine. It is the first program to implement parametric species richness modeling in a comprehensive, integrated and accessible fashion, and it also computes all existing coverage-based nonparametric estimates (with improved standard errors). A full manual is also provided.

In terms of future developments, we will next extend CatchAll to include a completely novel species richness estimation method based on fitting a linear model to ratios of adjacent frequency counts.[1] We will then incorporate objective Bayesian methods using reference and Jeffreys priors for the number of species.[16] Finally we will implement nonparametric maximum likelihood estimation, another new approach in the species richness problem.[9] These are computationally intensive procedures which will take some time to program. We welcome comments and suggestions regarding the current or potential future versions of CatchAll, which may be addressed to the author.

Acknowledgments

We deeply appreciate the advice and assistance of the following collaborators. Linda Woodard is the principal CatchAll programmer and Consultant in Cornell University's Center for Advanced Computing. Sean Connolly programmed the Excel display module and is a 2010 graduate of Cornell. Linda Amaral Zettler of the Marine Biological Laboratory is Secretariat of ICoMM. Slava Epstein of Northeastern University is co-investigator on the CatchAll grant. This research was conducted using the resources of the Cornell University Center for Advanced Computing, which receives funding from Cornell University, New York State, the National Science Foundation, and other leading public agencies, foundations, and corporations. The development of CatchAll is funded by NSF grant DEB-0816638 to JB. We thank James Foster of the University of Idaho for proposing the submission of this paper.

Appendix A. Computing the maximum likelihood estimates

We outline the expectation-maximization (EM) algorithm for computing the maximum likelihood estimates of the parameters in the mixture-of-two-exponentials (-geometrics) model, when the frequency count data is truncated on the right at τ. Extending the algorithm to higher numbers of components (three and four) is straightforward though not simple.

The observed data consists of the (nonzero) frequency counts $f_1, f_2, \ldots$, where f_i denotes the number of species observed i times in the sample. The relevant part of the log-likelihood of the data under the model is[9]

$$\sum_{i=1}^{\tau} f_i \log\left(u\left(\frac{1}{t_1}\right)\left(\frac{t_1}{1+t_1}\right)^i + (1-u)\left(\frac{1}{t_2}\right)\left(\frac{t_2}{1+t_2}\right)^i \right), \tag{A.1}$$

where $t_1, t_2 > 0, u \in (0,1)$. Our objective is to find (t_1, t_2, u) to maximize (A.1) given $f_1, f_2, \ldots$. We initialize u as $u^{(0)} = 1/2$, and t_1, t_2 as

$$t_1^{(0)} = \frac{\sum_{i=1}^{\lfloor 2\tau/3 \rfloor} i f_i}{\sum_{i=1}^{\lfloor 2\tau/3 \rfloor} f_i} - 1, \qquad t_2^{(0)} = \frac{\sum_{i=\lfloor \tau/3 \rfloor + 1}^{\tau} i f_i}{\sum_{i=\lfloor \tau/3 \rfloor + 1}^{\tau} f_i} - 1.$$

Now suppose we are at the kth step, $k = 0, 1, \ldots$, so that we have values $t_1^{(k)}, t_2^{(k)}, u^{(k)}$. Define

$$z_i^{(k)} := \frac{u^{(k)} \left(\frac{1}{t_1^{(k)}}\right) \left(\frac{t_1^{(k)}}{1+t_1^{(k)}}\right)^i}{u^{(k)} \left(\frac{1}{t_1^{(k)}}\right) \left(\frac{t_1^{(k)}}{1+t_1^{(k)}}\right)^i + (1-u^{(k)}) \left(\frac{1}{t_2^{(k)}}\right) \left(\frac{t_2^{(k)}}{1+t_2^{(k)}}\right)^i},$$

$i = 1, 2, \ldots$.

Update u:

$$u^{(k+1)} = \frac{\sum_{i=1}^{\tau} f_i z_i^{(k)}}{\sum_{i=1}^{\tau} f_i}$$

Update t_1, t_2:

$$t_1^{(k+1)} = \frac{\sum_{i=1}^{\tau} f_i i z_i^{(k)}}{\sum_{i=1}^{\tau} f_i z_i^{(k)}} - 1;$$

$$t_2^{(k+1)} = \frac{\sum_{i=1}^{\tau} f_i i (1 - z_i^{(k)})}{\sum_{i=1}^{\tau} f_i (1 - z_i^{(k)})} - 1.$$

Update z: $z_i^{(k)} \to z_i^{(k+1)}$.

Iterate to convergence. This yields MLEs (t_1, t_2, u), which are the key quantities required for all estimates, standard errors, fitted values, and goodness-of-fit statistics.[9]

References

1. I. Rocchetti *et al.*, forthcoming in *Annals of Applied Statistics* (2010).
2. R. A. Fisher *et al.*, *Journal of Animal Ecology* **12**, 44 (1943).
3. J. Bunge and M. Fitzpatrick, *Journal of the American Statistical Association* **88**, 364 (1993).
4. A. Chao and J. Bunge, *Biometrics* **58**, 531 (2002).
5. T. J. Shen *et al.*, *Ecology* **84**, 798 (2003).
6. R. Colwell, http://purl.oclc.org/estimates.
7. P. Schloss, *Applied and Environmental Microbiology* **75**, 7537 (2009).
8. J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, *Nature Methods* **7**, 335 (2010).
9. J. Bunge and K. Barger, *Biometrical Journal* **50**, 971 (2008).
10. http://www.maplesoft.com/ .
11. J. Bunge *et al.*, in preparation (2010).
12. http://icomm.mbl.edu/
13. S. M. Huse, D. M. Welch, H. G. Morrison, M. L. Sogin, *Environmental Microbiology* **12**, 1889 (2010).
14. http://icomm.mbl.edu/microbis/
15. A. Chao, *Biometrics* **43**, 783 (1987).
16. K. Barger and J. Bunge, forthcoming in *Journal of Bayesian Analysis* (2010).