BIOMETRICS 000, 1–??
000 2015
DOI: 000

Estimating Diversity via Frequency Ratios

Amy Willis*

Department of Statistical Science, Cornell University, Ithaca, New York, U.S.A.

**email:* adw96@cornell.edu

and

John Bunge

Department of Statistical Science, Cornell University, Ithaca, New York, U.S.A.

SUMMARY: We wish to estimate the total number of classes in a population based on sample counts, especially in the presence of high latent diversity. Drawing on probability theory that characterizes distributions on the integers by ratios of consecutive probabilities, we construct a nonlinear regression model for the ratios of consecutive frequency counts. This allows us to predict the unobserved count and hence estimate the total diversity. We believe that this is the first approach to depart from the classical mixed Poisson model in this problem. Our method is geometrically intuitive and yields good fits to data with reasonable standard errors. It is especially well-suited to analyzing high diversity datasets derived from next-generation sequencing in microbial ecology. We demonstrate the method's performance in this context and via simulation, and we present a dataset for which our method outperforms all competitors.

KEY WORDS: Alpha diversity; Biodiversity; Capture-recapture; Characterization of distributions; Microbial ecology; Species richness

1. Introduction

Our goal is to estimate the total number of classes C in a population, based on a sample of individuals from the population. This problem has many applications in the natural sciences, as well as in linguistics and computer science, but our particular interest is in microbial ecology: estimating the biodiversity (number of taxa or species richness) in a microbial community from a sample of DNA or RNA sequences. The rapid development of next generation sequencing technology has provided the opportunity to analyze very large microbial community composition datasets, often containing more than 10^9 sequences. Species richness analysis is a tool that frequently provides an indication of ecosystem health (Dethlefsen et al., 2008; Gao et al., 2013). However, classical approaches to the "species problem" perform poorly in the microbial context, because these datasets differ structurally from animal abundance datasets. Microbial datasets are characterized by a large number of rarely observed species, including a high "singleton" count (species observed exactly once), as well as a small number of very abundant species. The resulting large peak to the left and long tail to the right make inference especially challenging (Lladser et al., 2011; Bunge et al., 2014).

Let $f_1, f_2, \ldots$ denote the numbers of taxa observed once, twice, and so on, in the sample, and let f_0 denote the number of unobserved taxa, so that $C = f_0 + f_1 + f_2 + \ldots$. Using the observed data we wish to predict f_0 and hence estimate C. Our approach involves analysis of the *frequency ratios* f_{j+1}/f_j, as a function of j. We fit a nonlinear regression model to the ratios, projecting the fitted function downward to $j = 0$ so as to predict f_0 and estimate C, with associated standard errors and goodness-of-fit assessments.

The idea of using ratios of frequencies, or ratios of the probabilities of a mass function on the nonnegative integers, dates back at least to Katz (1945), who proved that for a probability distribution on the nonnegative integers with p_j mass on $j, j = 0, 1, \ldots$, the function $(j+1)p_{j+1}/p_j$ is linear in j only for the binomial, Poisson or negative

binomial distributions. The introduction of the ratio plot (f_{j+1}/f_j vs. j) in the species problem is due to Rocchetti et al. (2011), who exploited the Katz structure but found that a log-transformation was needed to fit the underlying linear model to data. Here we generalize that work by fitting a heteroscedastic, correlated nonlinear regression model to $(j, f_{j+1}/f_j)$, based on a theory of probability ratios due to Kemp (1968). This gives rise to a rich class of models which generate plausible estimates of C, as well as standard errors and a model selection procedure. Our method is geometrically intuitive, general (with respect to the underpinning probability model), stable (especially compared to maximum likelihood), and capable of identifying "classical" marginal distributions.

The traditional approach to the species problem works directly with the frequency counts. In this setup the sample counts of the taxa are modeled as C independent Poisson random variables where the Poisson means are an i.i.d. sample from some mixing distribution. The frequencies $(f_1, f_2, \ldots)$ then constitute a random sample from a zero-truncated mixed Poisson distribution, and estimation of C is based the likelihood function. This model has been studied from the frequentist, Bayesian, parametric, and nonparametric points of view (e.g., Böhning and Kuhnert (2006); Mao and Lindsay (2007); Bunge et al. (2012, 2014)), and improving the stability of mixed Poisson models motivated the ratio-based approach of Rocchetti et al. (2011) (see also Böhning et al. (2013, 2014)). In this paper we break away from the assumptions of the mixed Poisson model. We believe that this departure is as yet unexplored in the literature. This approach achieves greater flexibility in modeling, is underpinned by fewer assumptions, and permits simple diagnostics for model misspecification. Our method is formal rather than exploratory (*viz.*, we wish to obtain a richness estimate and a standard error), so we do not contrast it to heuristic approaches such as the "count metameter" of Hoaglin et al. (1985).

We are interested, then, in models for f_{j+1}/f_j, or equivalently p_{j+1}/p_j, as a function of j. There are three main lines of research on this topic. First, Katz' result as extended

by Kemp (1968), which is our primary model; we discuss it in detail below. Second, the Lerch distribution (Zörnig and Altmann, 1995; Johnson et al., 2005), which may be characterized by probability ratios of the form

$$\frac{p_{j+1}}{p_j} = \alpha \left(1 - \frac{1}{\beta + j + 1}\right)^{\nu},$$

$0 < \alpha < 1, \beta > 0, \nu \neq 0$. This arises as the stationary distribution of a birth and death process. Our investigations found that it does not generate a flexible class of statistical procedures and is difficult to fit numerically, and appears to present no advantage over the ratio-of-polynomials models. The third line of research on ratio characterization arises through distributions of randomly stopped sums (Pestana and Velosa, 2004), yielding probability ratios

$$\frac{p_{j+1}}{p_j} = \alpha + \frac{\beta}{\sum_{i=0}^{j} \gamma^i},$$

$\alpha, \beta \in \mathbb{R}, -1 < \gamma < 1$. We have not found any relevant interpretation or advantage for these models in this problem and so do not pursue them here.

In summary, we define a nonlinear model, based on ratios of polynomials, for the ratios of counts. This entails a generalized probability model for the count data which need not be mixed Poisson. We fit the models by nonlinear regression rather than maximum likelihood, for reasons we explain below. While the implementation is nontrivial due to the heteroscedastic and autocorrelated nature of the ratio data, the result is a flexible procedure which allows estimation of the total number of classes when sample sizes are large (i.e. $n \gtrsim 500$), as is typical of high-throughput sequencing datasets. In the following sections we discuss the statistical approach, and describe an R package which implements the method. We analyze several datasets and present simulation results.

2. Distributions based on ratios of probabilities

Let $\mathbf{p}$ denote a probability distribution on $\{0, 1, 2, \ldots\}$ with mass p_j on $j = 0, 1, 2, \ldots$, and $\sum_j p_j = 1$. A rich literature has examined characterization of distributions via ratios of adjacent probabilities $p_{j+1}/p_j, j = 0, 1, 2, \ldots$. The first broad theory was provided by Kemp (1968), who analyzed distributions with probability generating functions of the form

$$G(s) = \sum_{j=0}^{\infty} s^j p_j = \frac{{}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; \lambda s)}{{}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; \lambda)}, \tag{1}$$

where F is the generalized hypergeometric function and $a_1, \ldots, a_p, b_1, \ldots, b_q$ and λ are parameters. (As usual, $p_j = G^{(j)}(0)/j!$.) Analyzing the relevant parameter spaces was one of the main points of Kemp's original paper, and we do not reproduce her results here (see also Dacey (1972); Kemp (2010)). She demonstrated that for these distributions, ratios of probabilities have the form

$$\frac{p_{j+1}}{p_j} = \frac{(a_1 + j) \cdots (a_p + j)\lambda}{(b_1 + j) \cdots (b_q + j)(j + 1)}, \tag{2}$$

that is, rational functions of j. Tripathi and Gurland (1977) discuss $(p, q, a_1) = (2, 1, 1)$.

The results of Katz (1945) demonstrate that at least some distributions in the class defined by (1), which we call Kemp-type, are also mixed Poisson. Because of the prevalence of mixed Poisson distributions in the species problem literature, it is natural to ask whether this is true of all Kemp-type distributions. The answer is negative: for instance, terminating distributions can arise under the framework of Kemp (1968, Case (c)), while mixed Poisson distributions necessarily have full support. Furthermore, even if we restrict to Kemp distributions with full support on the nonnegative integers, it may be readily shown using the results of Puri and Goldie (1979) that these distributions need not be mixed Poisson. We conclude that the Kemp family provides an interesting direction of departure from mixed Poisson distributions. We will see in Sections 5 and 6 the advantages of this generalization when analyzing microbial data.

3. Frequency count ratio statistics

We begin by studying the joint distribution of the ratios $f_{j+1}/f_j, j = 0, 1, \ldots$. There are $C < \infty$ classes in the population. Assume that the ith class contributes X_i members to the sample, $X_i = 0, 1, \ldots$, and that $X_1, \ldots, X_C \sim$ i.i.d. $\mathbf{p}$, where $\mathbf{p}$ may have unbounded support. Then $f_j = \#\{X_i = j\}$, $j = 0, 1, \ldots$, and define the number of observed species $c = \sum_{j \geqslant 1} f_j$ and the number of observed individuals $n = \sum_{j \geqslant 1} j f_j$; f_0 is unobserved. We seek an approximation to the mean and covariance of $\{f_{j+1}/f_j\}_{j \geqslant 1}$.

Note first that the joint distribution of $f_0, f_1, \ldots$ is multinomial with (in general) unbounded support, with probability mass function (p.m.f.)

$$\frac{C!}{\prod_{j \geqslant 0} f_j!} \prod_{j \geqslant 0} p_j^{f_j}.$$

In a practical situation we can set the maximum frequency to be some fixed value τ. We therefore choose τ so that $\sum_{j > \tau} p_j$ is small, and we replace $\mathbf{p}$ by $\mathbf{q} \approx \mathbf{p}$ where $q_j = p_j/(1 - \sum_{j > \tau} p_j), j = 0, \ldots, \tau$ and $q_j = 0, j > \tau$. The p.m.f. of $f_0, f_1, \ldots, f_\tau$ then becomes an ordinary multinomial with bounded support,

$$\frac{C!}{\prod_{j=0}^{\tau} f_j!} \prod_{j=0}^{\tau} q_j^{f_j}.$$

It is not obvious how to analyze the moments of $\{f_{j+1}/f_j\}_{1 \leqslant j \leqslant \tau - 1}$ directly, but a Poisson approximation is available. One version is due to McDonald (1980), who compared the vector $\mathbf{f} := (f_1, \ldots, f_\tau)$ with $\mathbf{V} := (V_1, \ldots, V_\tau)$, where $V_1, \ldots, V_\tau$ are independent Poisson random variables with $E(V_i) = Cq_i$, and showed that

$$\sum_{\mathbf{a} \in \mathbb{R}^\tau} |\Pr(\mathbf{f} = \mathbf{a}) - \Pr(\mathbf{V} = \mathbf{a})| \leqslant 2C(1 - q_0)^2.$$

See Roos (1999a,b) for more complex but tighter bounds. Reapplying the approximation $\mathbf{q} \approx \mathbf{p}$, we proceed by treating $f_1, \ldots, f_\tau$ as independent Poisson random variables with $E(f_i) = Cp_i$.

We are interested, then, in the first and second (joint) moments of $\{f_{j+1}/f_j\}_{1 \leqslant j \leqslant \tau - 1}$. However, with positive probability any f_j may be zero. We therefore condition on all

f_j being nonzero, for $j = 1$ up to some maximum $J + 1 \leqslant \tau$. To avoid introducing new notation we now let f_j denote this zero-truncated version, so that finally we model $\{f_1, f_2, \ldots, f_{J+1}\}$ as independent zero-truncated Poisson random variables with Poisson parameter Cp_j.

Let μ_j and σ_j^2 denote the mean and variance of f_j, respectively.The delta method gives

$$E\left(\frac{f_{j+1}}{f_j}\right) \approx \frac{\mu_{j+1}}{\mu_j},$$

$j = 1, \ldots, J$, and

$$\begin{aligned}
&\mathrm{Cov}\left(\frac{f_2}{f_1}, \ldots, \frac{f_{J+1}}{f_J}\right) \\
&\approx \begin{bmatrix}
\frac{\mu_2^2\sigma_1^2}{\mu_1^4} + \frac{\sigma_2^2}{\mu_1^2} & & & & \\
-\frac{\mu_3\sigma_2^2}{\mu_2^2\mu_1} & \frac{\mu_3^2\sigma_2^2}{\mu_2^4} + \frac{\sigma_3^2}{\mu_2^2} & & & \\
0 & -\frac{\mu_4\sigma_3^2}{\mu_3^2\mu_2} & \frac{\mu_4^2\sigma_3^2}{\mu_3^4} + \frac{\sigma_4^2}{\mu_3^2} & & \\
\vdots & & \ddots & & \\
0 & \cdots & -\frac{\mu_J\sigma_{J-1}^2}{\mu_{J-1}^2\mu_{J-2}} & \frac{\mu_J^2\sigma_{J-1}^2}{\mu_{J-1}^4} + \frac{\sigma_J^2}{\mu_{J-1}^2} & \\
0 & \cdots & 0 & -\frac{\mu_{J+1}\sigma_J^2}{\mu_J^2\mu_{J-1}} & \frac{\mu_{J+1}^2\sigma_J^2}{\mu_J^4} + \frac{\sigma_{J+1}^2}{\mu_J^2}
\end{bmatrix}. \qquad (3)
\end{aligned}$$

The mean and variance of a zero-truncated Poisson random variable with (original) parameter λ are $\lambda/(1 - e^{-\lambda})$ and $\lambda(1 - e^{-\lambda} - \lambda e^{-\lambda})/(1 - e^{-\lambda})^2$ respectively. Therefore $\mu_j \approx Cp_j/(1 - e^{-Cp_j})$ so that

$$\frac{\mu_{j+1}}{\mu_j} \approx \frac{p_{j+1}}{p_j}\frac{1 - e^{-Cp_j}}{1 - e^{-Cp_{j+1}}},$$

and since C is typically large we regard f_{j+1}/f_j as a reasonable estimate of p_{j+1}/p_j. The expressions in the covariance matrix are more complicated. We have

$$\mathrm{Cov}\left(\frac{f_j}{f_{j-1}}, \frac{f_{j+1}}{f_j}\right) \approx -\frac{1}{C}\frac{p_{j+1}}{p_{j-1}p_j}\frac{1 - e^{-Cp_{j-1}}}{1 - e^{-Cp_{j+1}}}(1 - e^{-Cp_j} - Cp_j e^{-Cp_j}), \qquad (4)$$

$j = 2, \ldots, J$ and

$$\mathrm{Var}\left(\frac{f_{j+1}}{f_j}\right) \approx \frac{1}{C}\Big[\frac{p_{j+1}^2}{p_j^3}\frac{(1 - e^{-Cp_j})^2}{(1 - e^{-Cp_{j+1}})^2}\left\{1 - e^{-Cp_j} - Cp_j e^{-Cp_j}\right\}$$

$$+\frac{p_{j+1}}{p_j^2}\frac{(1-e^{-Cp_j})^2}{(1-e^{-Cp_{j+1}})^2}\left\{1-e^{Cp_{j+1}}-Cp_{j+1}e^{-Cp_{j+1}}\right\}\Big], \tag{5}$$

$j = 1, \ldots, J$. We will return to these when considering weights for nonlinear regression.

4. Nonlinear regression

Our initial model is based on (2), which we write as a ratio of polynomials in j. We use nonlinear regression, since no explicit likelihood is available in general, and estimation via empirical probability generating functions also presents difficulties in this context (Ng et al., 2013). Expanding and simplifying (2) then gives

$$\frac{f_{j+1}}{f_j} = \frac{\beta_0^* + \beta_1^* j + \ldots + \beta_p^* j^p}{1 + \alpha_1^* j + \ldots + \alpha_q^* j^q} + \epsilon_j. \tag{6}$$

where $\beta_0^*, \ldots, \beta_p^*, \alpha_1^*, \ldots, \alpha_q^*$ represents a more computationally convenient reparametrization of $a_1, \ldots, a_p, b_1, \ldots, b_q, \lambda$. In order to reduce the correlation between the parameter estimates we center j at $\bar{j}$, the (empirical) mean of j. Our final model is therefore

$$\frac{f_{j+1}}{f_j} = \frac{\beta_0 + \beta_1(j-\bar{j}) + \ldots + \beta_p(j-\bar{j})^p}{1 + \alpha_1(j-\bar{j}) + \ldots + \alpha_q(j-\bar{j})^q} + \epsilon_j, \tag{7}$$

where we assume that Cov$([\epsilon_j])$ is given by (3) – (5). We estimate $(\beta_0, \ldots, \beta_p; \alpha_1, \ldots, \alpha_q)$ by $(\hat{\beta}_0, \ldots, \hat{\beta}_p; \hat{\alpha}_1, \ldots, \hat{\alpha}_q)$ using a preimplemented nonlinear least squares solver (R Core Team, 2013).

Assume for the moment that we have selected a pair (p, q). The parameter estimation problem is then

$$\text{argmin}_{\beta_0,\ldots,\beta_p,\alpha_1,\ldots,\alpha_q}(\mathbf{F}-\mathbf{P})'\mathbf{W}^{-1}(\mathbf{P})(\mathbf{F}-\mathbf{P}), \tag{8}$$

(Seber and Wild, 1989, p.27), where

$$\mathbf{F} = \left[\frac{f_{j+1}}{f_j}\right], \mathbf{P} = \left[\frac{\beta_0 + \beta_1(j-\bar{j}) + \ldots + \beta_p(j-\bar{j})^p}{1 + \alpha_1(j-\bar{j}) + \ldots + \alpha_q(j-\bar{j})^q}\right],$$

$j = 1, \ldots, J$, and $\mathbf{W}(\mathbf{P})$ is the tridiagonal covariance matrix with diagonals given in (5) and off-diagonals in (4). We write $\mathbf{W}(\mathbf{P})$ to reflect the fact that the covariance structure depends on the true probabilities, which we approximate with the fitted values of

$\{p_1, \ldots, p_{J+1}\}$. Assuming a correctly specified model and certain integrability conditions on the errors, the conditions of White and Domowitz (1984) can be verified to demonstrate asymptotic consistency and normality of the solution to (8). We find that numerical convergence is almost never achieved when $\mathbf{W}(\mathbf{P})$ is tridiagonal, and so henceforth we approximate $\mathbf{W}(\mathbf{P})$ by its diagonal. Concurring with Rocchetti et al. (2011), simulations show this results in only a slight loss of precision.

The next question concerns the initial weighting scheme. Prior to model selection we do not have a form for (even the diagonal of) $\mathbf{W}(\mathbf{P})$. We considered various smooth functions for this including j^γ and $e^{\gamma j}$, and after much testing concluded that initial weights $1/j$ work well while remaining independent of model selection.

To find starting values for the parameters $(\beta_1, \ldots, \beta_p; \alpha_1, \ldots, \alpha_q)$ we use a sequential procedure, since under parametrization (7) all models are nested. The parameters of the lowest order model $q = 0, p = 1$ can be estimated using ordinary least squares. The starting values for model $p = q = 1$ are then $\hat{\beta}_{0,OLS}$, $\hat{\beta}_{1,OLS}$ and $\alpha_1 = 0$. This procedure is then repeated, using each model's estimates as the initial values for the next highest order model. If a model does not converge, the relevant parameter estimates are set to zero.

For $j = 0$, (7) gives

$$E\left(\frac{f_1}{f_0}\right) = \frac{\sum_{r=0}^{p} \beta_r (-1)^r \bar{j}^r}{1 + \sum_{r=1}^{q} \alpha_r (-1)^r \bar{j}^r}.$$

Rearranging and substituting parameter estimates for their unknown values, we obtain

$$\hat{f}_0 = f_1 \left(\frac{\sum_{r=0}^{p} \hat{\beta}_r (-1)^r \bar{j}^r}{1 + \sum_{r=1}^{q} \hat{\alpha}_r (-1)^r \bar{j}^r}\right)^{-1} = \frac{f_1}{\hat{b}_0},$$

where

$$\hat{b}_0 := \frac{\sum_{r=0}^{p} \hat{\beta}_r (-1)^r \bar{j}^r}{1 + \sum_{r=1}^{q} \hat{\alpha}_r (-1)^r \bar{j}^r},$$

and finally $\hat{C} = c + \hat{f}_0$.

Given the ability to computationally fit specific models, the question of model selection arises. We select the lowest order model meeting the following conditions. First, we

require $\hat{b}_0 > 0$ so that $\hat{f}_0 > 0$. (The issue of negative unobserved diversity estimators is as old as the problem itself: Fisher et al. (1943) recognized it when fitting the negative binomial distribution to Malayan butterfly data; see also Chao and Bunge (2002); Rocchetti et al. (2011)). Second, $1 + \alpha_1(j - \bar{j}) + \ldots + \alpha_q(j - \bar{j})^q$ can have no roots in $[0, J]$, so that (7) has no singularities in the relevant domain. Finally, the model must be computable, that is, must converge numerically. We advocate selecting the most parsimonious model that satisfies these requirements. If a model converges numerically, yields a positive $\hat{b}_0$, and has no singularities in the relevant domain we say it "satisfies all criteria" (SAC). Our final procedure, implemented in the R package *breakaway*, is described in Algorithm 1.

Algorithm 1 breakaway's model selection and reweighting procedure.

Require: frequency count table

if $\tau_{max} \geqslant 6$ **then**

weights $\leftarrow 1/j$

Fit (7) using [weights] for $p = 1, \ldots, 4$, $q = p - 1, p$ ▷ Total 8 models

if at least one model SAC **then**

while at least one model SAC **do**

current model $\leftarrow$ smallest model SAC

weights $\leftarrow$ (5) using fitted ratios $\{\hat{p}_j\}$ from current model as $\{p_j\}$

Fit (7) using [weights] for $p = 1, \ldots, 4$, $q = p - 1, p$

if current model does not SAC **then**

current model $\leftarrow$ smallest model SAC fit with weights $1/j$ ▷ Code 3

else Return the richness estimate, standard error and fitted ratios ▷ Code 2

else Return the WLRM ▷ Code 1

else "Not enough data to estimate richness"

In practice, Code 1 events rarely arise when frequency counts are accurately input. However, in the microbial case, chimeric filtering (e.g., Edgar (2013)) is generally required.

We now discuss error estimation. Applying the delta method, we find

$$\hat{\text{Var}}(\hat{f}_0) \approx f_1 \hat{b}_0^{-2} \left(1 - \frac{f_1}{n} + f_1 \hat{b}_0^{-2} \hat{\text{Var}}(\hat{b}_0)\right), \tag{9}$$

where $\hat{\text{Var}}(\hat{b}_0)$ is an empirical estimate of

$$\left\{\nabla \left(\frac{\sum_{r=0}^{p} \beta_r (-1)^r \bar{j}^r}{1+\sum_{r=1}^{q} \alpha_r (-1)^r \bar{j}^r}\right)\right\}^T \times \text{Cov}(\hat{\beta}_0, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots \hat{\alpha}_q) \times \nabla \left\{\frac{\sum_{r=0}^{p} \beta_r (-1)^r \bar{j}^r}{1+\sum_{r=1}^{q} \alpha_r (-1)^r \bar{j}^r}\right\},$$

and we treat the covariance between $\hat{b}_0$ and f_1 as negligible. This yields

$$\hat{\text{Var}}(\hat{C}) \approx \frac{n\hat{f}_0}{n+\hat{f}_0} + \hat{\text{Var}}(\hat{f}_0), \tag{10}$$

again treating the covariance term as negligible (as in Rocchetti et al. (2011)). A simulation study in the negative binomial case (Table 1) shows this approximation to be remarkably accurate.

Upon carrying out our estimation several outcomes are possible. First, the estimates of the parameters α_i and β_i may fall within the parameter spaces defined by Kemp (1968), or even within the Katz subset of the Kemp-type distributions, namely the negative binomial or Poisson. In this case we presume that the data is well-described by such a distribution; below we show a simulation of the negative binomial case indicating good performance and identification of that distribution. Second, the estimates may imply a legitimate terminating Kemp-type distribution. Third, the parameter estimates may not entail a distribution at all (the implied p_j would be negative); in this situation we can still use our method to estimate C but cannot give an interpretation to the probability model. Below we see that this behavior occurs on the Epstein dataset.

5. Simulations

Analyzing breakaway's behaviour for simulated negative binomial frequency counts demonstrates that when the true distribution is negative binomial, breakaway correctly infers this in more than 99% of cases, see Table 1. Note that Table 1 is based on simulating negative binomial counts near the boundary of the parameter space, and hence breakaway performs

well even in a near-pathological case. These choices of parameters lead to relatively high frequencies of rare species (singletons, doubletons, etc.), which is consistent with data structures observed in microbial diversity studies. For comparison we include the Chao-Bunge estimator, which was developed for the negative binomial; we see that breakaway has only a slightly larger standard error in exchange for its much greater generality.

[Table 1 about here.]

6. Applied data analysis

We demonstrate our approach on three microbial datasets, entitled Apples (Walsh et al., 2014), Soil (Schuette et al., 2010), and Epstein (S. Epstein, personal communication, February 28, 2014), all available as supplementary materials. We consider Apples a medium-diversity case, and Soil and Epstein as high-diversity.

Table 2 compares the results of breakaway with a number of competitors: the transformed and untransformed weighted linear regression model (tWLRM and uWLRM) of Rocchetti et al. (2011), the Chao-Bunge estimator (Chao and Bunge, 2002), the Chao (1987) lower bound (CLB), and 95% bootstrap confidence intervals for the nonparametric maximum likelihood (NPMLE) estimator of Norris and Pollock (1998) (c.f. also Wang and Lindsay (2005); Wang (2011)). Following standard practice we set $\tau = \tau_{max}$ for the WLRMs (where τ_{max} is the largest frequency before the first zero frequency) and $\tau = 10$ for the less robust Chao-Bunge estimator. By default, breakaway employs τ_{max} due to the (potentially) large number of parameters to be estimated, though the authors recommend confirming robustness of the estimate to the choice of τ when performing a richness analysis. The diagonal weighting structure for breakaway was employed as the tridiagonal model does not converge in any of these cases.

[Table 2 about here.]

We observe that in the medium diversity setting (Apples), breakaway's point estimate is comparable to other estimators, though its standard error is relatively large due to the larger number of parameters to be estimated. However, the advantage of breakaway is revealed in the high diversity cases. breakaway's standard error is less than half that of the uWLRM in the case of Soil. In the case of the Epstein dataset, the uWLRM and Chao-Bunge estimators both fail to produce any estimate. While breakaway's standard error is large relative to the estimate, we argue that the failure of both Chao-Bunge and the uWLRM to produce any estimate suggests that a high standard error is an honest assessment of the nature of the dataset. We cite the instability of the NPMLE as further evidence for this claim.

One appealing feature of regression-based estimators is that fit can be readily assessed, and the plausibility of the models employed by breakaway can be seen in Figure 1 in the case of the Soil dataset (ratio plots for the other datasets appear similar). This plot clarifies that the tWLRM standard errors in Table 2 are artificially low due to model misspecification. We argue that the added flexibility afforded by (7) replaces the need to log-transform to achieve a positive prediction for f_0.

Table 3 displays the empirical weights of the regression-based models for the Soil dataset. We observe that breakaway's weights, which are data-adaptive rather than based on the negative binomial model, appear to be a "middle ground" between the weights of the tWLRM and the uWLRM. We note that the uWLRM places more than 96% weight on the first ratio, providing an explanation for its poor fit as displayed in Figure 1.

[Table 3 about here.]

[Figure 1 about here.]

In the Soil, Apples, and Epstein datasets, the model with $p = q = 1$ was found to be sufficiently flexible to produce positive predictions for f_0, though other studied

datasets require higher order models. Parameter estimates for the Soil and Apples datasets correspond to distributional models with bounded support: if the support is considered bounded then normalization is possible and proper distributional models are implied. However, the parameter estimates for the Epstein dataset are not normalizable since the implied "probability structure" yields negative probabilities. We discuss the implications of this in Section 7. That breakaway correctly infers negative binomial distributions but rarely selects them when faced with real data is suggestive.

Similar to the uWLRM, NPML and Chao-Bunge estimators, breakaway does not guarantee a positive estimate. However, breakaway produces positive estimates in many situations where these estimators fail, and performs especially well in the high diversity situations that are typical of modern microbial datasets. The additional parameters in the ratio model contribute the flexibility required to produce positive estimates in these difficult cases and improve model fit. The weighting scheme is adaptive and thus more realistic than that of other regression-based estimates while preserving the appealing feature of a visual mechanism to confirm model specification. Estimates and standard errors are consistent with other estimators in datasets displaying low and medium diversity characteristics, and estimates produced by breakaway tend to exceed the Chao lower bound. The underpinning probabilistic structure is not confined to the mixed Poisson framework, and finally, the program is implemented in an R package freely available to the practitioner.

7. Conclusions and future directions

We have developed a diversity estimator based on fitting a class of distributions characterized by the functional form of their probability ratios. The method produces reasonable estimates with sensible standard errors in the medium and high diversity situations, and can outperform competitors in the high diversity situations that characterize modern microbial

datasets. Our estimator tends to produce results similar to nonparametric estimators when they exist. Diversity estimates and standard errors are consistent with simulation studies and the fitted structures are geometrically and intuitively plausible.

While this paper has focused on fitting ratios-of-polynomials models to frequency ratios via nonlinear regression, other probabilistically-based functional forms could be investigated. The Lerch distribution, which arises as the stationary distribution of a stochastic process, can be characterized by either the form of its frequency ratios or by its probability mass function, which exists in closed form (unlike general Kemp-type distributions). We attempted to estimate the parameters of a zero-truncated Lerch distribution via maximum likelihood and found that the algorithm did not converge for any of the datasets investigated above, which is a well-documented issue with direct estimation of parameters in the species problem (Bunge and Fitzpatrick, 1993; Bunge et al., 2014). Furthermore, only for the Apples dataset did nonlinear regression estimates of the Lerch parameters converge, suggesting that the Lerch distribution may not be flexible enough to model a broad range of diversity datasets. In the case of the Apples dataset, the Lerch nonlinear regression fits were similar to breakaway. However, the Lerch model is unstable, estimating $\hat{C} = 1727$ (1987) compared to breakaway's $\hat{C} = 1552$ (305).

Our method amounts to modeling the frequency count ratios via a low-dimensional functional representation. By parsimoniously fitting this low-dimensional structure, the method may depart from a distribution-based structure altogether, and we have presented a dataset for which estimation is only possible if this is permitted. We argue that the fitted empirical nonlinear regression function still permits estimation (prediction) of f_0 in this case, although we strongly advise visual confirmation of goodness of fit before proceeding. However, in many contexts a distributional model is implied if the support of the frequency count distribution can be regarded as bounded. Indeed, the right extremity

of the population frequency count distribution is the frequency of the most common taxon, which must be finite if the population size is believed to be finite.

Under our procedure, models may be excluded for reasons other than goodness of fit, and classical model selection diagnostics do not apply without accounting for the conditioning inherent in this procedure. Asymptotic properties have not been considered here due to the complexity arising from the algorithmic procedure. However, simulations in the negative binomial case support hypotheses of both consistency and normality. Comprehensive asymptotic analysis of breakaway is a topic for future research.

Finally, the sampling procedure used to construct frequency count tables in the microbial setting is complex, and many bioinformatic preprocessing tools distort and bias frequency tables (Caporaso et al., 2010; Edgar, 2013). Measurement error may be extreme, especially in the singleton count. Regression-based ratio methods are in an ideal position to adjust diversity estimates to account for this measurement error. This is an ongoing topic of the authors' research.

8. Supplementary Materials

The Apples, Epstein and Soil datasets are available via the Biometrics website on Wiley Online Library. The R package *breakaway* is available via CRAN.

Acknowledgements

The authors would like to thank two anonymous referees and an associate editor for many suggestions that improved both the details and style of this paper.

References

Böhning, D., Baksh, M. F., Lerdsuwansri, R., and Gallagher, J. (2013). Use of the ratio plot in capture–recapture estimation. *Journal of Computational and Graphical*

Statistics **22,** 135–155.

Böhning, D. and Kuhnert, R. (2006). Equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics* **62,** 1207–1215.

Böhning, D., Rocchetti, I., Alfó, M., and Holling, H. (2014). A flexible ratio regression approach for zero-truncated capture-recapture counts. Submitted manuscript.

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

Bunge, J., Willis, A., and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* **1,** x–x.

Bunge, J., Woodard, L., Böhning, D., Foster, J. A., Connolly, S., and Allen, H. K. (2012). Estimating population diversity with CatchAll. *Bioinformatics* **28,** 1045–1047.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7,** 335–336.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43,** 783–791.

Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics* **58,** 531–539.

Dacey, M. F. (1972). A family of discrete probability distributions defined by the generalized hypergeometric series. *Sankhyā Series B* **34,** 243–250.

Dethlefsen, L., Huse, S., Sogin, M. L., and Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota. *PLoS biology* **6,** e280.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature methods* **10,** 996–998.

Fisher, R. A., Corbet, S., and Williams, C. B. (1943). 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,** 42–58.

Gao, W., Weng, J., Gao, Y., and Chen, X. (2013). Comparison of the vaginal microbiota diversity of women with and without human papillomavirus infection: a cross-sectional study. *BMC infectious diseases* **13,** 271.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1985). *Exploring Data Tables, Trends and Shapes*. Wiley.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*, volume 444. John Wiley & Sons.

Katz, L. (1945). *Characteristics of Frequency Functions Defined by First Order Difference Equations*. PhD thesis, University of Michigan.

Kemp, A. W. (1968). A wide class of discrete distributions and the associated differential equations. *Sankhyā Series A* **30,** 401–410.

Kemp, A. W. (2010). Families of power series distributions, with particular reference to the Lerch family. *Journal of Statistical Planning and Inference* **140,** 2255–2259.

Lladser, M. E., Gouet, R., and Reeder, J. (2011). Extrapolation of urn models via poissonization: Accurate measurements of the microbial unknown. *PloS one* **6,**.

Mao, C. X. and Lindsay, B. (2007). Estimating the number of classes. *Annals of Statistics* **35,** 917–930.

McDonald, D. R. (1980). On the Poisson approximation to the multinomial distribution. *Canadian Journal of Statistics* **8,** 115–118.

Ng, C. M., Ong, S.-H., and Srivastava, H. M. (2013). Parameter estimation by hellinger type distance for multivariate distributions based upon probability generating functions. *Applied Mathematical Modelling* **37,** 7374–7385.

Norris, J. L. and Pollock, K. H. (1998). Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics* **5,** 391–402.

Pestana, D. D. and Velosa, S. F. (2004). Extensions of Katz–Panjer families of discrete

distributions. *REVSTAT–Statistical Journal* **2,**.

Puri, P. S. and Goldie, C. M. (1979). Poisson mixtures and quasi-infinite divisibility of distributions. *Journal of Applied Probability* **16,** 138–153.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rocchetti, I., Bunge, J., and Bohning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics* **5,** 1512–1533.

Roos, B. (1999a). Metric multivariate Poisson approximation of the generalized multinomial distribution. *Theory of Probability and Its Applications* **43,** 306–316.

Roos, B. (1999b). On the rate of multivariate Poisson convergence. *Journal of Multivariate Analysis* **69,** 120–134.

Schuette, U. M., Abdo, Z., Foster, J., Ravel, J., et al. (2010). Bacterial diversity in a glacier foreland of the high arctic. *Molecular Ecology* **19,** 54–66.

Seber, G. and Wild, C. (1989). *Nonlinear Regression*. Wiley.

Tripathi, R. and Gurland, J. (1977). A general family of discrete distributions with hypergeometric probabilities. *Journal of the Royal Statistical Society: Series B* **39,** 349–356.

Walsh, F., Smith, D. P., et al. (2014). Restricted streptomycin use in apple orchards did not adversely alter the soil bacteria communities. *Frontiers in Microbiology* **4,**.

Wang, J.-P. (2011). SPECIES: An R package for species richness estimation. *Journal of Statistical Software* **40,** 1–15.

Wang, J.-P. Z. and Lindsay, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100,** 942–959.

White, H. and Domowitz, I. (1984). Nonlinear regression with dependent observations. *Econometrica* **52,** pp. 143–162.

Zörnig, P. and Altmann, G. (1995). Unified representation of Zipf distributions. *Computational Statistics & Data Analysis* **19,** 461–473.

Received January 2015.

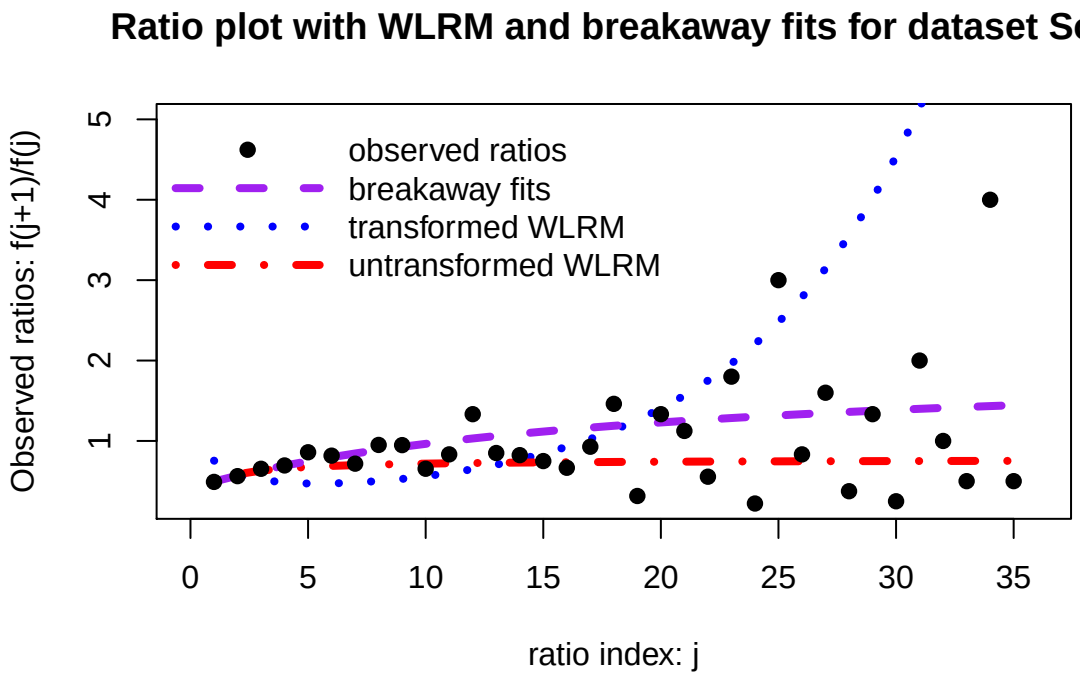

Figure 1. Comparison of fitted ratios for regression-based diversity estimators: the transformed and untransformed weighted linear regression model (WLRM), and breakaway. The above pertains to the Soil dataset.

Table 1

Mean estimated standard error (calculated using (10)) and actual standard error in estimating $\hat{C}$ when the true frequency count distribution is negative binomially distributed with probability parameter p, size parameter n and density $\binom{x+n-1}{x}p^n(1-p)^x$. Note that $\beta_0 = n(1-p) = 5$ while $\beta_1 = (1-p)$. The percentage of replications that resulted in breakaway correctly inferring the negative binomial (NB) distribution is also shown, along with the actual standard error for the Chao-Bunge (C-B) estimator with $\tau = 10$. Results are based on 10,000 replications.

True C	(prob, size)	s.ê.$(\hat{C})$	True s.e.$(\hat{C})$	% inferred NB	True s.e.(C-B)
50,000	$(0.99, 500)$	20.69	20.84	99.57	20.23
50,000	$(0.95, 100)$	19.56	19.73	100.00	19.06
5,000	$(0.99, 500)$	6.61	6.57	99.70	6.37
5,000	$(0.95, 100)$	6.24	6.20	99.99	5.98
500	$(0.99, 500)$	2.18	2.19	99.23	2.05
500	$(0.95, 100)$	2.10	2.09	99.25	1.94

Table 2

Comparison of the method breakaway with other diversity estimators: the transformed and untransformed weighted linear regression model (tWLRM and uWLRM), Chao-Bunge estimator, Chao lower bound (CLB), and 95% bootstrap confidence intervals (1000 resamples) for the nonparametric maximum likelihood estimator (NPMLE). Standard errors are given in parentheses where appropriate.

Dataset	Apples	Soil	Epstein
breakaway	1552 (305)	5008 (689)	2162 (1699)
uWLRM	1330 (77)	7028 (1743)	*
tWLRM	1179 (28)	3438 (87)	1849 (380)
Chao-Bunge	1377 (63)	4865 (257)	*
CLB	1241 (38)	3674 (84)	1347 (160)
NPMLE	(1049,2776)	(2885,38442)	(730,6936721)

Table 3

Comparison of empirical weights placed on the first through fifth data points for the regression-based diversity estimators: the transformed and untransformed weighted linear regression model (WLRM), and breakaway. The below pertains to the Soil dataset.

Model	w_1^2	w_2^2	w_3^2	w_4^2	w_5^2
Transformed WLRM	0.645	0.186	0.071	0.033	0.020
Untransformed WLRM	0.966	0.032	0.002	0.000	0.000
breakaway	0.738	0.186	0.051	0.015	0.005