Modeling Predator Preferences

Edward A. Roualdes and Simon Bonner
University of Kentucky

2014-09-30

Abstract

The modeling of a predator's prey selection has a breif history that includes many intuitive indices, few of which have both reasonable statistical justification and tractable asymptotic properties. Here, we provide a simple model that meets both of these criteria, while extending previous work to include an array of multiple species and time points. Further, we apply the Expectation-Maximization algorithm to cases where exact counts of the number of prey species eaten in a particular timer period is not observed. Simulations yada yada yada...

1 Introduction

Modeling a predator's food selection was given much attention for a short period of time; see Strauss [1979] and the references within. Many, if not all, of the indices developed, though intuitive, focused on a snapshot in time and only on one prey species, and they rarely had practicle asymptotic properties. This has left interested practicioners to the most computationally manageable of the techniques, and has completely ignored simultaneous testing across an array of both species and time.

The model presented here, maintains tractable asymptotic properties while being general enough to take into account an array of speices and time points. By modeling both time and any number of prey species, we are able to see a more detailed anlyais of the predator's eating preferences. Further, the simplicity of the model allows us to consider predators for which exact tallies of the number of each variety of prey species eaten within any given time period is not observed. Instead, we rely on the researcher being able to DNA sequence the contents of the predator's gut, and make a simple binary conclusion: yes this predator ate some of that prey species during this time period, or no they did not. Under certain situations, our models is able to accurately estimate parameters of interest despite having lost much information.

To showcase our model, we performed simulations for both scenarios when full count data from the predators' guts are observed, and when instead only binary repsonses, indicating specific prey species were eaten, are observed. Our `R` [Core Team, 2014] package, named `spiders`, is available on CRAN so that the broader community of researchers can similarly apply our methods.

2 Methods

2.1 Data

We assume data are collected in the following manner. Traps are dispersed, for T time periods, throughout the habitat of the predator and prey of interest. Prey species, indexed by $s \in \{1, \ldots, S\}$, are collected in the traps and counted at each time period. We assume these counts represent the number of prey species s the predator will encounter on average during time period $t \in \{1, \ldots, T\}$. The counts of prey species s caught during time period t are hypothesized to be independent draws from a Poisson distribution with rate parameter γ_{st}. We further assume that the number of prey species found in the gut of the trapped predators, also follows a Poisson distribution with rate λ_{st}. Here, the parameter λ_{st} represents the rate at which the predator ate the encountered prey species s during time period t. By modeling λ_{st} and γ_{st} we are able to test claims about a predator's eating preferences.

The use of Poisson distributions make the following implicit assumptions: 1) traps independently catch the prey species of interest, 2) predators eat, at a constant rate, prey species independently, 3) predators eat indepedent of each other. Our model's ability to accurately portray predators' habitats and underlying eating preferences dependes on the degree to which these assumptions are broken.

Let $X_{jst} \overset{\text{iid}}{\sim} \mathcal{P}(\lambda_{st})$ denote the number of prey species s that predator j ate during occurrence t where $j \in \{1, \ldots, J_t\}$. Let $Y_{ist} \overset{\text{iid}}{\sim} \mathcal{P}(\gamma_{st})$ denote the number of prey species s found in trap i during occurrence t, $i \in \{1, \ldots, I_t\}$. Formal statistical statements about the relative magnitudes of the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ offer insights to the relative rates at which predators eat particular prey species.

We consider five variations on the relative magnitude of $\lambda_{st}/\gamma_{st} = c_{st}$. These five hypotheses each allow c_{st} to vary by time, prey species, both, or neither. Because the five hypotheses are nested, a natural testing order is suggested in Figure 1.

1. $c_{st} = 1, s = 1, \ldots, S; t = 1, \ldots, T$
2. $c_{st} = c, s = 1, \ldots, S; t = 1, \ldots, T$
3. $c_{st} = c_s, s = 1, \ldots, S$
4. $c_{st} = c_t, t = 1, \ldots, T$
5. $c_{st} = c_{st}, s = 1, \ldots, S; t = 1, \ldots, T$

2.2 Fully Observed Data

The likelihood function that allows for estimation of these parameters is as follows. Since we assume X_{jst} is independent of Y_{ist} we can simply multiply the respective Poisson probability density functions together, and then form products over all s, t to get the likelihood.

$$L(x_{jst}, y_{ist}|\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \prod_{t=1}^{T}\prod_{s=1}^{S}\left\{\prod_{j=1}^{J_t} f_X(x_{jst}|\boldsymbol{\lambda}) \prod_{i=1}^{I_t} f_Y(y_{ist}|\boldsymbol{\gamma})\right\}. \tag{1}$$

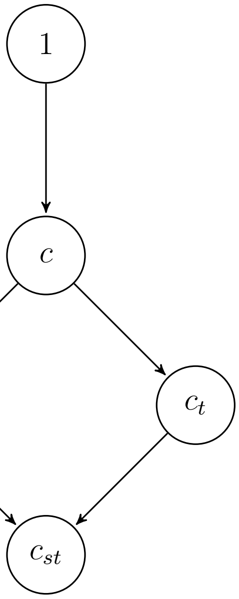

Figure 1: Hierarchy of hypotheses.

Writing all 5 hypotheses as $\lambda_{st} = c_{st}\gamma_{st}$, we can, in some cases find analytic solutions for the maximum likelihood estimates of c_{st} and γ_{st}. When the data are balanced $J_t = J$, $I_t = I$, and $c_{st} = c$ analytic solutions exist – what about $c_{st} = 1$. In all other cases, analytic solutions are not readily available and instead we rely on the fact that the log-likelihood $l(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \log L$ is concave. We maximize the log-likelihood by iteratively solving partial derivatives of l, with respect to c_{st} and γ_{st}, set equal to zero

$$\hat{c} = \frac{\sum_{s,t} X_{\cdot st}}{\sum_t J_t \sum_s \gamma_{st}}, \quad \hat{c}_t = \frac{\sum_s X_{\cdot st}}{J_t \sum_s \gamma_{st}}, \text{ or } \quad \hat{c}_s = \frac{\sum_t X_{\cdot st}}{\sum_t J_t \gamma_{st}}, \text{ and } \quad \hat{\gamma}_{st} = \frac{X_{\cdot st} + Y_{\cdot st}}{J_t c_{st} + I_t}.$$

where $X_{\cdot st} = \sum_{j=1}^{J_t} X_{jst}$ and $Y_{\cdot st} = \sum_{i=1}^{I_t} Y_{ist}$.

2.3 Unobserved Counts

Working with biologists who study spider eating preferences, we have found that not all predators' allow for easily counted prey species in their guts. As an alternative strategy, we can rely on the DNA sequencing of a sample from the predators' guts. If such sequencing returns a positive response, say a 1 if a particular predator ate prey species s and 0 otherwise, we can, albeit with some information lost, model predators' eating preferences with the above framework using the EM algorithm.

When the data X_{jst} are not observed, and instead a boolean response indicating if a predator ate any number of prey species s during time t is observed, we can still, to some degree, estimates the parameters of interest $\boldsymbol{\lambda}, \boldsymbol{\gamma}$. Because some information is observed, we can treat the counts as missing and use the EM algorithm to find the maximum likelihood estimates of the observed data likelihood.

We denote the binary response that the predator did in fact eat at least one prey species s in time period t by $Z_{jst} = 1(X_{jst} > 0)$. Now, the observed data are independent and identically distributed Bernoulli observations with parameter $p_{st} = 1 - \exp\{-\lambda_{st}\}$. Using this we can find the complete data likelihood by first noting that the conditional distribution of the unobserved data X_{jst} given $Z_{jst}, Y_{ist}, \boldsymbol{\lambda}, \boldsymbol{\gamma}$ is a truncated Poisson distribution

$$f_{X|Y,Z,\boldsymbol{\lambda},\boldsymbol{\gamma}}(x_{jst}) = \frac{\exp\{-\lambda_{st}\}\lambda_{st}^{x_{jst}}}{(1-\exp\{-\lambda_{st}\})x_{jst}!} 1(x_{jst} > 0) \quad \text{where} \quad \mathbb{E}_{[X|Y,Z]} X_{jst} = \frac{\lambda_{st} \exp\{\lambda_{st}\}}{\exp\{\lambda_{st}\} - 1}.$$

From this conditional distribution we get the joint distribution of X_{jst}, Z_{jst}

$$f_{X,Z|\boldsymbol{\lambda}}(x_{jst}, z_{jst}) = \begin{cases} \exp\{-\lambda_{st}\}, & x_{jst} = 0 \text{ and } Z_{jst} = 0 \\ \frac{\exp\{-\lambda_{st}\}\lambda_{st}^{x_{jst}}}{x_{jst}!}, & x_{jst} > 0 \text{ and } Z_{jst} = 1 \\ 0 & \text{otherwise} \end{cases}$$

The E-step of the EM algorithm is completed by computing the expected value of the complete data log-likelihood with respect to $f_{X|Y,Z,\boldsymbol{\lambda},\boldsymbol{\gamma}}(x_{jst})$ to get

$$\begin{aligned} \mathbb{E}l_{comp} &= \mathbb{E} \log f_{X,Z|\boldsymbol{\lambda}}(X_{jst}, z_{jst}) + \log f_{Y|\boldsymbol{\gamma}}(y_{ist}) \\ &= \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{J_t} \mathbb{E} \log f_{X,Z|\boldsymbol{\lambda}}(X_{jst}, z_{jst}) + \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{i=1}^{I_t} \log f_{Y|\boldsymbol{\gamma}}(y) \\ &= \sum_{s,t,j} (-\lambda_{st} + z_{jst}\mathbb{E}X_{jst} \log \lambda_{st}) + \sum_{s,t} (-I_t\gamma_{st} + Y_{\cdot st} \log I_t\gamma_{st}) + \text{const} \\ &= \sum_{s,t} (-J_t\lambda_{st} + z_{\cdot st}\mathbb{E}X_{jst} \log \lambda_{st}) + \sum_{s,t} (-I_t\gamma_{st} + Y_{\cdot st} \log I_t\gamma_{st}) + \text{const.} \end{aligned}$$

The EM algorithm requires iteratively solving $(\boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}) = \arg\max_{\boldsymbol{\lambda},\boldsymbol{\gamma}} \mathbb{E}l^k_{comp}$. Though no analytic solution to this maximization exists, one idea is to iteratively solve partial derivatives of $\mathbb{E}l_{comp}$ set equal to zero until convergence. In fact, as we only need find parameter values that increase the observed likelihood, we forgo fully iterating to find the maximum values of the parameters and instead perform just one step uphill within each EM iteration. This strategy is significantly less computationally intensive, thus generating a much faster generalized EM (GEM) algorithm.

This GEM algorithm accurately estimates the parameters when values of λ_{st} are relatively small, such that zeros are prevalent in the data Z_{jst}. In this case, not too much information is lost since estimation of $\mathbb{E}Z_{jst}$ can be estimated well by the proportion of observed zeros. On the other hand, if the predator consistently eats a given prey species, few to no zeros will show up in the observed data and $\mathbb{E}Z_{jst}$ is estimated to be nearly 1. The loss of information is best seen by attempting to solve for λ_{st} in the equation $1 = \mathbb{E}Z_{jst} = 1 - \exp\{-\lambda_{st}\}$; essentially λ_{st} is sent off to $+\infty$.

2.4 Testing

All hypotheses are evaluated via a likelihood ratio test (LRT), with statistic

$$\Lambda(X, Y) := -2 \log \frac{\sup L(\theta_0 | X, Y)}{\sup L(\theta_1 | X, Y)},$$

where θ_0, θ_1 represent the parameters estimated under the null and alternative hypotheses, respectively. It is well known that the asymptotic distribution of Λ is a χ^2_ρ distribution with ρ degrees of freedom. Under the EM algorithm we use $L_{obs}(Z, Y)$ as the likelihood in the calculation of Λ.

The degrees of freedom ρ are set equal to the number of free parameters available in the stated hypotheses under question. If we put the null hypothesis to be $H_0 : \lambda_t = c_t \gamma_t$, for all t and contrast this against $H_1 : \lambda_{st} = c_{st} \gamma_{st}$ then there are $\rho = 2(S \cdot T) - S \cdot T - T = S \cdot T - T$ degrees of freedom.

A set of hypotheses is determined by the p-value of the χ^2_ρ distribution. Hence, with a level of significance, α, the null hypothesis is rejected in favor of the alternative hypothesis if $\mathbb{P}(\chi^2_\rho > \Lambda) < \alpha$.

2.5 Testing c_{st}

After determining which model best fits the data, more detail can be extracted through a hypothesis test on the elements of c_{st}, or in vector notation as $\mathbf{c} \in \mathbb{R}^{S \cdot T}$. Let the elements of $\hat{\mathbf{c}}$ be the estimated values, $\hat{c}_{st}$, as estimated via the above maximum likelihood framework. Since $\hat{\mathbf{c}}$ is asymptotically normally distributed, any linear combination of the elements is also asymptotically normally distributed. For instance, let a be a vector of the same dimension of $\hat{\mathbf{c}}$. Then $a^t \hat{\mathbf{c}}$ is asymptotically distrbuted as $\mathcal{N}(a^t \mathbf{c}, a^t \Sigma a)$, where Σ is the covariance matrix of the asymptotic distrbution of $\hat{\mathbf{c}}$.

Suppose, for example, that the hypothesis c_s is determined to best fit the data with s ranging $s = 1, 2$. We can test to see whether or not the two species s_1 and s_2 are statistically equally preferred under the null hypothesis $\hat{c}_{s_1} = \hat{c}_{s_2}$. This hypothesis is alternatively written in vector notation as $a^t \hat{\mathbf{c}} = 0$, where $a = (1, -1)^t$. Tests of the following form $H_0 : a^t \mathbf{c} = \mu$ against any alternative of interest are then standard Z-tests. Similarly, confidence intervals of any level can be obtained if desired.

3 Simulations

Our simulations hypothesize two prey species, and five time points. Of the hierarchy of hypotheses, we simulate data under three null hypotheses: c, c_s, c_t. Sample sizes are randomly chosen from four overlapping levels. Let "small" sample sizes be randomly sampled numbers in $[20, 50]$, "medium" encompass $[30, 75]$, "large" $[50, 150]$, and "huge" $[100, 200]$. Hence, we randomly sample prey and predator gut count observations for each time period from one of the sample size levels, then cycle through all hypotheses. This is repeated for each level of sample size. We simulate 500 replicate datasets for each of the twelve scenarios above for both types of data, fully observed count data, X_{jst}, and for non-count data, when we observe only a binary response, $Z_{jst} = 1(X_{jst} > 0)$. A subset of the examples are shown here; the interested reader is referred to the supplementary materials for the complete set of simulation results.

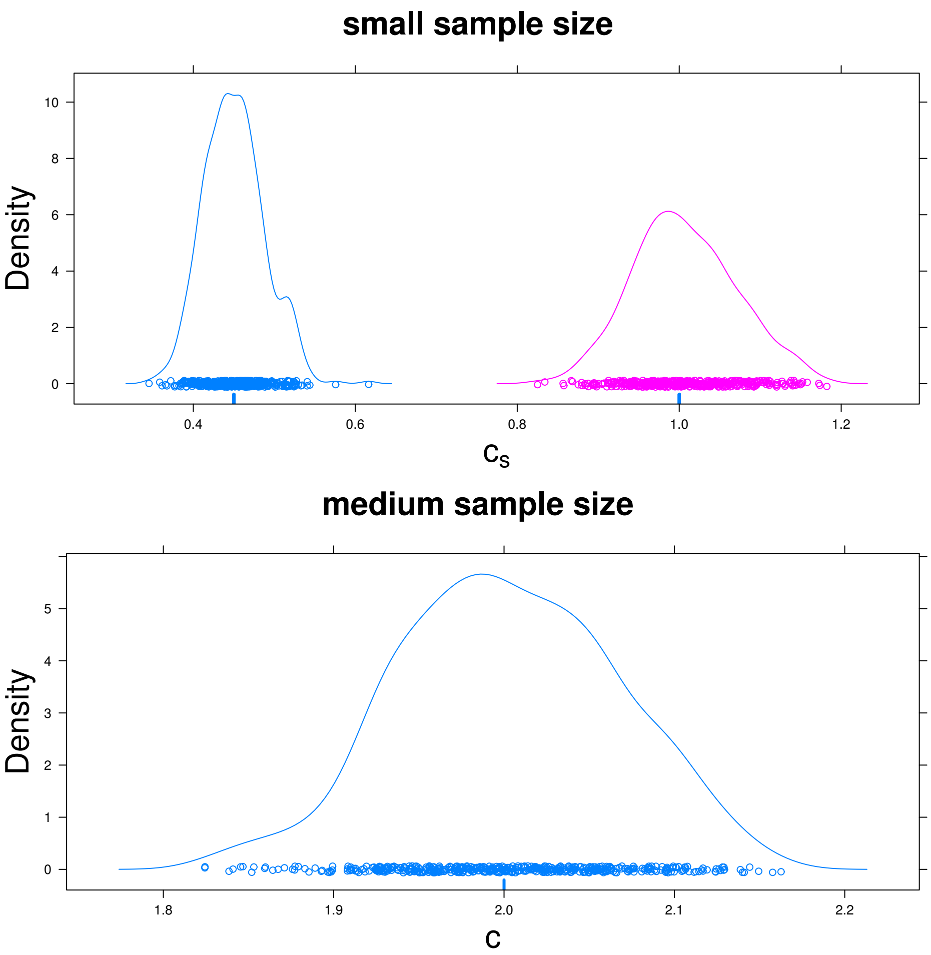

Figure 2: moar words

For all simulated data, the true parameter values for the rate at which prey species are encountered in the wild are fixed to be $\gamma_{st} = \pi$, $\forall s, t$. The values of λ_{st} are set with respect to each null hypothesis. For the hypothesis $c_{st} = c$, where predator preferences don't vary by either time or species, we consider $\lambda_{st} = 2\pi, \forall s, t$. Under the hypothesis c_s, the ratio of rates vary by species only, so we put $\lambda_{1t} = \sqrt{2}$ and $\lambda_{2t} = \pi$. Hence, $c_1 = \sqrt{2}/\pi \approx 0.45$ and $c_2 = 1$. For the second hypothesis, the ratio of rates vary by time t. Here, we put $\lambda_{st} = t$ for $t \in \{1, \ldots, 5\}$.

Figure 2 shows our model fitting to the fully observed count data. In the first display, the parameters $c_1 \approx 0.4501$ and $c_2 = 1$ are estimated as $\hat{c}_1 = 0.4508$ and $\hat{c}_2 = 1.007$, with standard errors of $\mathrm{se}(\hat{c}_1) = 0.037$ and $\mathrm{se}(\hat{c}_2) = 0.064$. The second display estimates the parameter $c = 2$ to be $\hat{c} = 2.002$ with standard error $\mathrm{se}(\hat{c}) = 0.065$.

As noted above we find that the EM algorithm accurately estimates the parameters when values of $\lambda_{st} = c_{st}\gamma_{st}$ are small. Figure 3 contains density plots of the EM algorithm's estimates for the hypotheses c_s, c_t for small and huge sample sizes, respectively. When data are simulated under the null hypothesis c_s, we find, even for the small sample size that point estimates are reasonably accurate. Though, when the values of λ_{st} are of sufficient size to

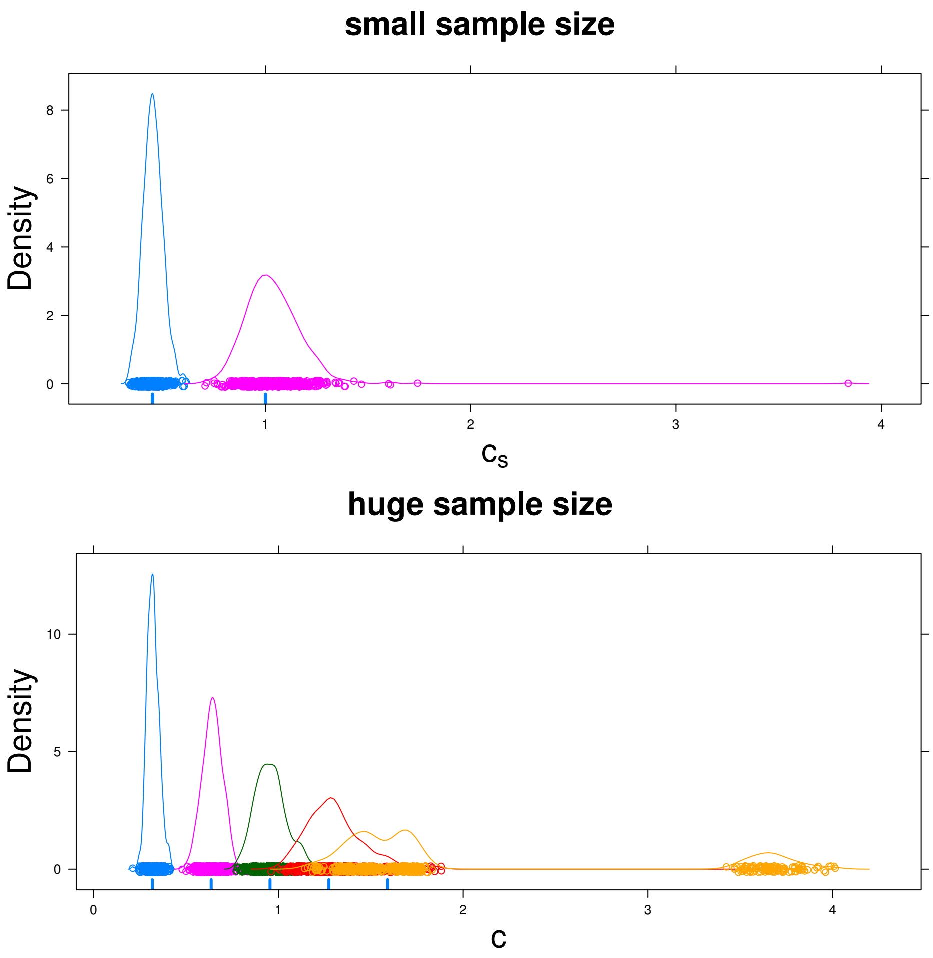

Figure 3: Density plots of the EM algorithm's estimates of c_s and c_t with small and huge sample sizes, respectively. The estimates near 4 for both of the plots represent samples for which few zeros appeared in the data and the EM algorithm had a particularly difficult time estimating the true parameter value.

make zeros in the simulated data less common, the algorithm occasionally over-estimates the true values. This effect is easily seen in Figure 3 for larger values of c_t despite the increased sample size, but it also seen, less dramatically, in the density plot for the hypothesis c_s. The plots of γ_{st} under the EM algorithm are not given as we do not consider missing data in the estimation of these parameters.

Because of the tendency for the EM algorithm to over-estimate some parameter values, bias for non-count data is significantly worse than it is for the count data. Table 1 displays the max absolute bias across all indices s, t of λ_{st} for each simulation scenario. In most cases, these numbers decrease, as expected, as the sample size increases.

4 Discussion

should we include mention of the spiders dataset?

		small	medium	large	huge
count	c_t	0.040	0.013	0.020	0.011
	c_s	0.028	0.015	0.010	0.005
	c	0.030	0.016	0.017	0.012
non-count	c_t	4.151	3.155	0.941	1.423
	c_s	0.138	0.096	0.044	0.032
	c	2.924	2.078	0.669	0.547

Table 1: Max absolute bias of all λ_{st} for each simulation scenario.

The model developed here advances the statistics literature available to determine predators' preferences by testing simultaneously across an array of multiple prey species and time. This is acheived via a simple, but statistically powerful, likelihood ratio test. Further testing of the ratio of rates for which predators eat to encounter prey species allows researchers to make specific conclusions about predators' preferences. For instance, rates acros time can be estimated to make statements about seasonal effects on a predator's eating habits, or relative rates across species groups allows for statements about the preferences for species s_1 to that of s_2.

When counts of predators' gut contents are not fully observed, and instead only a binary response indicating the existence of the prey species in the gut is observed, we are able to treat the counts as missing data. By modelling all of the observed data, both the binary responses and the number of prey species caught, and the missing count data, we are able to use the EM algorithm to extract as much information from the data as possible. Though this is nice in theory, in practice the success of this tweak to our original model is limited by the magnitude of the unknown parameters λ_{st}.

Further developments of our model could be beneficial. Taking into account other environmental variables that might effect a predators' eating habits, such as rain or temperature, say, might be advantageous.

An `R` package, named `spiders`, is available on CRAN at `http://cran.r-project.org/web/packages/spiders/index.html` and fits all the methods discussed above.

References

R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

Richard E Strauss. Reliability estimates for ivlev's electivity index, the forage ratio, and a proposed linear index of food selection. *Transactions of the American Fisheries Society*, 108(4):344–352, 1979.