# Modeling Predator Preferences

Edward A. Roualdes and Simon Bonner
University of Kentucky

2014-12-04

**Abstract**

The literature on modeling a predator's prey selection 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. A simulation study is provided to demonstrate the accuracy of our methods.

## 1 Introduction

The indices most commonly used to estimate a predator's food preferences, or selectivity, are relatively old [Ivlev, 1964, Jacobs, 1974, Chesson, 1978, Strauss, 1979, Vanderploeg and Scavia, 1979, Chesson, 1983], and yet continue to generate numerous publications; a quick search returns hundreds of publications that cite these fundamental papers, many of which are quite recent [Clements et al., 2014, Hansen and Beauchamp, 2014, Hellström et al., 2014, Lyngdoh et al., 2014, Madduppa et al., 2014]. These indices, though intuitive, lack statisical justification, focus on a snapshot in time, few allow more than one prey species to be considered, and practical asymptotic statistical properties are even more rare [Lechowicz, 1982]. We propose an intiutive statistical model to determine and statistically test predators' prey preferences across an array of time and multiple prey species.

A comprehensive overview by Lechowicz [1982], which was later summarized by Manly et al. [1992], details the benefits and faults of the most popular indices. According to these reviews, a majority of the indices give comparable results, save Strauss' linear index $L$, despite the fact that most of the methods differ by range and linearity of response. While Lechowicz [1982] recommends just one index, $E^*$ by Vanderploeg and Scavia [1979] as the "single best," albeit imperfect, index, Manly et al. [1992] instead takes the approach of suggesting against the subset of indices which don't "estimate any biologically meaningful value." $E^*$, an element of the Manly et al. [1992] suggested indices, is recommended on the basis of random feeding denoted by the index value 0, a possible range of $[-1, 1]$ (though $E^* = 1$ is nigh impossible), and because the index is based on the predator's choice of prey as a function of both the availability of the prey as well as the availability of all other prey.

1

The downside to this index is its lack of reasonable statistical properties [Lechowicz, 1982], thus making standard errors and hypothesis testing somewhat of a mystery. This is, in fact, a common fault amongst most of the indices.

To encourage more formal statistical inference, and simultaneously generalize predators' selectivity to animal resource selection, Manly et al. [1992] proposed the use of generalized linear models (GLM). The well established literature on GLMs allowed for hypothesis testing to replace the indices. The model we present here, while restricted to predators' eating preferences, is a comprimise between these two extremes, indices and GLMs. Our model offers formal hypothesis testing and inference similar to the GLMs of Manly et al. [1992], but also provides meaningfull single number summaries of the predator's dietary preferenes. To do this, we estimate the rate at which a predator consumes the prey of interest instead of estimating the proportion of consumed to available prey. An outcome of our model, is that we give up the somewhat arbitrary preference for an index to have 0 denote random eating and the range of $[-1, 1]$.

Our model enables formal hypothesis testing and statistical inference, while being general enough to perform statistical tests across multiple species and time points. This provides researchers a more detailed analysis of the predator's eating preferences. Further, because our model is based on Poisson distribution, members of the well studied exponential family, we are able to estimate the parameters of interest even when exact tallies of the number of each 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: this predator ate some of that prey species during this time period, or did not. Alongside our model, we offer an `R` Core Team [2014] package named `spiders` that fits all of 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. The number of prey species $s$ the predator will encounter on average during time period $t \in \{1, \ldots, T\}$ are considered random draws from a Poisson distribution with rate parameter $\gamma_{st}$. We further assume that the number of prey species found in the gut of the similarly trapped predators follows a Poisson distribution with rate $\lambda_{st}$. Here, the parameter $\lambda_{st}$ represents the rate at which the predator ate prey species $s$ during time period $t$. By modeling $\lambda_{st}$ and $\gamma_{st}$ 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 prey species independently, 3) predators eat independent of each other.

We denote the number of predators and the number of prey species caught, in each time period $t$, by $J_t$ and $I_t$, respectively. Let $X_{jst} \overset{\text{iid}}{\sim} \mathcal{P}(\lambda_{st})$ represent the number of prey species $s$ that predator $j$ ate during occurrence $t$, where $j \in \{1, \ldots, J_t\}$. Let $Y_{ist} \overset{\text{iid}}{\sim} \mathcal{P}(\gamma_{st})$ represent

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. $c_{st} = 1$

2. $c_{st} = c$

3. $c_{st} = c_s$

4. $c_{st} = c_t$

5. $c_{st} = c_{st}$

The first hypothesis states that predators and traps sample all prey species at the same rate. One imagines this is the case if the predator simply eats that which comes within its reach, thus suggesting a diet indifference. The second says that predators sample prey proportionally across all time periods. The third hypothesis says that predators sample different prey species at different rates, but each rate is steady across time. This implies that the predator expresses preferences for one prey species over another, but is unresponsive to changes due to time. Conversely, the forth hypothesis implies that each prey species is sampled similarly within each time period, while the rates across time are allowed to change. The fifth assumes a predator's selection varies by both time and prey species. This would make sense if environmental variables, say weather, or prey availability, and taste were affecting predators' selection strategies.

## 2.2 Fully Observed Count 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, and then form products over all $s, t$ to get the likelihood.

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

Writing all five hypotheses as $\lambda_{st} = c_{st} \gamma_{st}$, we can, in the simplest cases, find analytic solutions for the maximum likelihood estimates of $c_{st}$ and $\gamma_{st}$. Under the hypothesis $c_{st} = 1$, and when the data are balanced $J_t = J$, $I_t = I$, and $c_{st} = c$ analytic solutions exist. Namely, these solutions are

$$\hat{\gamma}_{st} = \frac{X_{\cdot st} + Y_{\cdot st}}{J_t + I_t}, \quad \text{and} \quad \hat{c} = \frac{I \sum_{s,t} X_{\cdot st}}{J \sum_{s,t} Y_{\cdot st}}, \quad \hat{\gamma}_{st} = \frac{X_{\cdot st} + Y_{\cdot st}}{I \left( \frac{\sum_{st} X_{\cdot st}}{\sum_{st} Y_{\cdot st}} + 1 \right)}$$

respectively, where $X_{\cdot st} = \sum_{j=1}^{J_t} X_{jst}$ and $Y_{\cdot st} = \sum_{i=1}^{I_t} Y_{ist}$.
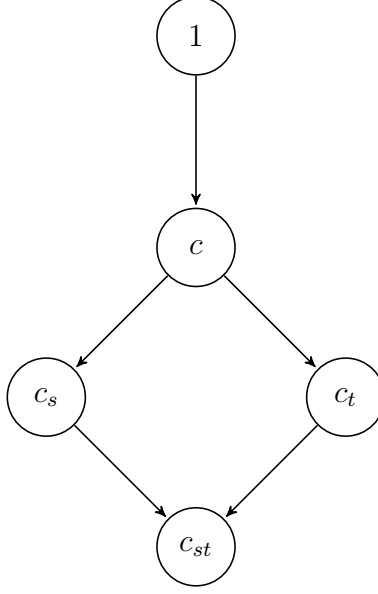
Figure 1: Hierarchy of hypotheses.

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, using coordinate descent [Luo and Tseng, 1992], by iteratively solving partial derivatives of $l$, with respect to $c_{st}$ and $\gamma_{st}$, set equal to zero

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

## 2.3 Unobserved Counts

Many times it may not be possible to count the number of individuals of each prey species that are in a predator's gut. Instead, it may only be possible to detect whether or not a predator consumed the prey species during a given period. In this case we can still make inference about the predators' preferences for the different prey species by using the EM algorithm to compute maximum likelihood estimates.

We denote the binary random variable indicating that the $j^{th}$ predator did in fact eat at least one individual of prey species $s$ in time period $t$ by $Z_{jst} = 1(X_{jst} > 0)$. The variables are independent Bernoulli observations with success probability $p_{st} = P(Z_{jst} = 1) = 1 - \exp\{-\lambda_{st}\}$. Despite not observing $X_{jst}$, we can compute maximum likelihood estimates of the parameters $\boldsymbol{\lambda}, \boldsymbol{\gamma}$ through the EM algorithm using the complete data log-likelihood

$$l_{comp}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \log f_{X,Y,Z}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} | \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \sum_{s=1}^{S} \sum_{t=1}^{T} \left[ \sum_{j=1}^{J_t} \log f_{X,Z}(x_{jst}, z_{jst} | \boldsymbol{\lambda}) + \sum_{i=1}^{I_t} \log f_Y(y_{jst} | \boldsymbol{\gamma}) \right].$$

The density of $Y_{jst}$ is exactly as in section 2.2 and so we focus on deriving the joint density of $X_{jst}$ and $Z_{jst}$. With the distribution of $Z_{jst}$ given above, we can compute $f_{X,Z}(x_{jst}, z_{jst} | \boldsymbol{\lambda})$

by noting that $X_{jst} = 0$ with probability 1 if $Z_{jst} = 0$, and that $[X_{jst}|Z_{jst} = 0]$ has a truncated Poisson distribution with density

$$f_{X|Y,Z,\boldsymbol{\lambda},\boldsymbol{\gamma}}(x_{jst}|z_{jst}) = \frac{\exp\{-\lambda_{st}\}\lambda_{st}^{x_{jst}}}{(1 - \exp\{-\lambda_{st}\})x_{jst}!}1(x_{jst} > 0)$$

and expected value

$$\mathbb{E}_{X|Y,Z}X_{jst} = \frac{\lambda_{st}\exp\{\lambda_{st}\}}{\exp\{\lambda_{st}\} - 1}.$$

The joint density of $X_{jst}, Z_{jst}$ is then

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

The EM algorithm works by iterating two steps, the E-step and M-step, until the optimum is reached [Dempster et al., 1977, McLachlan and Krishnan, 2007]. Let $k$ index the iterations in the EM algorithm so that $\boldsymbol{\lambda}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ denote the estimates computed on the $k^{th}$ M-step. The E-step consists of computing the expectation of $l_{comp}$ with respect to the conditional distribution of $X$ given the current estimates of the parameters

$$Q^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \mathbb{E}_{X|Y,Z,\boldsymbol{\lambda}^{(k)}}l_{comp}$$

in order to remove the unobserved data. The M-step then involves maximizing $Q = \mathbb{E}l_{comp}$ with respect to the parameters in the model to obtain updated estimates of the parameters,

$$(\boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}) = \arg\max_{(\boldsymbol{\lambda}, \boldsymbol{\gamma})} Q^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\gamma}).$$

These steps are then alternated until a convergence criterion monitoring subsequent differences in the parameter estimates/likelihood is met.

The calculation of $Q^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ is not difficult and is given by:

$$
\begin{aligned}
Q^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) &= \mathbb{E}\log f_{X,Z|\boldsymbol{\lambda}}(X_{jst}, z_{jst}) + \log f_{Y|\boldsymbol{\gamma}}(y_{ist}) \\
&= \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\mathbb{E}\log f_{X,Z|\boldsymbol{\lambda}}(X_{jst}, z_{jst}) + \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{i=1}^{I_t}\log f_{Y|\boldsymbol{\gamma}}(y) \\
&\propto \sum_{s,t,j}\left(-\lambda_{st} + z_{jst}\log\lambda_{st}\mathbb{E}X_{jst}\right) + \sum_{s,t}\left(-I_t\gamma_{st} + Y_{\cdot st}\log I_t\gamma_{st}\right) \\
&\propto \sum_{s,t}\left(-J_t\lambda_{st} + z_{\cdot st}\log\lambda_{st}\mathbb{E}(X_{jst}|\lambda_{st}^{(k)}, \gamma_{st}^{(k)})\right) + \sum_{s,t}\left(-I_t\gamma_{st} + Y_{\cdot st}\log I_t\gamma_{st}\right).
\end{aligned}
\tag{2}
$$

No analytic solution to the M-step exists, however, so we chose to maximize $Q$ with coordinate descent [Luo and Tseng, 1992]. In fact, as we only need find parameters that increase the value of $Q$ on each iteration, we forgo fully iterating to find the maximum and instead perform just one step uphill within each EM iteration. Since $Q^{(k)}$ is concave and

5

smooth in the parameters $\boldsymbol{\lambda}, \boldsymbol{\gamma}$, we are able to use the convergence of parameter estimates, $||(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\gamma}^{(k)}) - (\boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)})||_\infty < \tau$, for some $\tau > 0$, as our stopping criterion.

As we show in our simulation study, this generalized EM algorithm accurately estimates the parameters when values of $\lambda_{st}$ 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 $\lambda_{st}$ in the equation $1 = \mathbb{E}Z_{jst}1 - \exp\{-\lambda_{st}\}$. As the proportion of ones in the observed data increases, we expect $\lambda_{st}$ to grow exponentially large. When no zeros are present in the data, so that where only ones are observed, the likelihood can be made arbitrarily large by sending the parameter off to infinity.

## 2.4  Testing

The likelihood ratio test statistic is

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

where $\theta_0, \theta_1$ represent the parameters estimated under the null and alternative hypotheses, respectively. It is well known that the asymptotic distribution of $\Lambda$ is a $\chi_\rho^2$ distribution with $\rho$ degrees of freedom [Wilks, 1938]. When the observations $X_{jst}$ are not observed, we use $L_{obs}(Z, Y)$ as the likelihood in the calculation of $\Lambda$.

The degrees of freedom $\rho$ equal 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 $\chi_\rho^2$ distribution. Hence, with a level of significance, $\alpha$, the null hypothesis is rejected in favor of the alternative hypothesis if $\mathbb{P}(\chi_\rho^2 > \Lambda) < \alpha$.

## 2.5  Linear Transformations of $c_{st}$

After determining which model best fits the data, more detail can be extracted through a hypothesis test of 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 maximum likelihood estimates, $\hat{c}_{st}$, as found via the framework above. 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 distributed as $\mathcal{N}(a^t\mathbf{c}, a^t\Sigma a)$, where $\Sigma$ is the covariance matrix of the asymptotic distribution of $\hat{\mathbf{c}}$.

Suppose, for example, that the hypothesis $c_s$ is determined to best fit the data with $s$ ranging $s = 1, 2, 3$. We can test to see whether or not two species are statistically equally preferred under the null hypothesis $c_1 = c_2$. This hypothesis is alternatively written in vector notation as $a^t\mathbf{c} = 0$, where $a = (1, -1, 0)^t$. Tests of the following form $H_0 : a^t\mathbf{c} = \mu$ against any alternative of interest are then approximate $Z$-tests. Confidence intervals of any size are

178 similarly, readily obtained. Of course, one could also use a $t$ distribution as a small sample
179 size correction.

# 3   Simulations

181 Our simulations assume two prey species and five time points, throughout. Of the hierarchy
182 of hypotheses, we generate data under three models: $c, c_s, c_t$. Sample sizes for both prey
183 species and predator gut count observations are randomly chosen from four overlapping
184 levels: "small" sample sizes are randomly sampled numbers in $[20, 50]$, "medium" $[30, 75]$,
185 "large" $[50, 150]$, and "huge" $[100, 200]$. This is repeated for each level of sample size. We
186 simulate 500 replicate datasets for each of the twelve scenarios above for both types of data,
187 fully observed count data, $X_{jst}$, and for non-count data, when we observe only a binary
188 response, $Z_{jst} = 1(X_{jst} > 0)$. Each scenario is then fit with the true model that generated the
189 data. All simulations of non-count data use $\tau = 10^{-5}$ as the convergence tolerance. A subset
190 of the examples are provided here; the interested reader is referred to the supplementary
191 materials for the complete simulation results. For the simulations we used the R Core Team
192 [2014] package `BatchExperiments` by Bischl et al. [2014].
193     For all simulated data, the true parameter values for the rate at which prey species are
194 encountered in the wild are fixed to be $\gamma_{st} = \pi$, $\forall s, t$. The values of $\lambda_{st}$ are set with respect
195 to each data generating model. For model $c_{st} = c$, where predator preferences don't vary
196 by either time or species, we put $\lambda_{st} = 2\pi, \forall s, t$. Under model $c_s$, the ratio of rates vary by
197 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
198 last model, $c_t$, the ratio of rates vary by time $t$. Here, we put $\lambda_{st} = t$ for $t \in \{1, \ldots, 5\}$.
199     Figure 2 shows density plots of the estimates of $c_s, c$ when fitting the true model to the
200 fully observed count data generated under models $c_s$ and $c$. The plots provide evaluations
201 of parameter estimates under each scenario. In the first display, the parameters $c_1 \approx 0.45$
202 and $c_2 = 1$ are on average, across all 500 simulations, estimated as $\hat{c}_1 = 0.45$ and $\hat{c}_2 = 1.00$,
203 with standard errors of $\text{se}(\hat{c}_1) = 0.03$ and $\text{se}(\hat{c}_2) = 0.06$. The second display provides results
204 for model $c_{st} = c$. Averaging across all 500 simulations, the parameter $c = 2$ is estimated
205 as $\hat{c} = 2.00$. This is further seen in figure 3, where box plots of the parameter estimates,
206 centered at true parameter values, of the correct model fit to data generated from both $c_s$
207 and $c_t$ show empirically very little bias.
208     We next generated data with unobserved counts. As noted above under certain circum-
209 stances our unobserved counts model accurately estimates the parameters of interest, and
210 at other times can infinitely over-estimate parameters. To investigate this issue further, we
211 consider the same scenarios mentioned above, but now we reduce all of the count data down
212 to binary observations. For each scenario, we fit the unobserved counts model as if we knew
213 the true underlying model that generated the observed data.
214     Figure 4 contains density plots for all 500 replications of the data generating models $c_s, c_t$
215 with small and huge sample sizes, respectively. When data are generated under the model
216 $c_s$ and the true model fit to the non-count data, we find even for the small sample size that
217 point estimates are only very slightly biased. When parameter values are of sufficient size
218 to make zeros in the simulated data less common, the estimates from fitting the correct
219 model to the generated data are occasionally over-estimated. This effect is easily seen in the
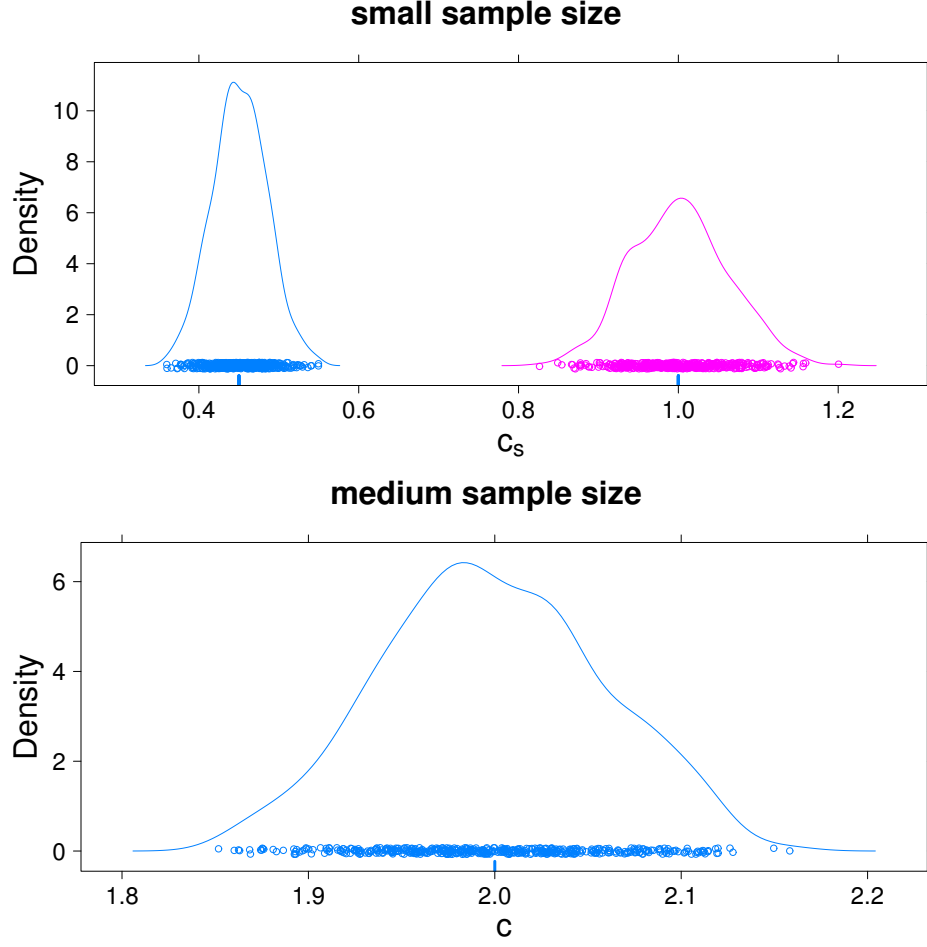
7

Figure 2: Density plots of all 500 estimates of fitting the true model to the data generated from models $c_s, c$ are shown with sample sizes small and medium, respectively.
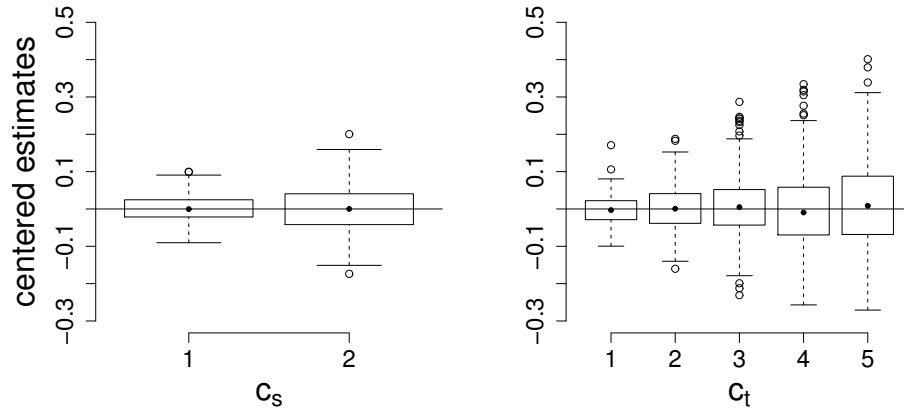


Figure 3: Shown are all 500 estimates, centered at the true parameter values, from fitting the true model to the data generated from models $c_s, c_t$ with sample sizes small and medium, respectively.
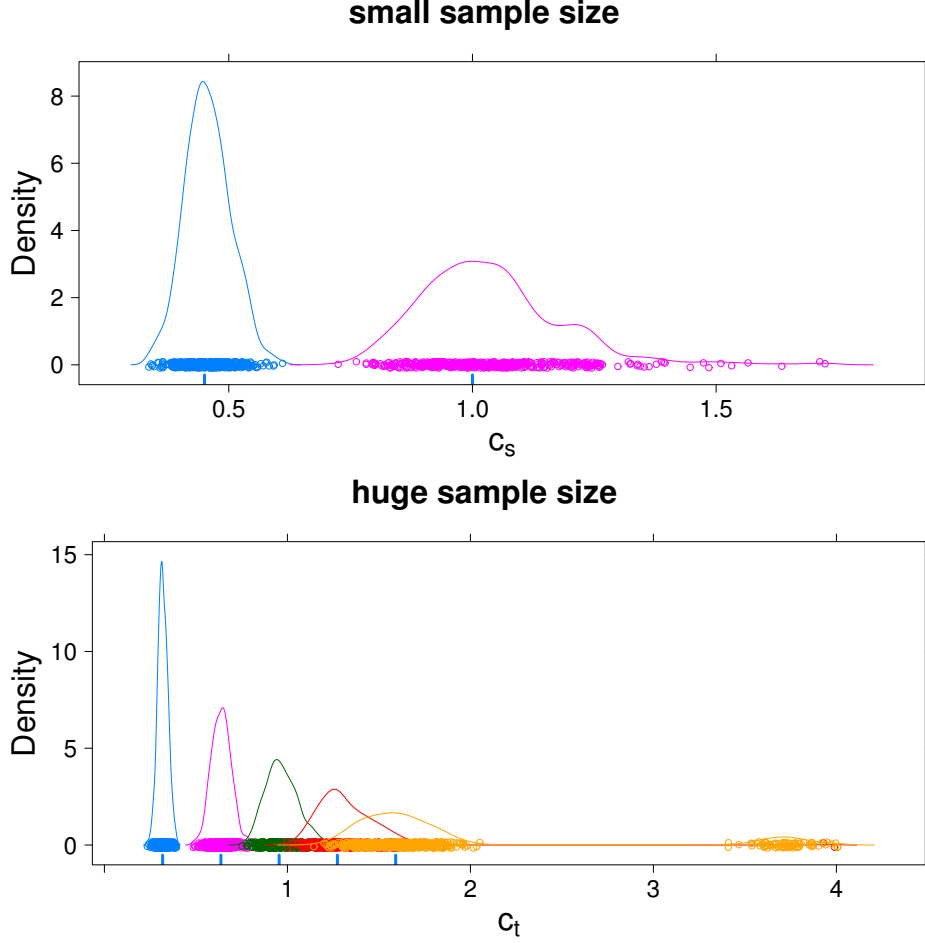
Figure 4: Density plots of all 500 estimates of fitting the true model to the data generated from models $c_s, c_t$, when counts are not observed, are shown with sample sizes small and huge, respectively.

bottom panel of figure 4 for larger values of $c_t$ despite the increased sample size, but is also seen, less dramatically, in the density plot for the $c_s$ generated data.

The cluster of estimates for $c_5$ between 3.5 and 4.0 in the bottom panel of figure 4 comes from datasets in which $X_{js5} > 0$ so that $Z_{js5} = 1$ for all $j, s$. For the data shown in figure 4, this happened 73 times out of the 500 replicated datasets. As mentioned above, the estimate of $c_5$ is infinite in this case. However, the EM algorithm will always provide a finite estimate for all parameters when it terminates. In this case, we set $\tau = 10^{-5}$ and it just happened that this caused the algorithm to terminate with $\hat{c}_5$ between 3.5 and 4.0. To confirm that this is due to the arbitrary choice of $\tau$, we repeated the algorithm with smaller values of $\tau$ for several datasets. As expected, $\hat{c}_5$ increased without bound as we refit the model with increasingly small values of $\tau$.

The over-estimation of parameters, a symptom of the loss of information due to the unobserved counts, can also be seen with box plots of the 500 point estimates centered at their respective true parameter values. Figure 5 contains box plots of the same scenarios in
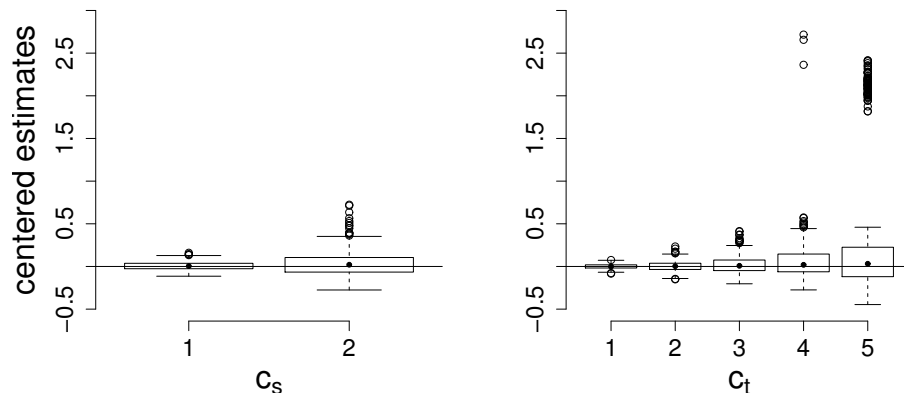
9

Figure 5: Shown are all 500 estimates, centered at the true parameter values, from fitting the true model to the data generated from models $c_s, c_t$, when counts are not observed, with sample sizes small and huge, respectively.

Figure 4. For the 73 cases in which $Z_{js5} = 1$ for all $j, s$ under model $c_t$ with the huge sample size, the bias is infinite since parameter estimates will, theoretically, be infinite. The finite bias shown in these plots is due to the finite estimates provided by the termination of the EM algorithm. Thus, conditional on a mixture of 0s and 1s in the data the corresponding estimators are unbiased, but when no 0s exist in the data the theoretical bias is infinite.

# 4    Real Data

We analyzed a dataset that was collected to investigate the eating preferences of the wolf spider, genus *Schizocosa*, towards the prey orders *Diptera* and *Collembola*. Predators were collected in traps within the deciduous Berea College Forest in Madison County, Kentucky, USA and had their gut-contents analyzed to determine whether or not the spiders ate any *Diptera* or *Collembola* within each time period. The left panel of figure 6 plots the percentage of spiders that had the prey in their guts across time. Prey were also collected in traps from the same area. These data were collected for each month from October 2011 to March 2013. On average, 69 spiders, and 111 and 297 *Diptera* and *Collembola*, respectively, were caught in each time period. The range of the sample sizes, across all 18 months is, 11 to 181 for caught spiders, from 7 to 322 for trapped *Diptera*, and from 101 to 755 for trapped *Collembola*. The right panel of figure 6 plots the total number of each order that was caught during each time period.

These data provide an example of our hierarchy of hypotheses. First, we tested model $c_{st} = c$ against $c_{st} = c_s$, to determine whether or not the wolf spider has different preferences for the two orders *Diptera* and *Collembola*. With, one degree of freedom, this likelihood ratio test indicated, $p - value < 0.0001$, that two parameters, one for each order, fits these data better than one parameter for both. Similarly, we tested whether or not there was a significant effect across time by testing model $c_{st} = c$ against $c_{st} = c_t$. Here, the likelihood
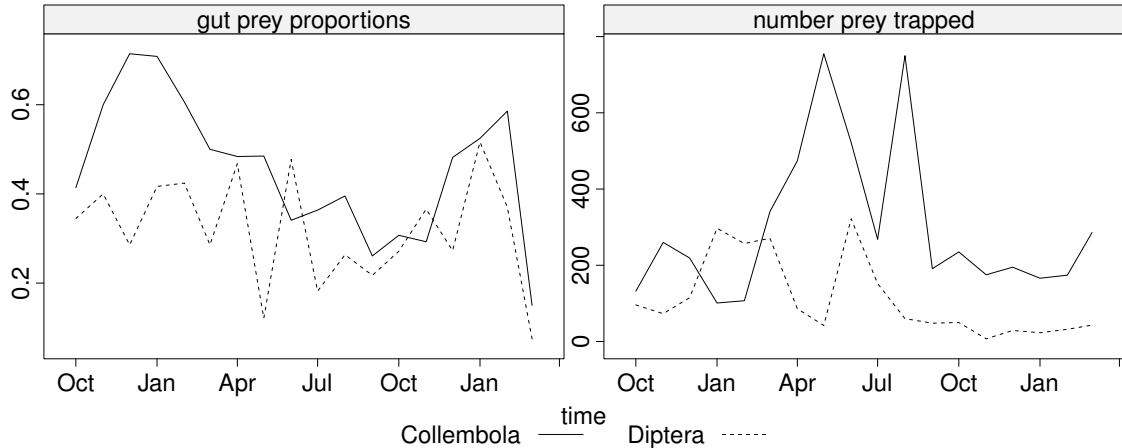
Figure 6: For both *Diptera* and *Collembola*, the plots show the proportion of spiders with these orders found in their guts, and the number of these orders trapped in each time period.

ratio test, with 17 degrees of freedom, implies that the wolf spiders of the Berea College Forest eat these prey orders at different rates across the months of the year, $p-value < 0.0001$. In fact, we find that the most parameter rich model, $\lambda_{st} = c_{st}\gamma_{st}$ fits these data better than is expected by chance, $p-value < 0.0001$. Model $c_{st}$ estimates 72 parameters in total; since, in this case, there are two prey of interest and 18 time periods, it takes 36 parameters to estimate each $c_{st}$ and $\gamma_{st}$. Figure 7 plots the point estimates and 95% confidence intervals of $c_{st}$, for both prey across all time periods.

With point estimates of $c_{st}$ under the model $\lambda_{st} = c_{st}\gamma_{st}$, we can test any number of linear contrasts. For instance, $c_{1t} = c_{2t}$, for $t \in \{1, \ldots, 18\}$. Using a level of significance of 0.05, and after making a Bonferroni multiple comparisons adjustment, the data can not say that the two prey are differently preferred in October, November, and December of 2011 and for March and July of 2012.

# 5 Discussion

The model developed here allows practitioners to determine predators' preferences by testing simultaneously across an array of multiple prey species and time. This is achieved 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 across 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 relative preferences for different species.

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 modeling 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
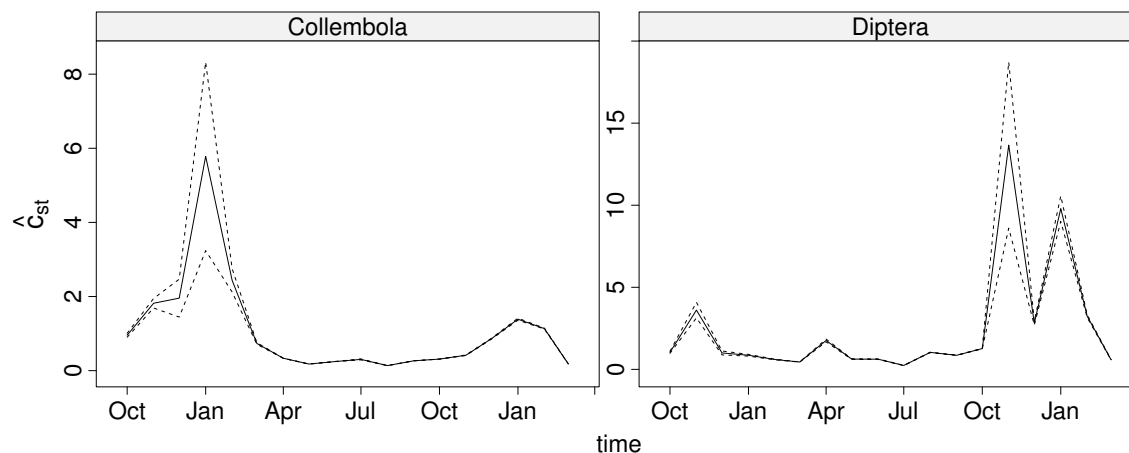
11

Figure 7: One plot for each prey order displays the point estimates and 95% confidence intervals as estimated from the model $c_{st}$.

this is nice in theory, in practice the success of this modification to our original model is limited by the magnitude of the unknown parameters $\lambda_{st}$.

Further developments of our model could be beneficial. Taking into account other environmental variables that might effect a predator's 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

Bernd Bischl, Michel Lang, and Olaf Mersmann. *BatchExperiments: Statistical experiments on batch computing clusters.*, 2014. URL `http://CRAN.R-project.org/package=BatchExperiments`. R package version 1.3.

Jean Chesson. Measuring preference in selective predation. *Ecology*, pages 211–215, 1978.

Jean Chesson. The estimation and analysis of preference and its relatioship to foraging models. *Ecology*, pages 1297–1304, 1983.

Hayley S Clements, Craig J Tambling, Matt W Hayward, and Graham IH Kerley. An objective approach to determining the weight ranges of prey preferred by and accessible to the five large african carnivores. *PloS one*, 9(7):e101054, 2014.

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/`.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

Adam G Hansen and David A Beauchamp. Effects of prey abundance, distribution, visual contrast and morphology on selection by a pelagic piscivore. *Freshwater Biology*, 59(11): 2328–2341, 2014.

Peter Hellström, Jesper Nyström, and Anders Angerbjörn. Functional responses of the rough-legged buzzard in a multi-prey system. *Oecologia*, 174(4):1241–1254, 2014.

Viktor Sergeevich Ivlev. Experimental ecology of the feeding of fishes. 1964.

Jürgen Jacobs. Quantitative measurement of food selection. *Oecologia*, 14(4):413–417, 1974.

Martin J Lechowicz. The sampling characteristics of electivity indices. *Oecologia*, 52(1): 22–30, 1982.

Zhi-Quan Luo and Paul Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72 (1):7–35, 1992.

Salvador Lyngdoh, Shivam Shrotriya, Surendra P Goyal, Hayley Clements, Matthew W Hayward, and Bilal Habib. Prey preferences of the snow leopard (panthera uncia): Regional diet specificity holds global significance for conservation. *PloS one*, 9(2):e88349, 2014.

Hawis H Madduppa, Neviaty P Zamani, Beginer Subhan, Unggul Aktani, and Sebastian CA Ferse. Feeding behavior and diet of the eight-banded butterflyfish chaetodon octofasciatus in the thousand islands, indonesia. *Environmental Biology of Fishes*, pages 1–13, 2014.

Bryan FJ Manly, Lyman L McDonald, Dana L Thomas, Trent L McDonald, and Wallace P Erickson. *Resource selection by animals*. Springer, 1992.

Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

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.

HA Vanderploeg and D Scavia. Two electivity indices for feeding with special reference to zooplankton grazing. *Journal of the Fisheries Board of Canada*, 36(4):362–365, 1979.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.