# Modeling Predator Preferences

Edward A. Roualdes and Simon Bonner
University of Kentucky

2014-10-13

**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 is provided to demonstrate the accuracy of our methods.

# 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 rarely obtained practical asymptotic properties. This leaves 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 multiple 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: this predator ate some of that prey species during this time period, or did not. Despite the data's lost information, under certain situations our model is able to accurately estimate the parameters of interest.

# 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 indepedent of each other.

We denote the number of predators and the numer 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. $c_{st} = 1, s = 1, \ldots, S; t = 1, \ldots, T$
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. $c_{st} = c_{st}, s = 1, \ldots, S; t = 1, \ldots, T$

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 indifferene. 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.

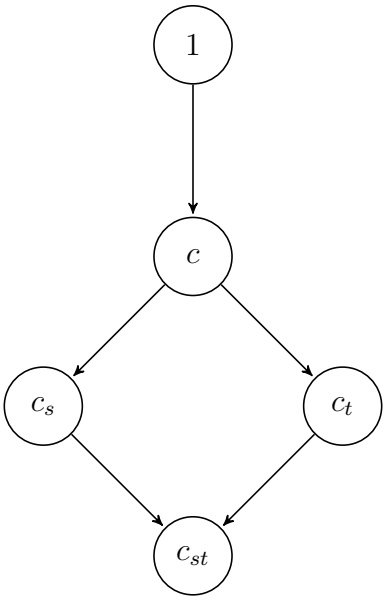


Figure 1: Hierarchy of hypotheses.

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

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 $Y_{\cdot st} = \sum_{i=1}^{I_t} Y_{ist}$.

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

Working with biologists who study spider foraging, 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 these incomplete data, model predators' eating preferences with the above framework using the EM algorithm.

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 $p_{st} = 1 - \exp\{-\lambda_{st}\}$. Despite not observing $X_{jst}$, the EM algorithm is able to find maximum likelihood estimates of the parameters $\boldsymbol{\lambda}, \boldsymbol{\gamma}$ using the complete data log-likelihood, $l_{comp}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \log f_{X,Y,Z}(x_{jst}, y_{ist}, z_{jst}|\boldsymbol{\lambda}, \boldsymbol{\gamma})$.

With the distribution of $Z_{jst}$ in hand, we calculate 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 EM algorithm works in two steps [Dempster et al., 1977, McLachlan and Krishnan, 2007]. To drop dependence on the unobserved data, the E-step consists of taking the expectation of $l_{comp}$ with respect to the conditional distribution of $X$ given the observed data and a current estimate of the parameters $\boldsymbol{\lambda}, \boldsymbol{\gamma}$, namely $f_{X|Y,Z,\boldsymbol{\lambda}}(x_{jst})$. The current estimate of the parameters are used within the calculation of $\mathbb{E}_{X|Y,Z,\boldsymbol{\lambda}}X_{jst}$ and not within the distribution functions of the underlying model. The M-step then maximizes $Q = \mathbb{E}l_{comp}$ with respect to the parameters in the model. The E-step and M-step are then iteratively updated until a convergence criterion is met.

Denoting iterations and current estimates of parameters of the EM algorithm with a superscript $(k)$, the two EM steps are E-step: $Q^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \mathbb{E}_{X|Y,Z,\boldsymbol{\lambda}^{(k)}} l_{comp}$, and M-step: $(\boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}) = \arg\max_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} Q^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$. The calculation of $Q^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ is not difficult and is given in equation 2.

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

In this case, no analytic solution to the M-step exists so we choose to maximize $Q$ with coordinate descent [Luo and Tseng, 1992]. In fact, as we only need find parameters that

increase the value of $Q$, 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 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.

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 $\mathbb{E}Z_{jst} = 1 = 1 - \exp\{-\lambda_{st}\}$. As the proportion of ones in the observed data increase, we expect $\lambda_{st}$ to grow exponentially large. When no zeros are present in the data, where only ones are observed, the maximum likelihood estimate can be made arbitrarily large by sending the parameter off to infinity.

## 2.4 Testing

The likelihood ratio test (LRT) 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 $\chi^2_\rho$ 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 $\chi^2_\rho$ distribution. Hence, with a level of significance, $\alpha$, 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 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 distrbuted as $\mathcal{N}(a^t\mathbf{c}, a^t\Sigma a)$, where $\Sigma$ 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 $s = 1, 2, 3$. 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 similarly, readily obtained.

# 3 Simulations

Our simulations assume two prey species and five time points, throughout. Of the hierarchy of hypotheses, we generate data under three models: $c, c_s, c_t$. 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 generate data for all models. 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, $Z_{jst} = 1(X_{jst} > 0)$. Each scenario is then fit with the true model that generated the data. A subset of the examples are provided here; the interested reader is referred to the supplementary materials for the complete set of simulation results.

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 $\lambda_{st}$ are set with respect to each data generating model. For the model $c_{st} = c$, where predator preferences don't vary by either time or species, we consider $\lambda_{st} = 2\pi, \forall s, t$. Under the model $c_s$, 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 last model, $c_t$, the ratio of rates vary by time $t$. Here, we put $\lambda_{st} = t$ for $t \in \{1, \ldots, 5\}$.

Figure 2 shows density plots of the estimates of $c_s, c$ when fitting the true model to the fully observed count data generated under models $c_s$ and $c$. The plots provide evaluations of parameter bias under each scenario. In the first display, the parameters $c_1 \approx 0.45$ and $c_2 = 1$ are on average, across all 500 simulations, estimated as $\hat{c}_1 = 0.45$ and $\hat{c}_2 = 1.00$, with standard errors of $\text{se}(\hat{c}_1) = 0.03$ and $\text{se}(\hat{c}_2) = 0.06$. The second display provides a similar conclusion. Averaging across all 500 simulations, the parameter $c = 2$ is estimated as $\hat{c} = 2.00$.

Though only a subset of the results are shown, when count data are fully observed, simulations show our model to have very little bias. This is further seen in figure 3, where box plots of the parameter estimates, centered at true parameter values, of the correct model fit to data generated from both $c_s$ and $c_t$ show empirically almost no bias.

When we generate data with unobserved count data, the story is slightly different. As noted above under certain circumstatnces our EM model accurately estimates the parameters of interest, and at different times can greatly over-estimate the parameters. To investigate this issue further, we consider the same scenarios mentioned above, but now we reduce all of our count data down to binary observations. For each scenario's generated data, we fit our EM algorithm as if we knew a priori the true underlying model that generated the observed data.

Figure 4 contains density plots of fitting the true EM model, for all 500 replications, of the data generating models $c_s, c_t$ with small and huge sample sizes, respectively. When

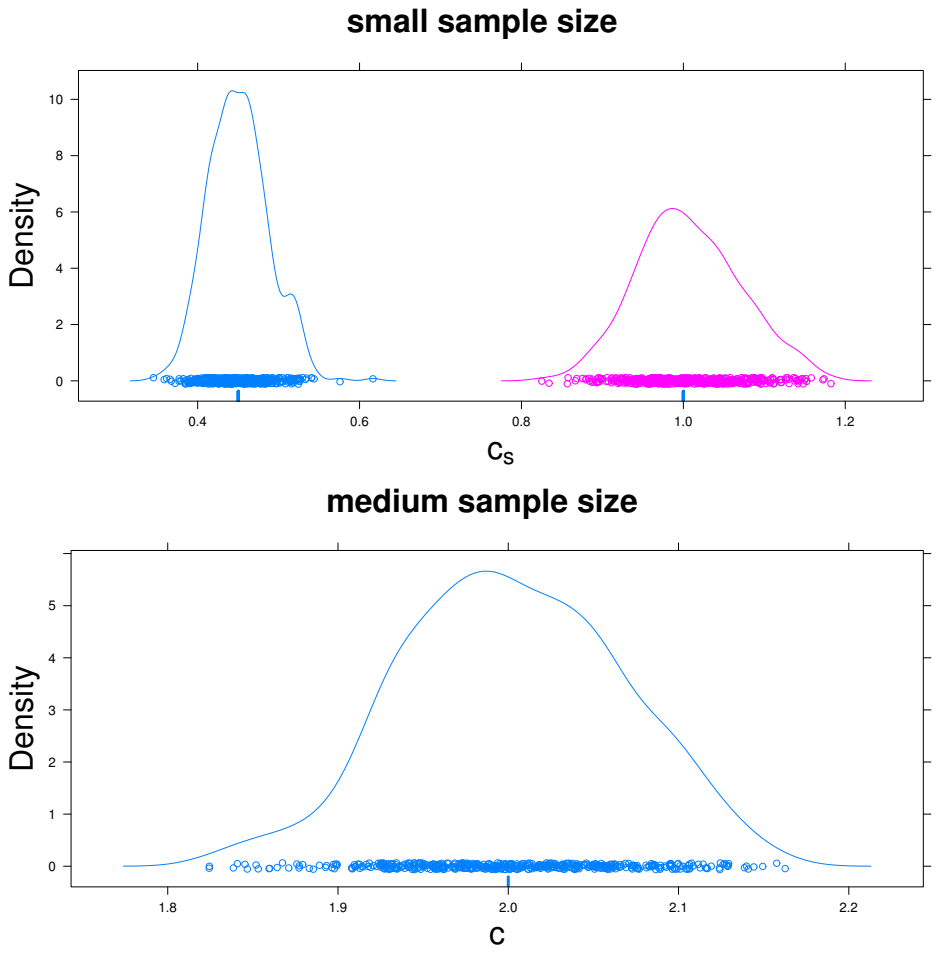


Figure 2: Density plots of all 500 estimates of fitting the true model to the data generated from models $c_s, c$ are shown with sample sizes small and medium, repsepectively.

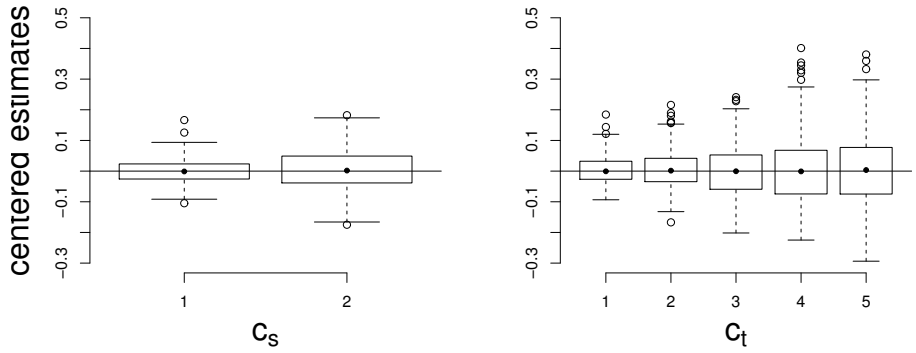


Figure 3: Shown are all 500 estimates, centered at the true parameter values, from fitting the true model to the data generated from models $c_s, c$ with sample sizes small and medium, repsepectively.

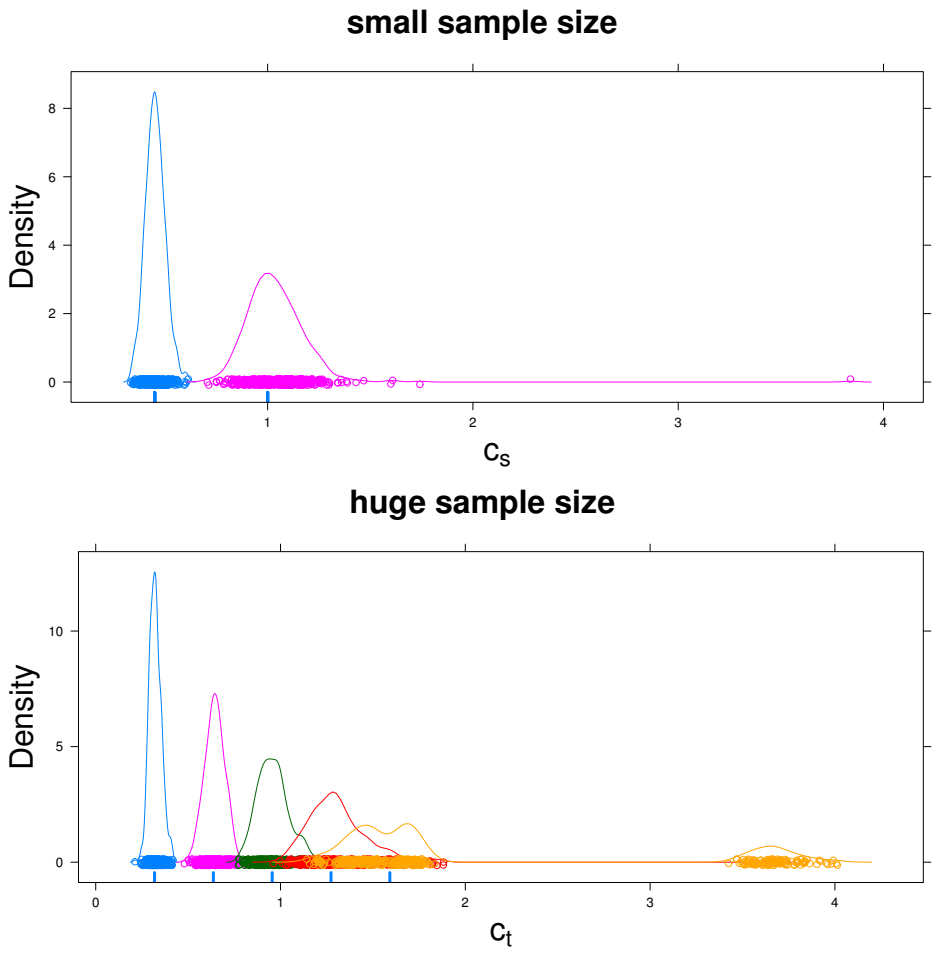


Figure 4: Density plots of all 500 estimates of fitting the true model to the data generated from models $c_s, c_t$ are shown with sample sizes small and huge, repsepectively.

data are generated under the model $c_s$ and the true model fit to the data, we find even for the small sample size that point estimates are reasonably accurate. When parameter values are of sufficient size to make zeros in the simulated data less common, the estimates from fitting the correct model to the generated data are occassionally over-estimated. This effect is easily seen in 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 over-estimation of parameters, a symptom of the loss of information due to the unobserved count data, 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 Figure 4, but with the small and huge sample sizes. It is clear that as the parameters values increase and zeros in the observed data $Z_{jst}$ become less prevalent, over-estimation occurs more frequently. Because of the tendency for the EM algorithm to over-estimate some parameter values, bias for non-count data is worse than it is for the count data.

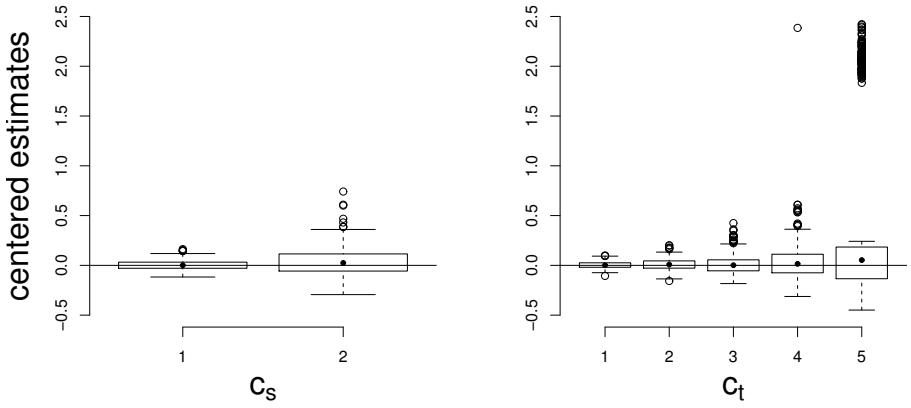


Figure 5: Shown are all 500 estimates, centered at the true parameter values, from fitting the true model to the data generated from models $c_s, c_t$ are shown with sample sizes small and huge, repsepectively.

# 4 Real Data

Interest in the predator genus Schizocosa and the prey orders? Diptera, Collembola, Ensifera, for time periods October 2011 to March 2013.

# 5 Discussion

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 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 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 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 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

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.

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.

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.

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.