Noname manuscript No.
(will be inserted by the editor)

Formal Modelling of Predator Preferences using Molecular Gut-Content Analysis

Edward A. Roualdes · Simon J. Bonner · Thomas D. Whitney · James D. Harwood

the date of receipt and acceptance should be inserted later

Abstract The literature on modelling a predator's prey selection describes 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 data from multiple species and time points. Further, we apply the expectation-maximisation algorithm to compute estimates if exact counts of the number of prey species eaten in a particular time period are not observed. We conduct a simulation study to demonstrate the accuracy of our method, and illustrate the utility of the approach for field analysis of predation using a real data set, collected on wolf spiders using molecular gut-content analysis.

Keywords electivity · expectation-maximization · predator-prey interactions · generalist predators · food web analysis

1 Introduction

The indices most commonly used to describe 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 many applied papers continue to use them. A quick search of papers published in 2014 returns hundreds of publications that cite these fundamental papers, a few being 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 the statistical rigour of a full model, focus on a snapshot in time, and rarely allow more than one prey species to be considered (Lechowicz, 1982). Other authors are using the Monte Carlo based method of Agustí et al (2003) (see Davey et al, 2013; King et al, 2010), but this is also unable to take into account multiple prey species across multiple time points. We propose an intuitive statistical model to estimate and statistically test differences in a predator's prey preferences across an array of time points and between multiple prey species.

A comprehensive overview by Lechowicz (1982), later summarised by Manly et al (2002), details some of the benefits and faults of the most popular indices. According to these reviews, a majority of the indices give comparable results, save Strauss's linear index L, despite the fact that most of the methods differ by range and linearity of response. While Lechowicz recommends one index, E^* by Vanderploeg and Scavia (1979) as the "single best" (Lechowicz, 1982), albeit imperfect, index, Manly et al. instead take the approach of excluding the subset of indices which do not "estimate any biologically meaningful value" (Manly et al, 2002). Lechowicz (1982) recommends the index E^*, an element of the Manly et al (2002) suggested indices, because the index value 0 denotes random feeding, the index has a range restricted

E.A. Roualdes · S. J. Bonner
Department of Statistics, University of Kentucky, Rm. 311 Multidisciplinary Science Building, 725 Rose Street, Lexington, KY 40536-0082,
E-mail: edward.roualdes@uky.edu

T.D. Whitney
Department of Entomology, University of Kentucky, Lexington, KY 40546, USA *Present address: Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA 30602, USA*

J.D. Harwood
Department of Entomology, University of Kentucky, Lexington, KY 40546, USA

to $[-1, 1]$ (though $E^* = 1$ is nigh impossible), and 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 number of available prey types (assumed known). The downside to this index is its lack of reasonable statistical properties (Lechowicz, 1982), thus making the computation of standard errors and thus hypothesis testing difficult. This is, in fact, a common fault amongst most of the indices.

Manly et al (2002) recognized the need for more formal statistical inference and proposed the use of generalised linear models (GLM). The well established literature on GLMs allows for formal hypothesis testing. Much of this work however focuses on general resource selection and doesn't directly address the problems of food selection. Our model focuses specifically on the problem of predators' food selection, offers formal inference and hypothesis testing similar to that of GLMs Manly et al (2002), and provides meaningful single number summaries of the predators' food preferences. The model we are proposing 1) maintains the intuitiveness of the indices summarised by Lechowicz (1982); Manly et al (2002), 2) offers formal statistical justification similar to GLMs, 3) accommodates data collected from multiple prey species across multiple time points, and 4) provides parameters estimates even when exact counts of prey species eaten are not fully observed.

Our model assumes that numbers of each prey species captured and the numbers consumed by an individual predator follow independent Poisson distributions with separate rates that might possibly vary between species and over time. We are able to estimate the parameters of interest even when exact counts of each prey species eaten within any given time period are not observed. When exact counts are unobserved, we instead rely on the researcher being able to detect prey DNA within the predator's gut (Schmidt et al, 2014; Raso et al, 2014; Madduppa et al, 2014) and make a simple binary conclusion: this predator ate some of that prey species during this time period, or did not.

This paper is organised as follows. Section 2 describes our statistical model, for both fully observed count data and for the non-observed count data for which we use the expectation-maximisation (EM) algorithm to estimate parameters of interest, and the statistical tests used to make statements about the parameters of interest. In Section 3, we offer a simulation study that demonstrates the accuracy of our methods. Section 4 provides a real data set, which investigates the eating preferences of wolf spiders (Araneae: Lycosidae), found in the Berea College Forest in Madison County, Kentucky, USA, to demonstrate how those interested in assessing trophic interactions with gut-content analyses could apply our methods. A brief discussion concludes the paper in Section 5. Alongside our model, we offer an `R` (Core Team, 2014) package named `spiders` that implements the methods discussed.

2 Methods

We assume that samples of both the predators and prey are captured from the study area on T occasions. Depending on the species involved and the design of the study the predators and prey may be sampled in the same way or using different methods. In the spider experiment described in Section 4, for example, prey are captured using pitfall traps that are dispersed throughout the study area whereas the predators (spiders) were captured by hand. Predators and prey species are collected and counted at each time period. We denote the number of predators and the number of prey species caught in each time period $t \in \{1, \ldots, T\}$ by J_t and I_t, respectively. Prey species will be indexed by $s \in \{1, \ldots, S\}$. Let X_{jst} represent the number of prey species s that predator j ate during period t, where $j \in \{1, \ldots, J_t\}$. Let Y_{ist} represent the number of prey species s found in trap i during period t, $i \in \{1, \ldots, I_t\}$.

The number of prey species s that predator j ate during period t is assumed to follow a Poisson distribution with rate parameter λ_{st}, $X_{jst} \overset{\text{iid}}{\sim} \mathcal{P}(\lambda_{st})$. The parameter λ_{st} represents the rate at which the predator ate prey species s during time period t. The number of prey species s found in trap i during period t is assumed to follow a Poisson distribution with rate parameter γ_{st}. The use of Poisson distributions make the following implicit assumptions: 1) traps independently catch the prey species of interest, 2) predators eat independently of each other.

By modelling $\lambda_{st}, \gamma_{st}$ we are able to test claims about a predator's eating preferences. Formal statistical statements about the relative magnitudes of the parameters $\boldsymbol{\lambda} = (\lambda_{11}, \ldots, \lambda_{ST})^t$ and $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{ST})^t$ offer insights to the relative rates at which predators eat particular prey species. We consider five variations on the relative magnitude of $c_{st} = \lambda_{st}/\gamma_{st}$. These five hypotheses each allow c_{st} to vary by time, prey species, both, or neither.

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 the relative rate of sampling for the predator and the traps is the same for all species on all occasions. One imagines this is the case if the prey move randomly and the predator simply eats prey which comes within its reach, thus suggesting no selection for a particular prey item. The second states that predators sample prey proportionally across all time periods. The third hypothesis states 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 fourth hypothesis implies that each prey species is sampled similarly within each time period, while the rates across time are allowed to change. The fifth hypothesis assumes a predator's selection varies by both time and prey species. This would make sense if environmental and biological variables, such as weather, prey availability, and/or palatability were affecting predators' selection strategies.

[Fig. 1 about here.]

Because the five hypotheses are nested, a natural testing order is suggested in Figure 1. For instance, suppose interest lies in the predator's dietary preferences with respect to prey species s_1 and s_2 across three time periods. The hypothesis that the predator has the same preference for the two prey species can be tested with

$$H_0 : \lambda_t = c_t\gamma_t, \forall t$$
$$H_1 : \lambda_{st} = c_{st}\gamma_{st}.$$

The alternative here allows for the predators preferences to vary independently for the two species and across the three time points. Similarly, the hypotheses

$$H_0 : \lambda_s = c_s\gamma_s, \forall s$$
$$H_1 : \lambda_{st} = c_{st}\gamma_{st}$$

would test if the predator's eating varies by prey species, but not by time, against the most parameter rich model c_{st}. The above two tests should be performed before any of the simpler methods are tested.

2.1 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 obtain the likelihood

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

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 (MLEs) of c_{st}, γ_{st}, and by invariance λ_{st}. Such solutions exist for the simplest model, $c_{st} = 1$, for which only γ_{st} need be estimated, and for the second simplest model, $c_{st} = c$, provided that the data are balanced so that $J_t = J$ and $I_t = I$. Respectively, these solutions are

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

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 to maximise the likelihood numerically. To compute MLEs, we maximise the log-likelihood, using coordinate descent (Luo and Tseng, 1992), by iteratively solving the likelihood equations (i.e. the equations obtained by setting the partial derivatives of the likelihood with respect to the parameters equal to zero). This yields the following equations for updating c in the models c, c_t, and c_s respectively

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

The equation for updating $\hat{\gamma}_{st}$ is

$$\hat{\gamma}_{st} = \frac{X_{\cdot st} + Y_{\cdot st}}{J_t c_{st} + I_t},$$

where c_{st} would be replaced by c, c_t, or c_s depending on the chosen model.

2.2 Unobserved Counts

In many applications, such as DNA-based gut-content analysis, it is not possible to count the number of individuals of each prey species that are in a predator's gut. Instead, it is only possible to detect whether or not a predator consumed the prey species during a given time period, based on the rate at which prey DNA decays in the predator gut (Greenstone et al, 2013). In this case we can still make inference about the predators' preferences for the different prey species by using the expectation-maximisation (EM) algorithm to compute MLEs.

We denote the binary random variable indicating whether the j^{th} predator did in fact eat at least one individual of prey species s in time period t by $Z_{jst} = 1(X_{jst} > 0)$. Given the Poisson assumptions above, these 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

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

The density of Y_{jst} is exactly as in Section 2.1 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 maximising $Q^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ with respect to the parameters in the model to obtain updated estimates of the parameters,

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

These steps are 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:

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

No analytic solution to the M-step exists, however, so we again chose to maximise Q with coordinate descent (Luo and Tseng, 1992). In fact, as we only need to find parameters that increase the value of Q on each iteration, we forgo fully iterating the coordinate descent algorithm to find the maximum and instead perform just one step uphill within each EM iteration (Givens and Hoeting, 2012). 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.

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

2.3 Testing

The likelihood ratio test statistic is

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

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 $\varLambda$ is a χ^2_ρ distribution with ρ degrees of freedom (Wilks, 1938). The degrees of freedom ρ 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. When the observations X_{jst} are not observed, we use $L_{obs}(\boldsymbol{\lambda}, \boldsymbol{\gamma} | Z, Y)$ as the likelihood in the calculation of $\varLambda$

$$L_{obs}(\boldsymbol{\lambda}, \boldsymbol{\gamma} | Z, Y) = \prod_{s,t}^{S,T} \left\{ \prod_{j=1}^{J_t} f_Z(z_{jst} | \boldsymbol{\lambda}) \prod_{i=1}^{I_t} f_Y(y_{ist} | \boldsymbol{\gamma}) \right\} .$$

The level of significance α is used to reject the null hypothesis in favour of the alternative hypothesis if $\mathbb{P}(\chi^2_\rho > \varLambda) < \alpha$.

2.4 Linear Transformations of c_{st}

After determining which model best fits the data, more detail may be extracted through specific 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 $\mathcal{N}(a^t \mathbf{c}, a^t \varSigma a)$, where $\varSigma$ is the covariance matrix of the asymptotic distribution of $\hat{\mathbf{c}}$. Tests of the 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. 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 the first two species are 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 $a = (1, -1, 0)^t$.

3 Simulation Study

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 for both prey species and predator gut count observations are randomly chosen from four overlapping levels: "small" sample sizes are randomly sampled numbers in $[20, 50]$, "medium" $[30, 75]$, "large" $[50, 150]$, and "larger" $[100, 200]$. This is repeated for each level of sample size. We simulate 500 replicate data sets 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 fitted with the true model that generated the data. All simulations of non-count data use $\tau = 10^{-5}$ as the convergence tolerance. A subset of the examples are provided here; the interested reader is referred to the supplementary materials for the complete simulation results.

[Fig. 2 about here.]

For all simulated data, the true parameter values for the rate at which prey species are encountered in the wild are fixed to be $\gamma_{st} = \pi \approx 3.14, \forall s, t$. The values of λ_{st} are set with respect to each data generating model. For model $c_{st} = c$, where predator preferences don't vary by either time or species, we put $\lambda_{st} = 2\pi, \forall s, t$. Under 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 $t \in \{1, \ldots, 5\}$.

[Fig. 3 about here.]

We consider results when the correct model is fit to the simulated data. Figure 2 shows the density plot of the estimates of c_s when fitting the true model to the fully observed count data generated under models c_s, while Figure 3 shows the same for the estimates of c when data is generated under model c. The plots provide evaluations of parameter estimates under each scenario. For model c_s in Figure 2, 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 sample standard deviations of $\mathrm{SD}(\hat{c}_1) = 0.03$ and $\mathrm{SD}(\hat{c}_2) = 0.06$. Figure 3 provides results for model $c_{st} = c$. Averaging across all 500 simulations, the parameter $c = 2$ is estimated as $\hat{c} = 2.00$ with sample standard deviation $\mathrm{SD}(\hat{c}) = 0.06$. This is further seen in Figure 4, where box plots of the parameter estimates, centred at true parameter values, of the correct model fit to data generated from both c_s and c_t show empirically very little bias.

[Fig. 4 about here.]

We next generated data with unobserved counts. As noted above under certain circumstances our unobserved counts model accurately estimates the parameters of interest, and at other times can infinitely over-estimate parameters. To investigate this issue further, we consider the same scenarios mentioned above, but reduce all of the count data down to binary observations. For each scenario, we fit the unobserved counts model as if we knew the true underlying model that generated the observed data.

Figures 5 and 6 contain density plots of the estimates of c_s, c_t for all 500 replications of the data generating models c_s, c_t with the small and the larger sample sizes, respectively. When data are generated under the model c_s and the true model is fit to the non-count data, we find even for the small sample size that point estimates are only very slightly biased. 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 occasionally over-estimated. This effect is easily seen in Figure 6 for the two greatest 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.

[Fig. 5 about here.]

The cluster of estimates for c_5 between 3.5 and 4.0 in Figure 6 comes from data sets in which $Z_{js5} = 1$ for all j, s. For the data shown in Figure 6, this happened 73 times out of the 500 replicated data sets. 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 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 τ, we repeated the algorithm with smaller values of τ for several data sets. As expected, $\hat{c}_5$ increased without bound as we refit the model with increasingly small values of τ.

[Fig. 6 about here.]

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 centred at their respective true parameter values. Figure 7 contains box plots of the same scenarios in Figures 5 and 6. For the 73 cases in which $Z_{js5} = 1$ for all j, s under model c_t with the larger 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 appear to be unbiased, but when no 0s exist in the data the theoretical bias is infinite.

[Fig. 7 about here.]

4 Application

To illustrate these methods, we analysed a data set collected to investigate the feeding preferences of two species of wolf spider, Schizocosa ocreata and Schizocosa stridulans (Araneae: Lycosidae). Every $6 - 12$ days, 10 to 40 spiders were hand-collected between October 2011 and April 2013 within Berea College Forest in Madison County, Kentucky, USA. Spiders were removed from the leaf litter using an aspirator, placed in separate 1.5 mL microcentrifuge tubes filled with 95% EtOH, and preserved at −20℃ until DNA extraction. In parallel, we also surveyed availability of forest floor prey using pitfall traps ($n = 32$). For the analysis, both species of Schizocosa were pooled and the number of spiders and prey were analysed by month. On average, 69 spiders, 111 Diptera, and 297 Collembola were caught in each time period. The range of the sample sizes across all 18 months was 11 to 181 for caught spiders, 7 to 322 for trapped Diptera, and 101 to 755 for trapped Collembola. Figure 8 plots the total number of each order that was caught during each time period.

[Table 1 about here.]

To determine whether spiders had consumed dipterans and/or collemblans, we conducted a molecular analysis of their gut-contents. First, DNA from spiders was extracted using Qiagen DNEasy®Tissue Extraction Kit (Qiagen Inc., Chatsworth, California, USA) following the animal tissue protocol outlined by the manufacturer, with minor modifications. Whole bodies of the spiders were first crushed to release prey DNA from within their alimentary canal for extraction. The 200μL extractions were stored at −20℃ until PCR. Second, order-specific primers from the literature were used to detect the DNA of Collembola and Diptera within the guts of the spiders. Primer pairs designed by Sint et al (2012), targeting the 18S rDNA gene, were used to detect Collembola predation Table 5. A PCR cycling protocol for 12.5μL reactions containing 1x Takara buffer (Takara Bio Inc., Shiga, Japan), 0.2 mM dNTPs, 0.2μM of each primer, 0.625 U Takara Ex TaqTM and 1.5μL of template DNA, using BioRad PTC−200 and C1000 thermal cyclers (Bio-Rad Laboratories, Hercules, California, USA), was optimised as follows: 95℃ for 1 minute, followed by 35 cycles of 94℃ for 30 seconds, 61.2℃ for 90 seconds, and 72℃ for 60 seconds. Primer pairs designed by Eitzinger et al (2014), targeting the 18S rDNA gene, were used to detect Diptera predation Table 5. PCR cycling protocol for 12.5μL reactions with Takara reagents (as above) and 2μL of template DNA was optimised as follows: 95℃ for 1 minute, followed by 40 cycles of 94 for 45 seconds, 60℃ for 45 seconds, and 72℃ for 45 seconds. Both primer pairs were tested for cross-reactivity against a range of prey and predator species from the field site and in all cases, no amplification of DNA was observed, confirming suitable specificity of the primers for this study. Lastly, electrophoresis of 10μL of each PCR product was later conducted to determine success of DNA amplification using 2% Seakem agarose (Lonza, Rockland, Maine, USA) stained with 1x GelRed™nucleic acid stain (Biotium, Hayward, California, USA). This procedure allowed us to determine a presence or an absence of Diptera and Collembola DNA within each spider.

[Fig. 8 about here.]

[Fig. 9 about here.]

These data provide an example of our hierarchy of hypotheses. From the bottom of the graph in Figure 1, we tested the most parameter rich model $c_{st} = \lambda_{st}/\gamma_{st}$ against models c_s, c_t. In both cases, the more parameter rich model fits these data better than is expected by chance; $H_0 : c_s$ versus $H_1 : c_{st}$ gives p-value < 0.0001 and $H_0 : c_t$ versus $H_1 : c_{st}$ gives 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 γ_{st}. Figures 10, 11 plot the point estimates and 95% confidence intervals of c_{st}, for both prey across all time periods.

[Fig. 10 about here.]

[Fig. 11 about here.]

With point estimates of c_{st} under the model $c_{st} = \lambda_{st}/\gamma_{st}$, we can test any number of linear contrasts. For instance, the hypotheses $c_{1t} = c_{2t}$, for $t \in \{1, \ldots, 18\}$ state that wolf spiders equally prefer the orders Diptera and Collembola at each of the 18 time points. 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 earliest estimators of predators' dietary preferences, summarised by Lechowicz (1982) and Manly et al (2002), produced one number summaries that rarely considered more than one prey species, did not take into account changes across time, and had minimal statistical justification. Instead, the indices developed were justified by arguing in favor of each index's unique, and claimed "optimal," properties. Further, these indices could not handle unobserved count data.

A more recent method for testing prey preferences was presenteby by Agusti et al (2003). This method attempts to test the hypothesis that individuals are eating prey at random using a test based on Monte Carlo simulation (more specificially the bootstrap). Unfortunately, we believe that the method described is flawed specifically because it does not attempt to model the distribution of the number of prey items eaten, as we have done. The null hypothesis tested by Agusti et all (2003) is that the proportion of prey species $j = 1, \ldots, J$ in the mesocosm, denoted π_j, is equal to the proportion of prey species j in the predator's guts, denoted by p_j, at all of the sampling times. This is tested by repeatedly simulating the gut contents of spiders in each sample under this hypothesis and then comparing summary statistics computed from the observed data and the simulated data to generate a p-value. The problem with this approach is that the relationship between π_j and p_j depends on the number of prey items eaten by each predator. If each predator had eaten exactly one prey item then the null hypothesis of Agusti et al (2003) would be equivalent to random sampling of prey. However, the relationship between π_j and p_j is more complicated when predators can have more than one prey item in their gut. Given that a predator has m prey items in its gut, randomly sampled from the mesocosm, the marginal probability that at least one item from prey species j is present is $1 - (1 - \pi_j)^m$. The marginal probability that an individual has as least one item from prey species j in its gut is then $\sum_{m=1}^{\infty}(1 - (1 - \pi_j)^m)P(m)$, where $P(m)$ is the probability that a predator has a total of m species in its gut. The presence of species in a predators gut contents also cannot be simulated without modeling the number of items in the gut for exactly the same reason. This shows the importance of the distribution of prey items in the gut in testing the hypothesis of random sampling and because of this it is unclear what hypothesis is being tested by Agusti et al (2003). Monte Carlo simulation methods similar to this could be implemented using the output from our model, but these would be computationally intensive in comparison to the likelihood ratio tests which can be computed almost immediately.

Throughout this paper we assume the complete data were generated from underlying Poisson distributions. Though this assumption is not easily tested, and other factors affecting our model may distort our methods, e.g. sampling issues, environmental variables, differential decay rates, we do not believe a better assumption is readily available to handle all the cases we considered. The Poisson distribution provides reasonable biological interpretation to the available data, data obtained by ecologists interested in prey electivity. One could hypothesize a more complex distribution that similarly has good biological interpretation, say a distribution with more than one parameter. Such a distribution could theoretically perform better than our model when full count data are observed and the sampled data were not generated from a Poisson distribution – if the data were generated from a Poisson distribution then our likelihood ratio test ensures we are using the most powerful test available. But when only unobserved count data is available, any distribution with more than one parameter is unidentifiable. The Poisson distribution on the other hand allowed us to use a relatively simple EM algorithm to estimate parameters of interest when only unobserved count data are available. Future development of our methods might consider more complex distributions to handle variables not presently considered in the case of fully observed count data.

The methods we have presented offer ecologists a formal statistical framework to model and test predators feeding preferences across an array of time points and prey species. Building from intuitive assumptions, we have developed methodologies for analyzing counts of prey items consumed and have extended these methods for analyzing simple presence/absence data. We have provided computationally efficient methods for fitting these models using the EM algorithm and developed powerful tests for a hierarchy of hypotheses based on the likelihood ratio. We are not aware of previous methods that have offered such abilities. These methods are available to researchers via the R package spiders which is available at `http://cran.r-project.org/web/packages/spiders/index.html`.

Acknowledgements The information reported in this paper (No. $15 - 08 - 008$) is part of a project of the Kentucky Agricultural Experiment Station and is published with the approval of the Director. Support for this research was provided by the University of Kentucky Agricultural Experiment Station State Project KY008055 and the National Science Foundation Graduate Research Fellowship Program.

References

Agustí N, Shayler S, Harwood JD, Vaughan I, Sunderland K, Symondson WOC (2003) Collembola as alternative prey sustaining spiders in arable ecosystems: prey detection within predators using molecular markers. Molecular Ecology 12(12):3467–3475

Bischl B, Lang M, Mersmann O (2014) BatchExperiments: Statistical experiments on batch computing clusters. URL `http://CRAN.R-project.org/package=BatchExperiments`, r package version 1.3

Chesson J (1978) Measuring preference in selective predation. Ecology 59(2):211–215

Chesson J (1983) The estimation and analysis of preference and its relationship to foraging models. Ecology 64(5):1297–1304

Clements HS, Tambling CJ, Hayward MW, Kerley GI (2014) An objective approach to determining the weight ranges of prey preferred by and accessible to the five large african carnivores. PloS one 9(7):e101,054

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

Davey JS, Vaughan IP, Andrew King R, Bell JR, Bohan DA, Bruford MW, Holland JM, Symondson WO (2013) Intraguild predation in winter wheat: prey choice by a common epigeal carabid consuming spiders. Journal of Applied Ecology 50(1):271–279

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39(1):1–38

Eitzinger B, Unger EM, Traugott M, Scheu S (2014) Effects of prey quality and predator body size on prey dna detection success in a centipede predator. Molecular ecology 23(15)

Givens GH, Hoeting JA (2012) Computational statistics, vol 708. John Wiley & Sons

Greenstone MH, Payton ME, Weber DC, Simmons AM (2013) The detectability half-life in arthropod predator–prey research: what it is, why we need it, how to measure it, and how to use it. Molecular ecology 23(15)

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

Hellström P, Nyström J, Angerbjörn A (2014) Functional responses of the rough-legged buzzard in a multi-prey system. Oecologia 174(4):1241–1254

Ivlev VS (1964) Experimental ecology of the feeding of fishes. London

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

King RA, Vaughan IP, Bell JR, Bohan DA, Symondson WO (2010) Prey choice by carabid beetles feeding on an earthworm community analysed using species-and lineage-specific PCR primers. Molecular Ecology 19(8):1721–1732

Lechowicz MJ (1982) The sampling characteristics of electivity indices. Oecologia 52(1):22–30

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

Lyngdoh S, Shrotriya S, Goyal SP, Clements H, Hayward MW, Habib B (2014) Prey preferences of the snow leopard (*Panthera uncia*): Regional diet specificity holds global significance for conservation. PloS one 9(2):e88,349

Madduppa HH, Zamani NP, Subhan B, Aktani U, Ferse SC (2014) Feeding behavior and diet of the eight-banded butterflyfish *Chaetodon octofasciatus* in the Thousand Islands, Indonesia. Environmental Biology of Fishes pp 1–13

Manly B, McDonald L, Thomas D, McDonald T, Erickson W (2002) Resource selection by animals: statistical analysis and design for field studies. Nordrecht, The Netherlands: Kluwer

McLachlan G, Krishnan T (2007) The EM algorithm and extensions, vol 382. John Wiley & Sons

Raso L, Sint D, Mayer R, Plangg S, Recheis T, Brunner S, Kaufmann R, Traugott M (2014) Intraguild predation in pioneer predator communities of alpine glacier forelands. Molecular ecology 23(15)

Roualdes EA, Bonner S (2014) spiders: Fits predator preferences model. R package version 1.0

Schmidt JM, Barney SK, Williams MA, Bessin RT, Coolong TW, Harwood JD (2014) Predator–prey trophic relationships in response to organic management practices. Molecular ecology 23(15)

Sint D, Raso L, Traugott M (2012) Advances in multiplex PCR: balancing primer efficiencies and improving detection success. Methods in Ecology and Evolution 3(5):898–905

Strauss RE (1979) 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

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

Wickham H (2011a) The split-apply-combine strategy for data analysis. Journal of Statistical Software 40(1):1–29, URL http://www.jstatsoft.org/v40/i01/

Wickham H (2011b) testthat: Get started with testing. The R Journal 3:5–10, URL http://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf

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

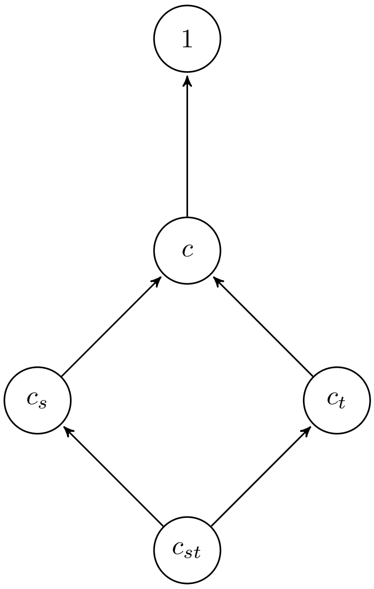

Fig. 1 This hierarchy of hypotheses suggests the order in which the discussed models should be tested. One begins with the most complex models at the bottom and sequentially, following the arrows, tests simpler hypotheses using the formal test described in Section 2.3 until a final model is established.

Small Sample Size

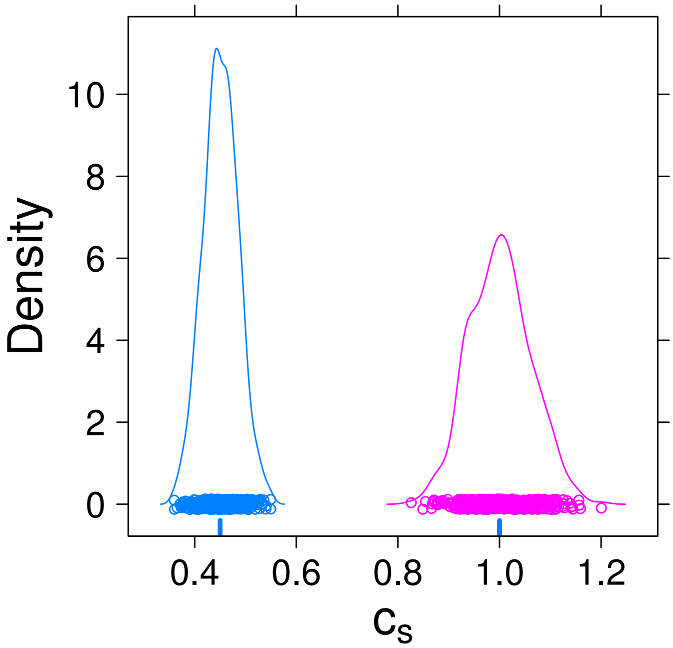

Fig. 2 The density plot of all 500 estimates of fitting the true model to the data generated from models c_s with the small sample size is shown. Each element of c_s is colour coded for clarity, and ticks on the x-axis show the true parameter values.

Medium Sample Size

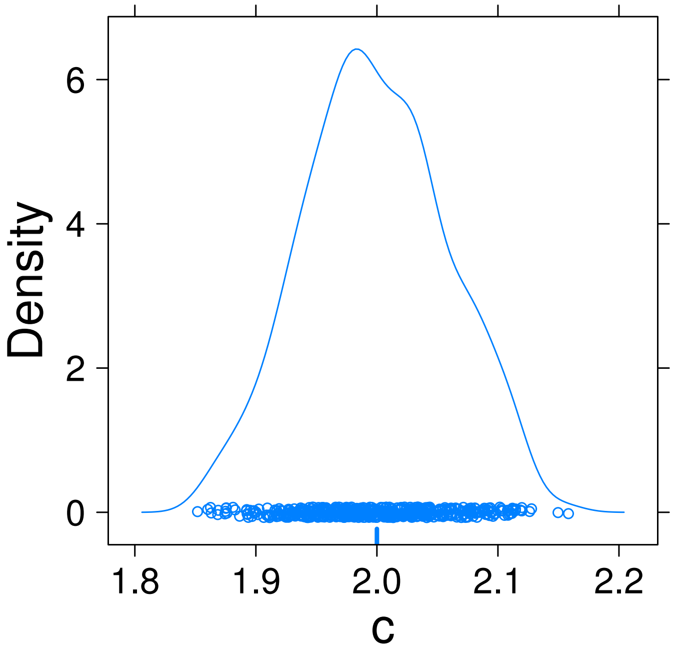

Fig. 3 The density plot of all 500 estimates of fitting the true model to the data generated from model *c* with the medium sample size is shown. A tick on the x-axis shows the true parameter value.

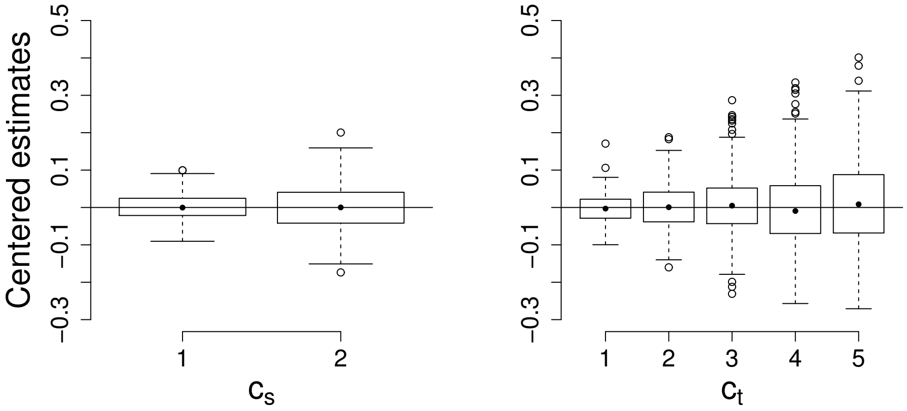

Fig. 4 Shown are all 500 estimates, cent-red at the true parameter values, from fitting the true model to the data generated from models c_s, c_t with sample sizes small and medium, respectively.

Small Sample Size, Non-Count Data

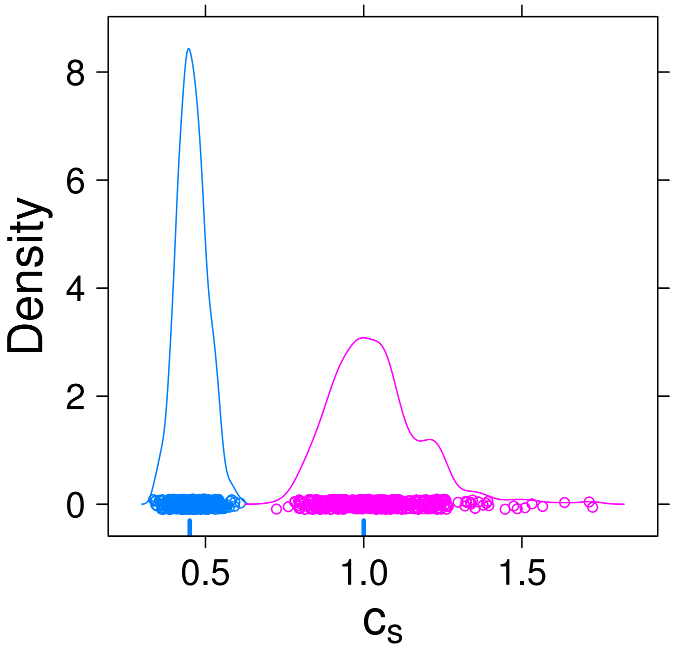

Fig. 5 The density plot of all 500 estimates of fitting the true model to the data generated from model c_s, when counts are not observed, is shown with the small sample size. Each element of c_s is colour coded for clarity, and ticks on the x-axis show the true parameter values.

Larger Sample Size, Non-Count Data

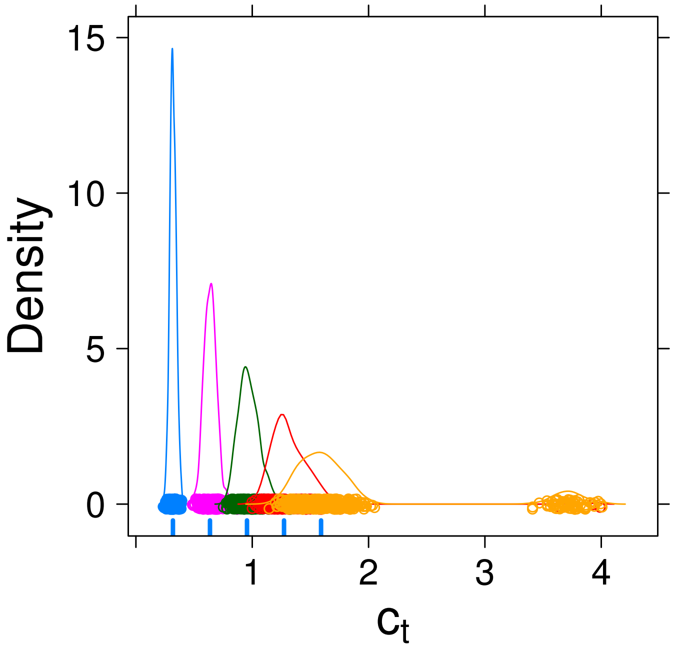

Fig. 6 The density plot of all 500 estimates of fitting the true model to the data generated from model c_t, when counts are not observed, is shown with the larger sample size. Each element of c_t is colour coded for clarity, and ticks on the x-axis show the true parameter values.

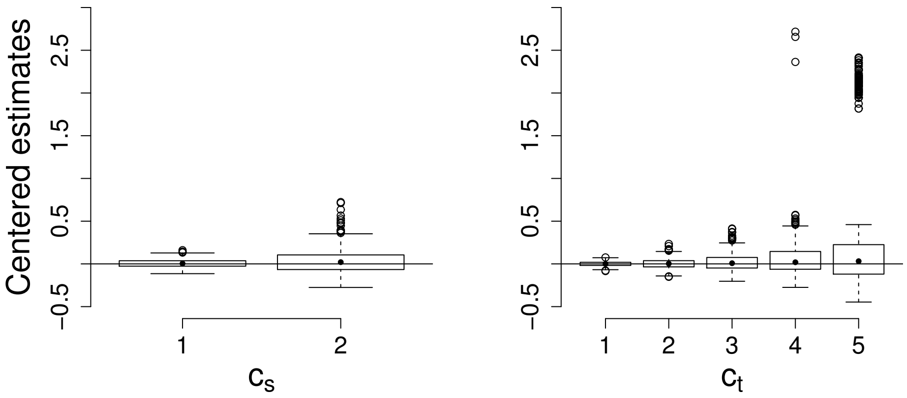

Fig. 7 Shown are all 500 estimates, centred at the true parameter values, from fitting the true model to the data generated from models c_s, c_t, when counts are not observed, with sample sizes small and larger, respectively.

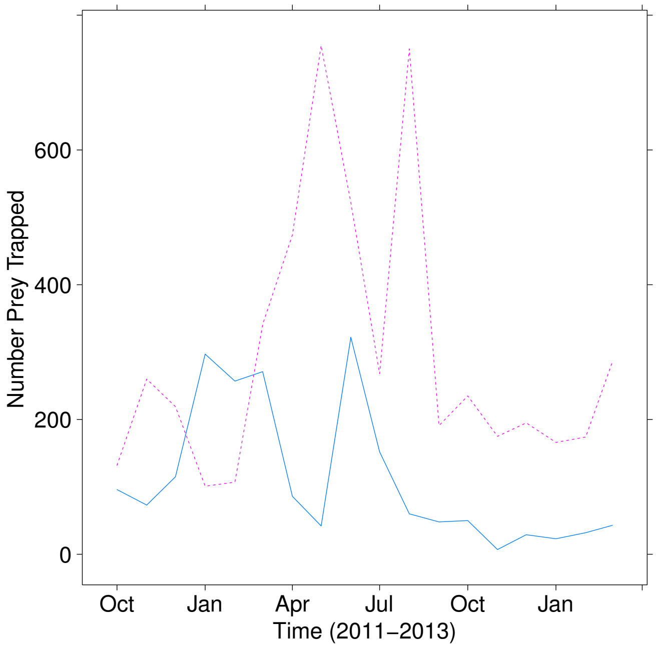

Fig. 8 For both Collembola (pink/dashed) and Diptera (blue/solid), the plot shows the number of the prey trapped in each time period.

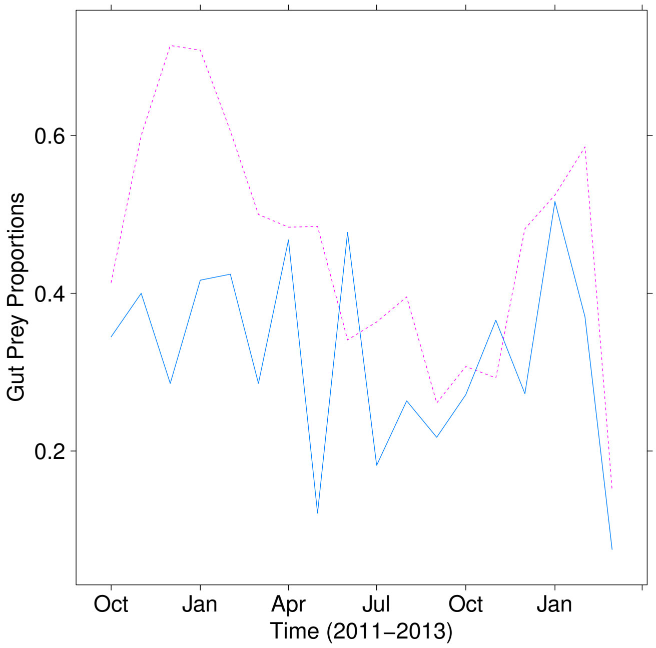

Fig. 9 For both Collembola (pink/dashed) and Diptera (blue/solid), the plot shows the prey proportions in the sampled wolf spiders' guts in each time period.

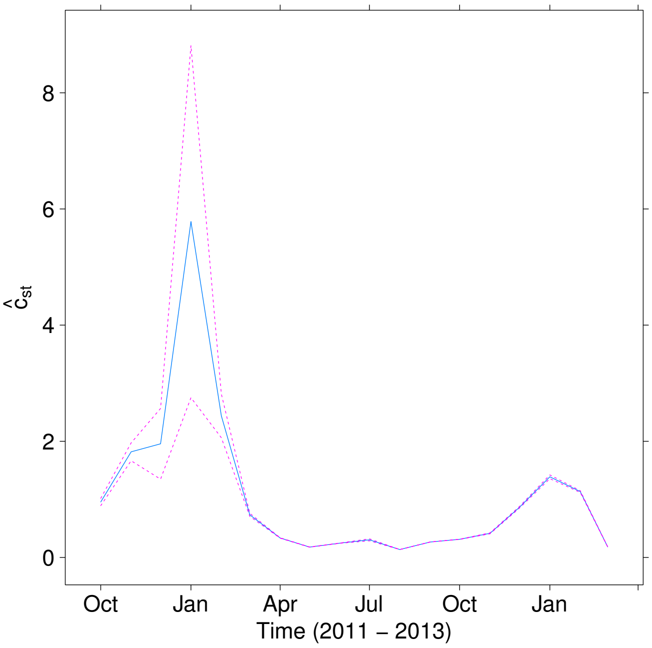

Fig. 10 Point estimates (blue/solid) and 95% confidence intervals (pink/dashed) as estimated from the model c_{st} for Collembola.

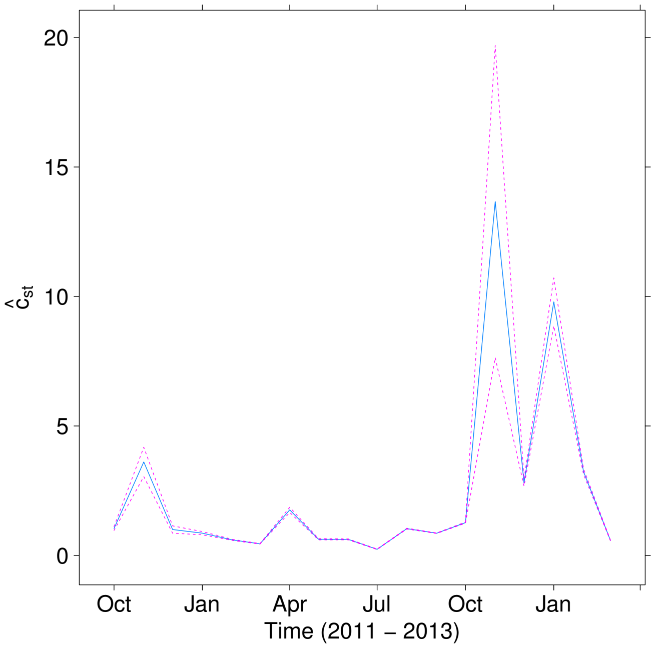

Fig. 11 Point estimates (blue/solid) and 95% confidence intervals (pink/dashed) as estimated from the model c_{st} for Diptera.

Target group	**Primer names and sequences** $5' - 3'$	**Size (bp)**	**Source**
Collembola	Col3F: GGACGATYTTRTTRGTTCGT	228	Sint et al (2012)
	Col-gen-A246: TTTCACCTCTAACGTCGCAG		
Diptera	DIPS16: CACTTGCTTCTTAAATrGACAAATT	198	Eitzinger et al (2014)
	DIPA17: TTyATGTGAACAGTTTCAGTyCA		

Table 1 Targeted prey orders, primer names and sequences, size of amplicon, and source of design for the detection of prey taxa within the guts of Schizocosa spiders. Both primer sets were used in singleplex PCR assays.