# Predator Preferences Model

S. Bonner and E. A. Roualdes

July 20, 2014

## 1 Generic Model

Strauss' statistic $L_{st}$ [Str79] is the difference between $r_{st}$ the proportion of prey species $s$ found in the gut of a predator during occurrence $t$ and $p_{st}$ the proportion of prey species $s$ found in the habitat of said predator during occurrence $t$. When the statistic $L_{st} := r_{st} - p_{st}$ is equal to zero we say the predator ate prey species $s$ randomly during time $t$. A positive difference, $L_{st} > 0$, indicates preferential eating of prey species $s$, and negative values indicate aversion to prey species $s$ at time $t$. We can replicate and add to the modeling considered by Strauss with the following set up.

Let $X_{jst} \sim \mathcal{P}(\lambda_{st})$ denote the number of prey species $s$ found in the gut of predator $j$ during occurrence $t$; $j \in \{1, \ldots, J\}, s \in \{1, \ldots, S\}, t \in \{1, \ldots, T\}$. For now we ignore the fact that we only observe a 1 if predator $j$ ate prey species $s$ during occurrence $t$ or a zero if the predator did not. Further, $Y_{ist} \sim \mathcal{P}(\gamma_{st})$ will denote the number of prey species $s$ found in trap $i$, $i \in \{1, \ldots, I\}$, hypothesized to represent the habitat of predator $j$, during time occurrence $t$. Interest lies in the null hypothesis

$$H_0 : \boldsymbol{\lambda}_t = c\boldsymbol{\gamma}_t, \forall t \tag{1}$$

where $c \in \mathbb{R}$. This hypothesis suggests that the predator's eating preferences are independent of time, i.e. the ratio of the rate at which the predator eats prey species $s$ to the rate at which prey species $s$ is found in the habitat is constant across time. Strauss' $L$ is a special case of this framework where only a particular combination of prey species $s$ and time $t$ is considered, and $c$ is equal to the ratio of total number of prey found in the gut of the predator to the total number of prey found in the habitat.

We first consider the parameters of this model under the null hypothesis. Here, $S \cdot T + 1$ parameters will be estimated, $\gamma_{st}$ for $(s,t) \in \{S \times T\}$ and $c$.

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

The parameters under the general alternative hypothesis, $H_1 : \boldsymbol{\lambda}_t \neq c\boldsymbol{\gamma}_t, \forall t$, are simply the maximum likelihood estimates.

$$\hat{\lambda}_{st} = \frac{X_{\cdot st}}{J} \text{ and } \hat{\gamma}_{st} = \frac{Y_{\cdot st}}{I}$$

Null hypothesis (1) above is tested with the standard likelihood ratio statistic

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

where $\theta_0 = \{\boldsymbol{\lambda}_t = c\boldsymbol{\gamma}_t, \forall t : (\gamma_{st}, c) \in \mathbb{R}^{ST+1}\}$ represents the likelihood under the null hypothesis, and $\theta_1 = \{\boldsymbol{\lambda}_{st} \neq c\boldsymbol{\gamma}_{st} : (\lambda_{st}, \gamma_{st}) \in \mathbb{R}^{2ST}\}$ represents the likelihood under the alternative hypothesis. Then $\Lambda \dot{\sim} \chi^2_{ST-1}$ and null hypothesis (1) is rejected when $P(\chi^2_{ST-1} > \Lambda) < \alpha$.

## 1.1 Unequal Trap Schedules

Consider a situtation where the traps (catching prey species) were left out for a differing length of time within each occurrence, e.g. trap $i$ in month $t$ was left out for 6 days, but trap $i'$ in month $t$ was left out for 3 days. We expect to catch an unequal number of prey in each trap simply because one trap was left out for a longer time within month $t$.

Put $I_t := \sum_i numDays_{it}$ to be the total number of days all traps were left out during month $t$. Then the complete data log-likelihood under this scenario is as follows.

$$l = -J\sum_{st}\lambda_{st} + \sum_{st} X_{\cdot st}\log\lambda_{st} - \sum_t I_t \sum_s \gamma_{st} + \sum_{st} Y_{\cdot st}\log\gamma_{st} + constant$$

If there exists a differing number of predators in each time period, we can similarly model this by indexing $J$ by $t$, say $J_t$.

## 1.2 Non-Count Gut Data

With smaller animals it is not always possible to observe a count of prey species eaten by a predator species. Instead, a binary response of whether or not DNA of said prey species exists in the gut of the predator is observed. In this case, we can treat the count of eaten prey species as missing and maximize the likelihood via the EM algorithm.

We model what information we do observe. Denote this binary response by $Z_{jst}$, which takes on the value 1 if the $j^{th}$ predator ate prey species $s$ in time $t$ and 0 otherwise. Hence, $\Lambda(X,Y)$ is still calculable via the EM algorithm as follows. Write out the complete data likelihood,

$$L_{comp}(X,Z,Y|\boldsymbol{\lambda},\boldsymbol{\gamma}) = \prod_t^T\prod_s^S\prod_j^J f_{X,Z}(x,z|\boldsymbol{\lambda})f_Y(y|\boldsymbol{\gamma}) \tag{3}$$

Take the expectation of $\log L_{comp}$ with respect to the distribution of the missing data given the observed data and parameters,

$$f_{X|Y,Z,\boldsymbol{\lambda},\boldsymbol{\gamma}}(x) = \frac{f_{X,Y,Z}(x,y,z|\boldsymbol{\gamma},\boldsymbol{\lambda})}{f_{Y,Z}(y,z|\boldsymbol{\lambda},\boldsymbol{\gamma})}.$$

Then maximize $\mathbb{E}_{X|Y,Z}l_{comp}$ with respect to the parameters $\gamma_{st}, \lambda_{st}, c$, just as above since here $X_{jst}$ is simply replaced by its expectation.

# 2 Simulations

In order to simulate and test data, we need to install an R package named spiders. To install the package spiders, you can use devtools, or download the files from the GitHub webpage.

```
## library(devtools) install_github('spiders',
## 'roualdes')
library(spiders)
```

We can then simulate some data by using the function spiders::simData. Input to this function includes the number of predator species observed at each time point, the number of traps used at each time point, the number of prey species of interest, the number of time points for which measurements were taken, and rate parameters $\lambda_{st}, \gamma_{st}$ for which the predator ate and observed, respectively, prey species $s$ at time $t$. For now, we assume that the number of predators and traps is fixed across time points. Below we specify rate parameters that will be proportional across time; we expect to fail to reject null hypothesis (1) here.

```
Predators <- 77
Traps <- 87
PreySpecies <- 2
Times <- 5
ST <- Times * PreySpecies
l <- matrix(1, nrow = Times, ncol = PreySpecies)
g <- matrix(2, nrow = Times, ncol = PreySpecies)
```

Simulate some data as follows.

```
fdata <- simPref(PreySpecies, Times, Predators, Traps, l,
    g)
```

Then we can fit the model and calculate a p-value from these data. The output of the model fit is below.

```
(prefs <- predPref(fdata$eaten, fdata$caught))

$gAlt
$gAlt$gamma
  preySpecies1 preySpecies2
1        2.000        1.874
2        2.184        2.011
3        1.471        1.989
4        2.092        1.966
5        2.103        2.046
```

```
$gAlt$lambda
  preySpecies1 preySpecies2
1       0.8571       1.0000
2       0.8961       0.8571
3       1.0000       0.7013
4       0.8571       1.0519
5       1.0130       1.1299


$null
$null$gamma
  preySpecies1 preySpecies2
1        1.943        1.943
2        2.097        1.951
3        1.659        1.838
4        2.008        2.040
5        2.113        2.145

$null$c
[1] 0.4745


$loglikH1
[1] 9496

$loglikH0
[1] 9488

$numPredators
[1] 77 77 77 77 77

$numTraps
[1] 87 87 87 87 87

$Lambda
[1] 15.86

$df
[1] 9

$p.value
[1] 0.06977
```

An example where we reject the null hypothesis would have rate parameters that are differently proportional across time. Consider the following.

```
l <- matrix(1:ST, nrow = Times, ncol = PreySpecies)
g <- matrix(2 * (ST:1), nrow = Times, ncol = PreySpecies)
fdata <- simPref(PreySpecies, Times, Predators, Traps, l,
    g)
predPref(fdata$eaten, fdata$caught)$p.value

[1] 0
```

# Tests

## Complete, Balanced Data

We can test our functions by simulating many datasets with the same underlying parameters, fitting the model, and plotting the distribution of the estimates. Let's first set the values of the parameters to be constant across time and species.

```
l <- matrix(1.5, nrow = Times, ncol = PreySpecies)
g <- matrix(2 * pi, nrow = Times, ncol = PreySpecies)
```

Using the function `spiders::testPref` we simulate $M = 250$ datasets based on the above parameters and fit the model to each. By default, only 4 of the $S * T = 10$ parameters of each $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are stored.

```
M <- 250
system.time(out <- testPref(PreySpecies, Times, Predators,
    Traps, l, g, M = M))

   user  system elapsed
  5.430   0.041   5.482
```

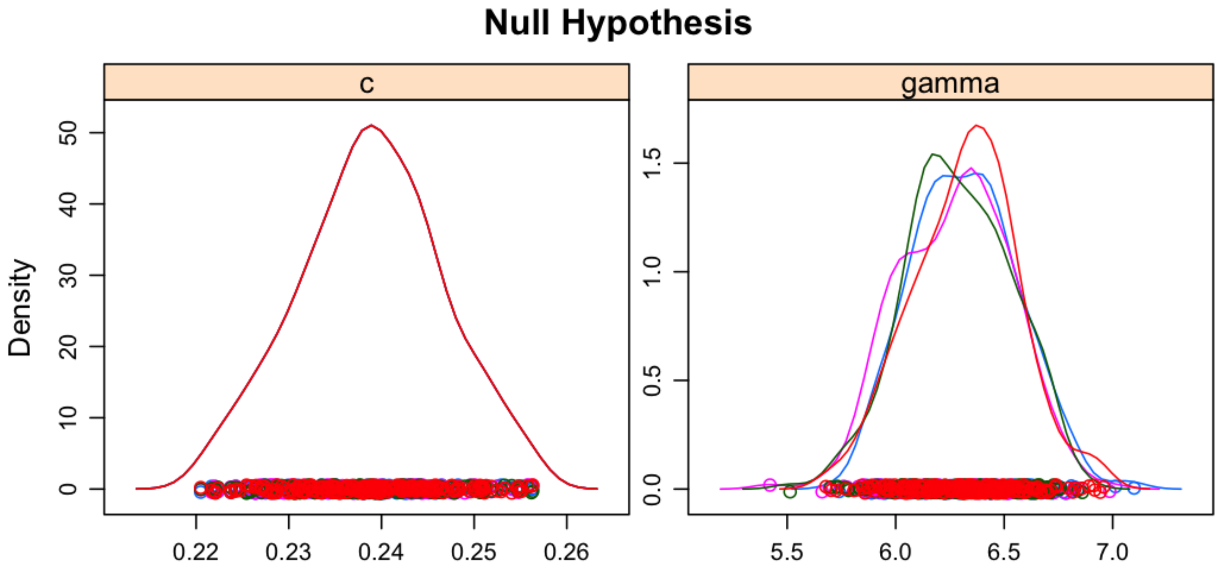
Also output from this function is the number of iterations within `spiders::predPref`, in this case since we are fitting the complete, balanced data so there is only

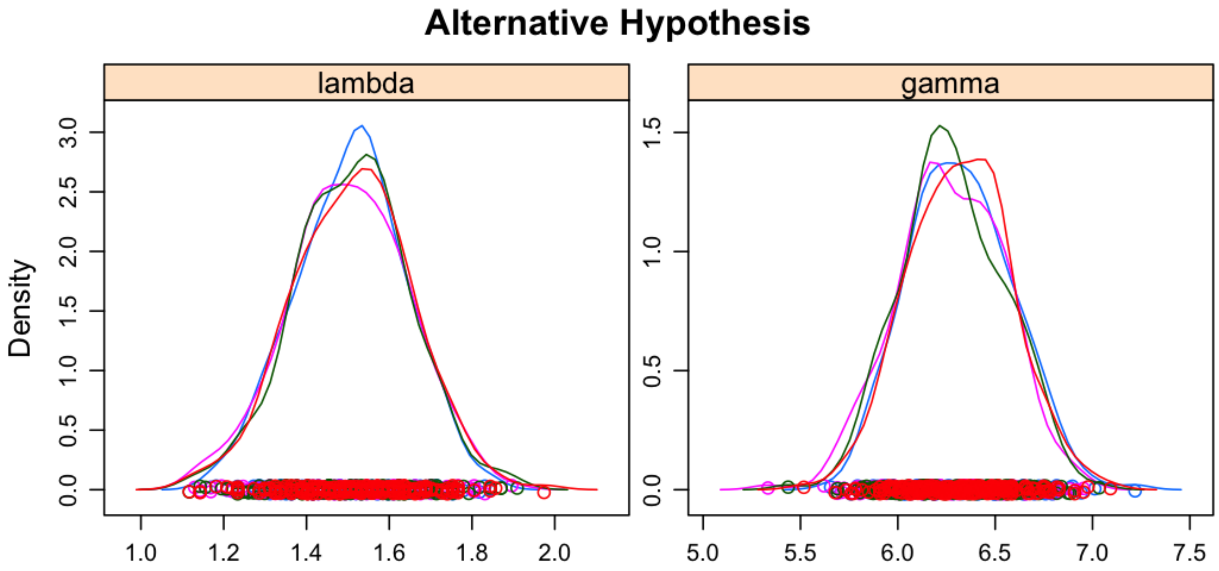
```
out$iters[M, ]

null gAtl
   1    1
```

for every simulated dataset.

Since we know the value of the underlying parameters, we can plot these data and visually inspect our algorithm's performance using the function `spiders::plotTestPref`.

```
plotTestPref(out$null, out$alt)
```



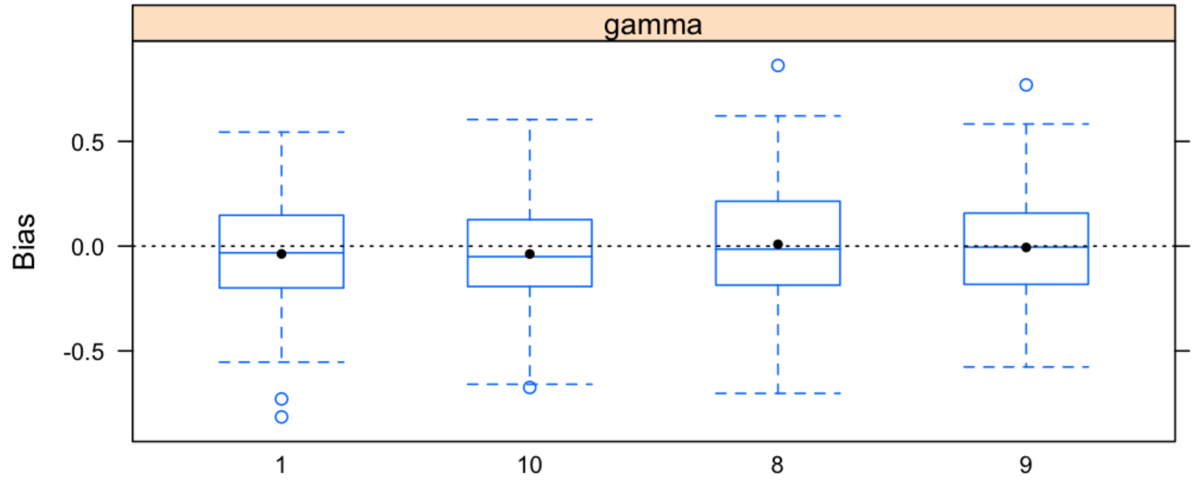


To see that $c$ is estimated accurately, we use the fact that $\gamma_{st} = \gamma$ for all $\{s, t\}$ by averaging over the 4 values of $\boldsymbol{\gamma}$ that were sampled: $\bar{\hat{c}} * \bar{\hat{\gamma}} = 1.505$.

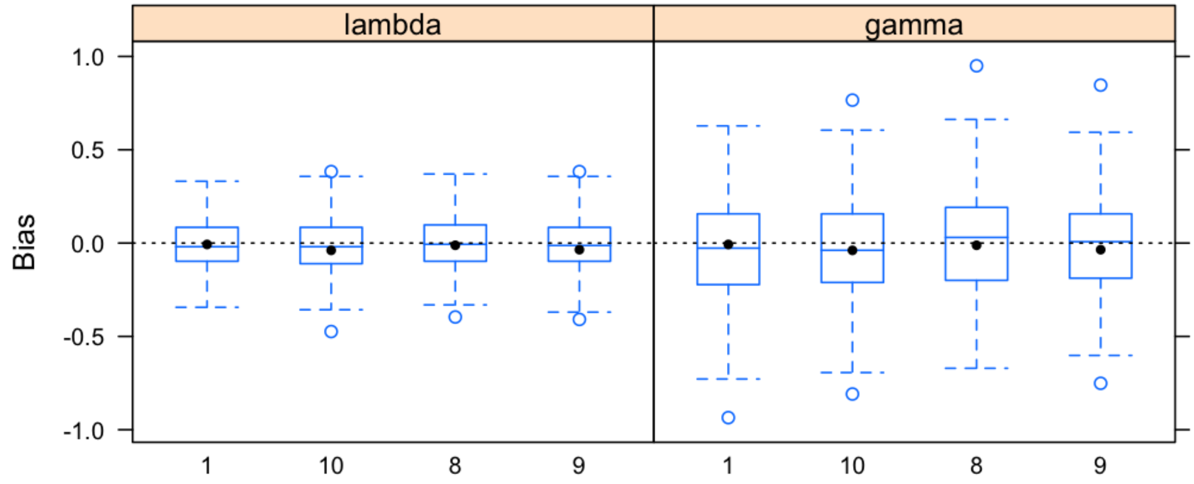
Further, we can calculate and plot the estimated bias. The boxplot contains the same randomly sampled parameters of $\lambda_{st}, \gamma_{st}$ as before. Here, the indices on the abcissa represent the enumeration of the set $\{(T, S)\}$ indexing $T$ first, e.g. index 8 for would refer to the parameter value for the third time point of the second prey species since we have 5 time points and 2. Averaged over all simulations, both mean (dot) and median (line), for each parameter value, are shown.

```
bias <- calcBias(out$null, out$alt, l, g)
plotBias(bias)
```

Null Hypothesis



Alternative Hypothesis



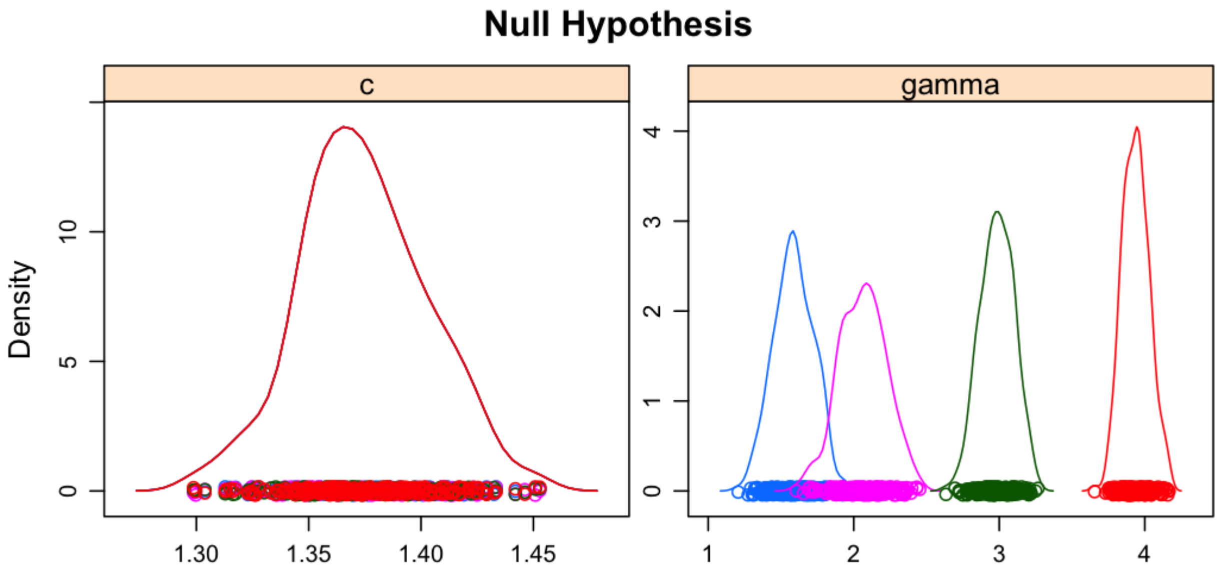
### Complete, Unbalanced Data

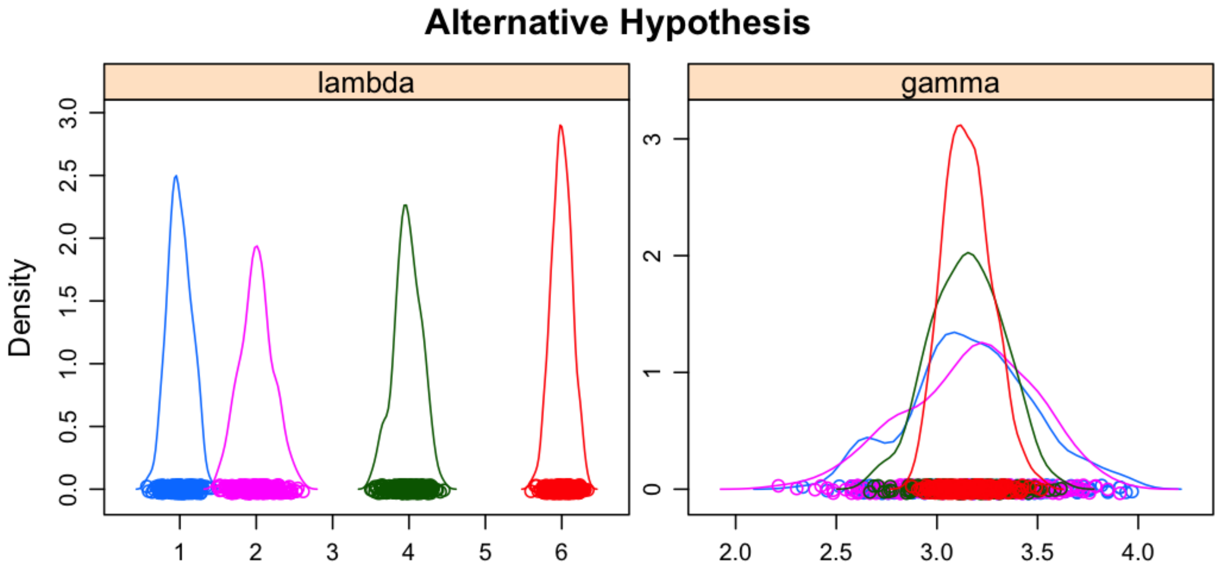
Here, we simply have to simulate an uneven number of predators and prey species caught in each time period. We do this and also create a more fun set of parameter values.

```
Predators <- 4 * c(11, 22, 33, 44, 77)
Traps <- 3 * Predators/4
l <- matrix(cbind(1:5, 2:6), nrow = Times, ncol = PreySpecies)
g <- matrix(pi, nrow = Times, ncol = PreySpecies)
```

Then we can simulate the data and plot it just as before.

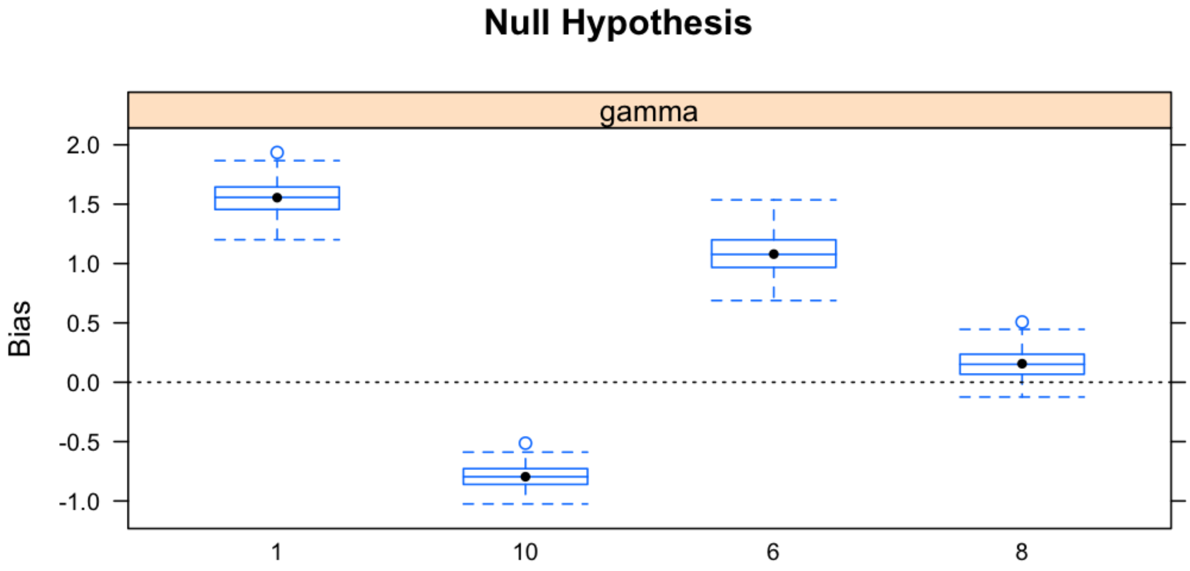
```
system.time(out <- testPref(PreySpecies, Times, Predators,
    Traps, l, g, M = M))

   user  system elapsed
 11.587   0.053  11.646

plotTestPref(out$null, out$alt)
```

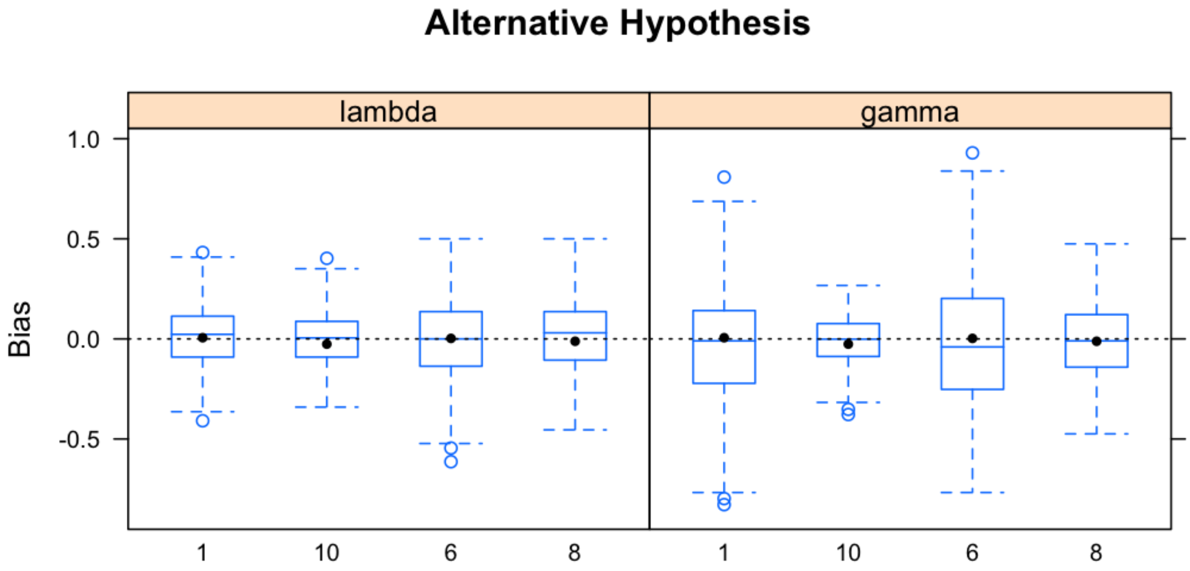




It is fairly obvious, by the bias plots, when the null hypothesis is false.

```
bias <- calcBias(out$null, out$alt, l, g)
plotBias(bias)
```



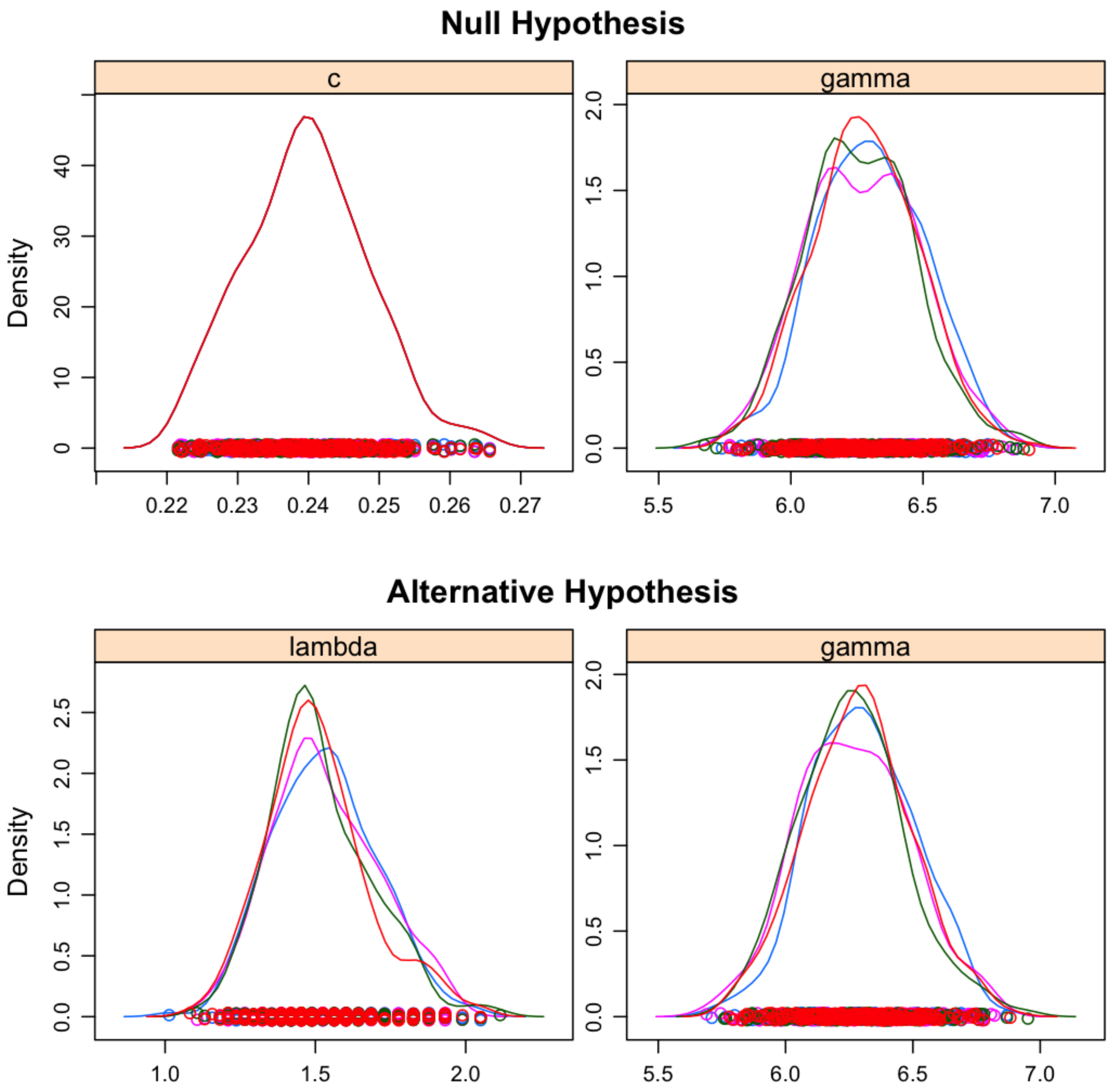


### Non-Count Gut Data

Similarly, we can test the EM algorithm using the same funciton, but declaring the variable `EM` to be true.

```
Predators <- 124
Traps <- 145
l <- matrix(1.5, nrow = Times, ncol = PreySpecies)
g <- matrix(2 * pi, nrow = Times, ncol = PreySpecies)
system.time(out <- testPref(PreySpecies, Times, Predators,
    Traps, l, g, M = M, EM = TRUE))
```

```
   user  system elapsed 
 34.314   0.087  34.405 

plotTestPref(out$null, out$alt)
```



The same as before, $c$ is estimated fairly well: $\bar{\hat{c}} * \bar{\hat{\gamma}} = 1.5058$. Also, we can check how many iterations of the EM algorithm the last simulated dataset required

```
out$iters[M, ]

null gAtl 
  13   19 
```

# A Details

## A.1 Complete, Balanced Data

With balanced, count data the log-likelihood is

$$l = -J \sum_{s,t} \lambda_{st} + \sum_{s,t} X_{\cdot st} \log \lambda_{st} - I \sum_{s,t} \gamma_{st} + \sum_{s,t} Y_{\cdot st} + \text{const.}$$

Under $H_0 : \lambda_{st} = c\gamma_{st}$ we find solutions of first derivatives of $l$ with respect to $c, \gamma_{st}$ set equal to zero to be

$$c = \frac{\sum_{s,t} X_{\cdot st}}{J \sum_{s,t} \gamma_{st}} \quad \text{and} \quad \gamma_{st} = \frac{X_{\cdot st} + Y_{\cdot st}}{Jc + I}.$$

Solving these equations simultaneously gives

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

Under $H_1 : \lambda_{st} \neq \gamma_{st}$ we find the maximum likelihood estimates to be

$$\hat{\lambda}_{st} = \frac{X_{\cdot st}}{J} \quad \text{and} \quad \hat{\gamma_{st}} = \frac{Y_{\cdot st}}{I}.$$

## A.2 Complete, Unbalanced Data

With unbalanced, count data the log-likelihood is

$$l = -\sum_t J_t \sum_s \lambda_{st} + \sum_{s,t} X_{\cdot st} \log J_t \lambda_{st} - \sum_t I_t \sum_s \gamma_{st} + \sum_{s,t} Y_{\cdot st} \log I_t \gamma_{st}.$$

Under $H_0 : \lambda_{st} = c\gamma_{st}$ we find solutions of first derivatives of $l$ with respect to $c, \gamma_{st}$ set equal to be

$$c = \frac{\sum_{s,t} X_{\cdot st}}{\sum_t J_t \sum_s \gamma_{st}} \quad \text{and} \quad \gamma_{st} = \frac{X_{\cdot st} + Y_{\cdot st}}{cJ_t + I_t}.$$

Since we can't solve these equations simultaneously, we exploit the fact that the log-likelihood is concave; each term is concave and convexity is closed under addition. Hence, we can iteratively maximize the profile log-likelihood until convergence.

Under $H_1 : \lambda_{st} \neq \gamma_{st}$ we find the maximum likelihood estimates to be

$$\hat{\lambda}_{st} = \frac{X_{\cdot st}}{J_t} \quad \text{and} \quad \hat{\gamma}_{st} = \frac{Y_{\cdot st}}{I_t}.$$

## A.3 Non-Count Data

When the count data are not observed we can fall back on treating the counts as missing, so long as we do observe indicators telling us when the predator species did eat prey species $s$.

### E-Step

For this we need the conditional distribution of the missing data given the observed data

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

Since the joint density of $X_{jst}, Z_{jst}$ is

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

it is easy to calculate the expected value of the complete data log-likelihood

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

with respect to $f_{X|Y,Z,\boldsymbol{\lambda},\boldsymbol{\gamma}}$, where the expectation conditions on all of the remaining parameters.

## M-Step

We maximize $\mathbb{E}l_{comp}$ with respect to its parameters $\lambda_{st}, \gamma_{st}, c$, thus completing one step of the EM algorithm.

## Hypotheses Under EM

Under $H_0 : \lambda_{st} = c\gamma_{st}$, we iteratively solve the following equations within on step of the EM algorithm

$$\hat{\gamma}_{st} = \frac{z_{\cdot st}\mathbb{E}X_{jst} + Y_{\cdot st}}{cJ_t + I_t} \qquad \text{and} \qquad \hat{c} = \frac{\sum_{s,t} z_{\cdot st}\mathbb{E}X_{jst}}{\sum_t J_t \sum_s \gamma_{st}}.$$

Under $H_1$ we complete one step of the EM algorithm via

$$\hat{\lambda}_{st} = \frac{z_{\cdot st}\mathbb{E}X_{jst}}{J_t} \qquad \text{and} \qquad \hat{\gamma}_{st} = \frac{Y_{\cdot st}}{I_t}.$$

# References

[Str79] 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** (1979), no. 4, 344–352.