Multivariate analyses for UHURU Invert data

By Jacob Levine

Jacob Levine

January 16, 2022

Contents

1. Abundance

1.1 Modelling approach

Let's start with the abundance data. The basic idea here is to fit a model which takes into account a possible correlation structure among the responses (order abundance) in order to avoid the typical pitfalls of multiple testing (i.e. fucked type 1 error rates). Because the responses are counts, we should also use some type of Poisson GLM. A nice way to accomplish both these things is to fit a Multivariate Poisson LogNormal Model. The multivariate Poisson LogNormal model is given by the following:

$$\begin{aligned} Z_i \tilde{} N(\mu_i, \Sigma) \\ Y_{i,j}|Z_{i,j} \ P(exp(Z_{i,j})) \\ \mu_i = x_i^T \theta_j \end{aligned} \tag{1}$$

where $Y_{i,j}$ is the number of individuals of order j observed at site i, μ_i is the vector of expected log abundance at site i and is calculated from the covariates, x_i, and estimated coefficients, θ_j. Σ gives the correlation structure among orders, and Z_i is referred to as the latent vector. We can fit this class of models using the `VGAM` package in R.

We are interested primarily in the effect of treatment on order abundance, though also need to control for the experimental design. There are several ways to do this, but the easiest is probably to estimate a fixed effect for each unique replicate (6 in total). This gives us two covariates: Treatment ('C', 'M', or 'T') and Replicate ('1N', '2N', '3N', '1S', '2S', '3S').

1.2 Loading the data

Now we can load in some data and get it intoa nice format=:

```
library(VGAM)

## load data
uhurua <- read.csv("UHURU_summary_metadata_abundance.csv", row.names = 1, header = TRUE

removeWords <- function(str, stopwords) {

  x <- unlist(strsplit(str, "_"))
  paste(x[!x %in% stopwords], collapse = " ")

}

colnames(uhurua) <- apply(X = matrix(colnames(uhurua)), MARGIN = 1, FUN = removeWords,
uhurua[,16:31]
```

```
    Acari Aranae Blattodea Coleoptera Diptera Gatropoda Hemiptera Hymenoptera
N1C     4     28         0         21     128         0       115        1319
N1M     3     35         0         59      61         0       181         486
N1T     4     34         2         39      80         1       147         938
N2C     4     23         0         96      93         0       204        2830
N2M     6     27         1         59     113         0       320        2813
N2T     0     13         0         46      66         0       108        1828
N3C     3     30         0         41      32         0       337         629
N3M     4     43         0         81     107         0       453         845
N3T     1     42         0         28      40         0       126        1028
S1C     5     10         0         56     101         1       482         378
S1M     3     15         0        100     119         0      1248         258
S1T     1     36         0         88     131         0       231         254
S2C    16     17         0         88     226         0       430         349
S2M     7     36         0        176     126         0      1595         425
S2T     3     27         0        117      70         2       236         351
S3C     4     33         0         64      68         1       475         173
S3M     0     27         0         89      42         1       388         662
S3T     3     38         0        109      96         0      1122         168
    Isopoda Lepidoptera Mantodea Neuroptera Orthoptera Phasmatodea Psocoptera
```

N1C	0	29	5	0	47	7	0
N1M	0	35	1	2	56	6	0
N1T	0	34	0	0	55	23	0
N2C	0	67	0	0	35	7	0
N2M	0	42	0	13	29	8	0
N2T	0	47	0	0	35	22	7
N3C	0	13	0	14	101	15	0
N3M	0	34	0	15	22	8	0
N3T	0	0	2	5	17	27	0
S1C	0	31	10	9	23	6	0
S1M	0	65	2	81	24	3	0
S1T	0	36	10	3	22	8	0
S2C	5	46	25	14	29	8	0
S2M	1	145	2	147	19	4	0
S2T	0	50	7	9	17	3	0
S3C	1	29	5	15	25	5	0
S3M	1	26	6	16	15	3	0
S3T	1	66	8	37	21	8	0

	Solifugae
N1C	6
N1M	0
N1T	0
N2C	0
N2M	0
N2T	1
N3C	0
N3M	0
N3T	0
S1C	0
S1M	0
S1T	0
S2C	0
S2M	0
S2T	0
S3C	0
S3M	0
S3T	0

Some of the species seem to have very sparse data (Solifugae, Psocoptera, Isopoda, Gatropoda, Blattodea). I think it is probably best that we remove

these from our analysis (by just not fitting models on them) for the time being as the model fits for them will likely be weak and its unlikely we would glean anything terribly exciting about them anyways.

1.3 Fitting the model

```
## make sure replicate is factor and not numeric
uhurua$Replicate <- as.factor(uhurua$Replicate)

## fit a vglm
abund_model <- vglm(cbind(Acari,
                          Aranae,
                          Coleoptera,
                          Diptera,
                          Hemiptera,
                          Hymenoptera,
                          Lepidoptera,
                          Mantodea,
                          Neuroptera,
                          Orthoptera,
                          Phasmatodea) ~ Treatment + Block + Block:Replicate,
                    family = "poissonff",
                    data = uhurua)

## extract information we want
summary <- summary(abund_model)
coef_table <- summary@coef3

## utility function to make the names nicer to read
rename <- function(x) {

  split <- unlist(strsplit(x, ":"))
  num <- split[length(split)]
  spp <- colnames(summary@y)[as.numeric(num)]
  if (length(split) > 2) {

    newname <- paste0(spp, ":", split[1], ":", split[2])

  }
```

```
  else {

    newname <- paste0(spp, ":", split[1])

  }

  return(newname)

}

## employ our utility function
rownames(coef_table) <- apply(X = matrix(rownames(coef_table), ncol = 1), MARGIN = 1, F
simple_coef <- coef_table[!grepl("Replicate", rownames(coef_table)),]
simple_coef <- data.frame(simple_coef[order(rownames(simple_coef)),])
for (i in 1:nrow(simple_coef)) {

  p.value <- simple_coef[i, "Pr...z.."]
  if (p.value < 0.05) clar <- "*"
  else clar <- " "
  simple_coef[i, "clarity"] <- clar

}
colnames(simple_coef) <- c("Estimate", "std.error", "z.value", "p.value", "clarity")

## check out the results
simple_coef[,c(1,2,4,5)]
```

```
                          Estimate  std.error       p.value clarity
Acari:(Intercept)       1.71873433 0.32342288  1.071261e-07       *
Acari:BlockS           -0.20067070 0.44946657  6.552620e-01
Acari:TreatmentM       -0.44802472 0.26693827  9.327230e-02
Acari:TreatmentT       -1.09861229 0.33333333  9.812898e-04       *
Aranae:(Intercept)      3.28124760 0.12432197 1.642412e-153       *
Aranae:BlockS          -0.46383711 0.16340967  4.532720e-03       *
Aranae:TreatmentM       0.26072626 0.11205659  1.997924e-02       *
Aranae:TreatmentT       0.29826418 0.11115465  7.289386e-03       *
Coleoptera:(Intercept)  3.46872517 0.10197393 1.300942e-253       *
Coleoptera:BlockS       0.71804473 0.11181110  1.345440e-10       *
Coleoptera:TreatmentM   0.43242092 0.06712146  1.176261e-10       *
```

Coleoptera:TreatmentT	0.15415068	0.07123314	3.046203e-02	*
Diptera:(Intercept)	4.63080695	0.06835276	0.000000e+00	*
Diptera:BlockS	0.26607484	0.08103379	1.025249e-03	*
Diptera:TreatmentM	-0.13176928	0.05747846	2.187690e-02	*
Diptera:TreatmentT	-0.29387404	0.06011325	1.015179e-06	*
Hemiptera:(Intercept)	4.70409886	0.05123311	0.000000e+00	*
Hemiptera:BlockS	1.48764006	0.05260495	6.181386e-176	*
Hemiptera:TreatmentM	0.71708739	0.02698935	1.538112e-155	*
Hemiptera:TreatmentT	-0.03638577	0.03157674	2.491992e-01	
Hymenoptera:(Intercept)	6.89758257	0.02184322	0.000000e+00	*
Hymenoptera:BlockS	-1.12558603	0.03857667	3.700133e-187	*
Hymenoptera:TreatmentM	-0.03385297	0.01892884	7.370609e-02	
Hymenoptera:TreatmentT	-0.21774252	0.01987663	6.308737e-28	*
Lepidoptera:(Intercept)	3.27726339	0.11660782	8.508956e-174	*
Lepidoptera:BlockS	0.29783444	0.13334106	2.550751e-02	*
Lepidoptera:TreatmentM	0.47868675	0.08679290	3.482339e-08	*
Lepidoptera:TreatmentT	0.08040043	0.09456748	3.952189e-01	
Mantodea:(Intercept)	1.17958135	0.42052430	5.031262e-03	*
Mantodea:BlockS	1.29928298	0.46056619	4.786585e-03	*
Mantodea:TreatmentM	-1.40876722	0.33634988	2.809303e-05	*
Mantodea:TreatmentT	-0.51082562	0.24343217	3.586709e-02	*
Neuroptera:(Intercept)	-1.29578035	0.71874835	7.141508e-02	
Neuroptera:BlockS	3.83945231	0.71466964	7.771895e-08	*
Neuroptera:TreatmentM	1.66188439	0.15126270	4.426118e-28	*
Neuroptera:TreatmentT	0.03774033	0.19429176	8.459835e-01	
Orthoptera:(Intercept)	4.23977003	0.09211991	0.000000e+00	*
Orthoptera:BlockS	-0.82848853	0.14429784	9.383210e-09	*
Orthoptera:TreatmentM	-0.45473616	0.09953271	4.907333e-06	*
Orthoptera:TreatmentT	-0.44268782	0.09916742	8.042674e-06	*
Phasmatodea:(Intercept)	2.31305639	0.20679255	4.807583e-29	*
Phasmatodea:BlockS	-0.75030559	0.29428100	1.078395e-02	*
Phasmatodea:TreatmentM	-0.40546511	0.22821773	7.562435e-02	
Phasmatodea:TreatmentT	0.63965850	0.17838818	3.360887e-04	*

This should read like your standard summary output table. `Estimate` gives the estimated coefficient for the covariate-spp pairing described by the row name. `std.error` gives the standard error, `p.value` the Wald test p-value, and `clarity` gets a star when $p < 0.05$. It is a bit hard to pick out patterns staring at a table like this, so lets try visualizing it.

1.4 Visualizing model predictions

```
## first create some fake data
fake.data <- data.frame(Treatment = c("C", "C", "M", "M", "T", "T"),
                        Block = rep(c("N", "S"), times = 3),
                        Replicate = rep("1"), times = 6)

## generate predictions for the fake data
predictions <- predict(abund_model, newdata = fake.data, se.fit = T)

## make data.frame longform for easier plotting
p.data <- do.call("rbind", replicate(11, fake.data, simplify = FALSE))
p.data$species <- rep(colnames(summary@y), each = 6) ## attach species information

## calculate 95% Wald CIs (these are not prediction intervals!!)
p.data$predictions <- matrix(predictions$fitted.values, ncol = 1)
p.data$ci.lower <- matrix(as.matrix(predictions$fitted.values) - 1.97*as.matrix(predict
p.data$ci.upper <- matrix(as.matrix(predictions$fitted.values) + 1.97*as.matrix(predict

## transform from the scale of the linear predictors to the response scale (bit of weir
p.data$tr.predictions <- abund_model@family@linkinv(p.data$predictions)
p.data$tr.ci.lower <- abund_model@family@linkinv(p.data$ci.lower)
p.data$tr.ci.upper <- abund_model@family@linkinv(p.data$ci.upper)

## generate plots
plotlist <- list()
for (spp in unique(p.data$species)) {

  plotlist[[spp]] <- second_axis(ggplot(data = p.data[p.data$Block == "N" & p.data$spec
                                        aes(x = Treatment, y = tr.predictions)) +
                        geom_point(size = 2, color = "#43a2ca") +
                        ylab("predicted abundance") +
                        geom_errorbar(aes(ymin = tr.ci.lower, ymax = tr.ci.upper), size
                        theme_jabo() + ## my custom theme (see /plot_utility.R)
                        theme(legend.position = "none",
                              axis.title = element_blank()) +
                        ggtitle(spp))

}
```

```
## align plots
align_plots(plotlist[[1]], plotlist[[2]], plotlist[[3]],
          plotlist[[4]], plotlist[[5]], plotlist[[6]],
          plotlist[[7]], plotlist[[8]], plotlist[[9]],
          plotlist[[10]], plotlist[[11]], align = c("hv"))

## print plots in a grid
plot_grid(plotlist[[1]], plotlist[[2]], plotlist[[3]],
          plotlist[[4]], plotlist[[5]], plotlist[[6]],
          plotlist[[7]], plotlist[[8]], plotlist[[9]],
          plotlist[[10]], plotlist[[11]])
```

I don't know the ecology well enough to make a nice interpretation of these. Some of them jump out at me as being intuitive though. For example, the more mammal-dependent orders (Diptera and Acari (ticks?)) decrease when mammals are excluded. Orthoptera also decreases, perhaps because the vegatation becomes less grassy? I am probably making up stories here so I will leave it to the people who know better. I think these Confidence intervals are conservative (they are Wald CIs, not corrected ones), so that is probably why the trends appear less clear than the model output table might suggest.