Multivariate analysis Final Report

Ankita

2022-05-22

This document investigates the roles the different environmental variables play in shaping the distribution of the communities and individual taxa Further, analyzes why species differ to environmental response using functional traits of each species. gllvm package is used to analyze multivariate species data.

Packages used

```
# Packages used
#install.packages("mvabund")
#install.packages("gllvm")
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("lmom")
library(lmom)
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.6-2
```

```
library(mvabund)
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
pacman::p_load(data.table, dplyr, gllvm, ggplot2, lattice, magrittr, mvabund, stringr, readxl)
pacman::p_load(ape, Hmsc, data.table, magrittr, dplyr, gllvm, stringr)
```

The data file contains three objects: `macro` containing 51 macroinvetebrates occurences (presence-absence) of taxa across 536 sites `env` containing the 6 environmental variables across 330 sites. The environmental variables (env) includes: max_depth, pH, conductivity, alitude & catchment area. `traits` containing the five

trait variables.The following traits are included: 1. voltinism: semi(< 1 generation/y), uni(1 generation/y)or bi or multi(> 1 generation/y) 2. size : small -1, medium -2, large- 3 3. respiration : Tegument(teg), gills(gil), pls.spi(plastron.spiracle) 4.locomotion(habit): crawl, swim, burrow 5.feed strategy : shredder, predator, herbivore , gatherer, filter, parasite 6.ovip (reproduction): aqu (clutches,free),ovo (ovoviviparity), ter (terrestrial)

Loading the dataset

```
env <- readRDS("environment.rds")
macro <- readRDS("macroinvertebrates.rds")
traits <- readRDS("traits.rds")
```

Selecting/filtering the dataset

```
##using semi_join to print only the rows of ID that have a matching site for species data
macro <- semi_join(macro, env, by = "sample_id")
macro <- macro[order(macro$sample_id),]# sorting species data
macro <- macro[, 2:51]
macro_inv <- data.frame(rbind(macro))
#sorting environment data
env <- env[order(env$sample_id),]
env <- env[,2:7] #remove site ID
env_list <- data.frame(rbind(env)) # converting data.table to data.frame
```

```
#creating new dataframe by combining  3 dataset into one
df_list <- list(macro_inv, env_list, traits)
names(df_list) <- c("macroinv", "env", "traits")

# filtering the species dataset
inv <- colnames(macro)
inv %<>% str_replace_all("\\.", "\\ ") %>%
  str_replace_all("sp\\ .*", "") %>%
  str_trim()
names(df_list$macroinv) = inv
```

#multi dimensional scaling

```
# Compute CCA
#canonical correspondence analysis
spe.ca <- cca(macro ~ max_depth + temperature + conductivity + pH + cat_area + altitude, data= env)
spe.ca
```

```
## Call: cca(formula = macro ~ max_depth + temperature + conductivity + pH
## + cat_area + altitude, data = env)
##
##               Inertia Proportion Rank
## Total          3.3482     1.0000
## Constrained    0.3437     0.1027    6
## Unconstrained  3.0044     0.8973   49
## Inertia is scaled Chi-square
##
## Eigenvalues for constrained axes:
##    CCA1    CCA2    CCA3    CCA4    CCA5    CCA6
## 0.23369 0.04682 0.03470 0.01615 0.00733 0.00505
##
## Eigenvalues for unconstrained axes:
##     CA1     CA2     CA3     CA4     CA5     CA6     CA7     CA8
```

```
## 0.20039 0.16855 0.15281 0.13492 0.12939 0.11032 0.10291 0.09669
## (Showing 8 of 49 unconstrained eigenvalues)
```

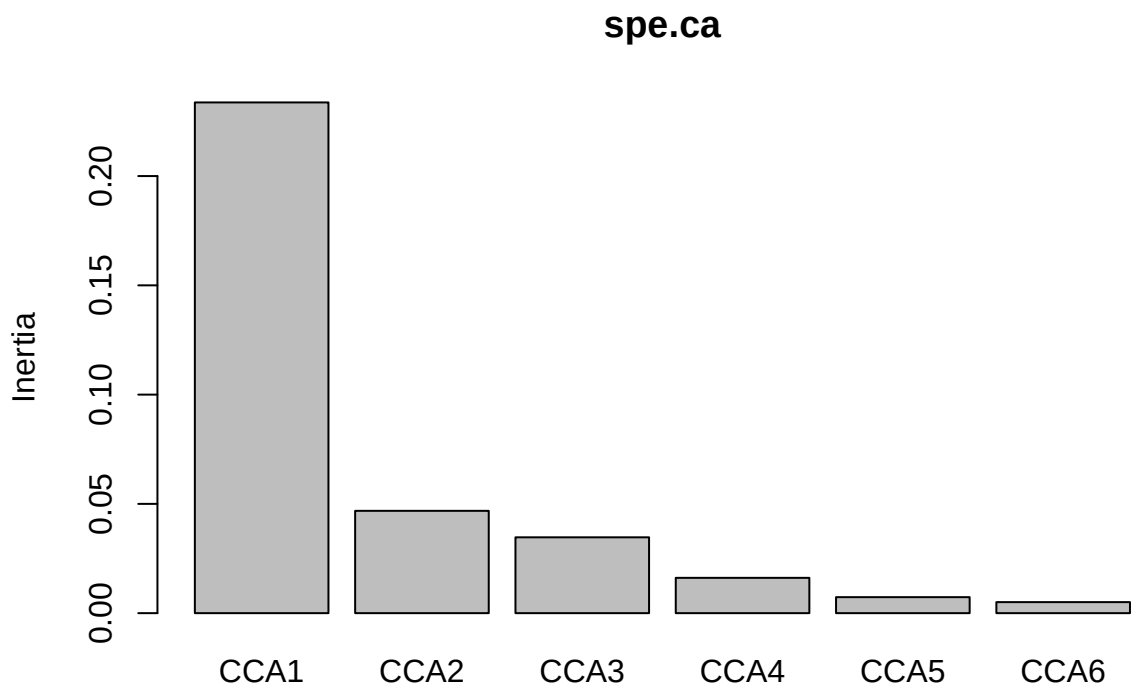
```
# anova #check whether CCA ordination is significant
anova(spe.ca)
```

```
## Permutation test for cca under reduced model
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = macro ~ max_depth + temperature + conductivity + pH + cat_area + altitude, data
##          Df ChiSquare      F Pr(>F)
## Model     6   0.34375 6.1592  0.001 ***
## Residual 323   3.00445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(spe.ca, by="axis") # axes are significant
```

```
## Permutation test for cca under reduced model
## Forward tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = macro ~ max_depth + temperature + conductivity + pH + cat_area + altitude, data
##          Df ChiSquare       F Pr(>F)
## CCA1      1   0.23369 25.1239  0.001 ***
## CCA2      1   0.04682  5.0333  0.003 **
## CCA3      1   0.03470  3.7309  0.017 *
## CCA4      1   0.01615  1.7365  0.194
## CCA5      1   0.00733  0.7877  0.872
## CCA6      1   0.00505  0.5429  0.961
## Residual 323   3.00445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

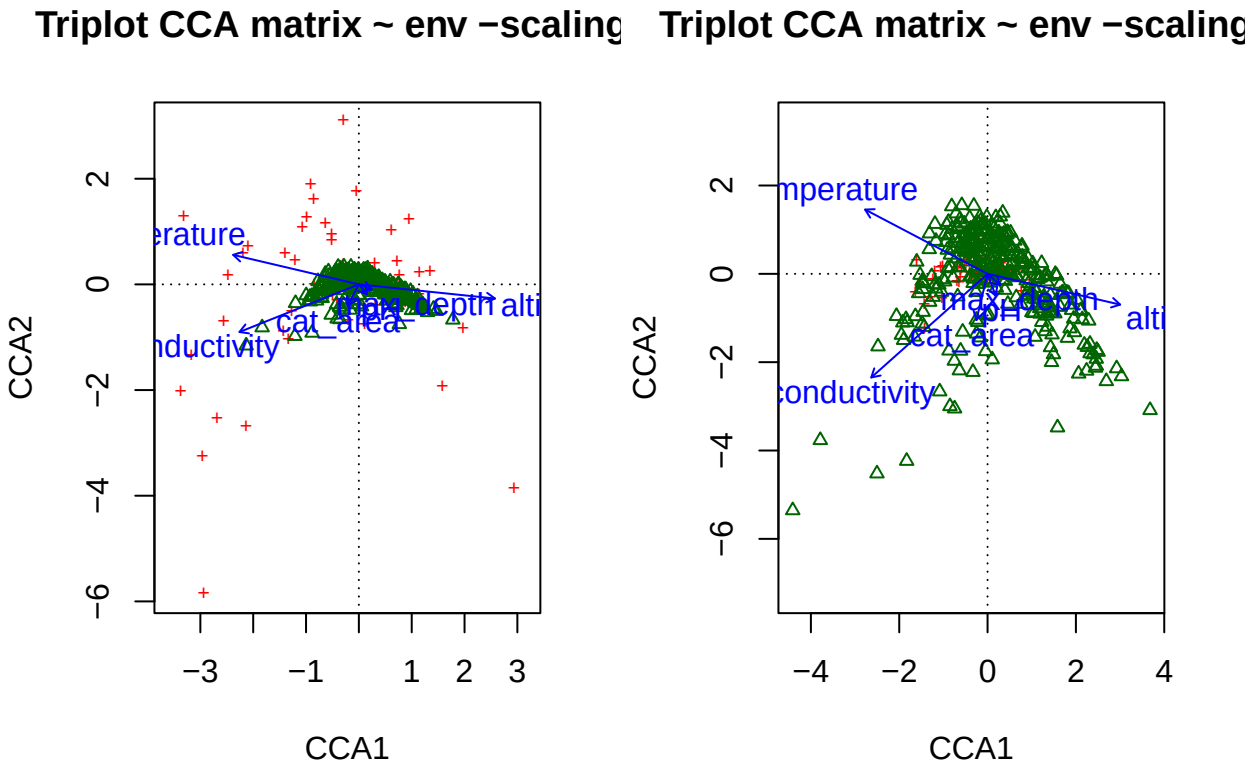
```
# plotting
screeplot(spe.ca)
```




```
# Plotting eigenvalues and % of variance for each axis
(ev2 <- spe.ca$CA$eig)
```

```
##         CA1         CA2         CA3         CA4         CA5         CA6
## 0.200387230 0.168551964 0.152814236 0.134920172 0.129386022 0.110315095
##         CA7         CA8         CA9        CA10        CA11        CA12
## 0.102907395 0.096691643 0.094487742 0.090226826 0.084203665 0.079682599
##        CA13        CA14        CA15        CA16        CA17        CA18
## 0.077208389 0.076028334 0.074038437 0.071251625 0.069535972 0.064570546
##        CA19        CA20        CA21        CA22        CA23        CA24
## 0.063851820 0.059879267 0.058354287 0.054775478 0.054076527 0.051144925
##        CA25        CA26        CA27        CA28        CA29        CA30
## 0.050994809 0.049404188 0.047724064 0.046971576 0.044069782 0.043105035
##        CA31        CA32        CA33        CA34        CA35        CA36
## 0.039115441 0.038282709 0.037190225 0.035832447 0.033730206 0.032848034
##        CA37        CA38        CA39        CA40        CA41        CA42
## 0.031379347 0.029210011 0.027749433 0.026213089 0.024907212 0.023643752
##        CA43        CA44        CA45        CA46        CA47        CA48
## 0.023037356 0.021955052 0.020893355 0.017938813 0.015954556 0.013156821
##        CA49
## 0.009848182
```

```
# CA biplots
par(mfrow=c(1,2))
plot(spe.ca, scaling=1, display=c('sp', 'lc', 'cn'), main='Triplot CCA matrix ~ env -scaling 1')
plot(spe.ca, display=c('sp', 'lc', 'cn'), main="Triplot CCA matrix ~ env -scaling 2")
```



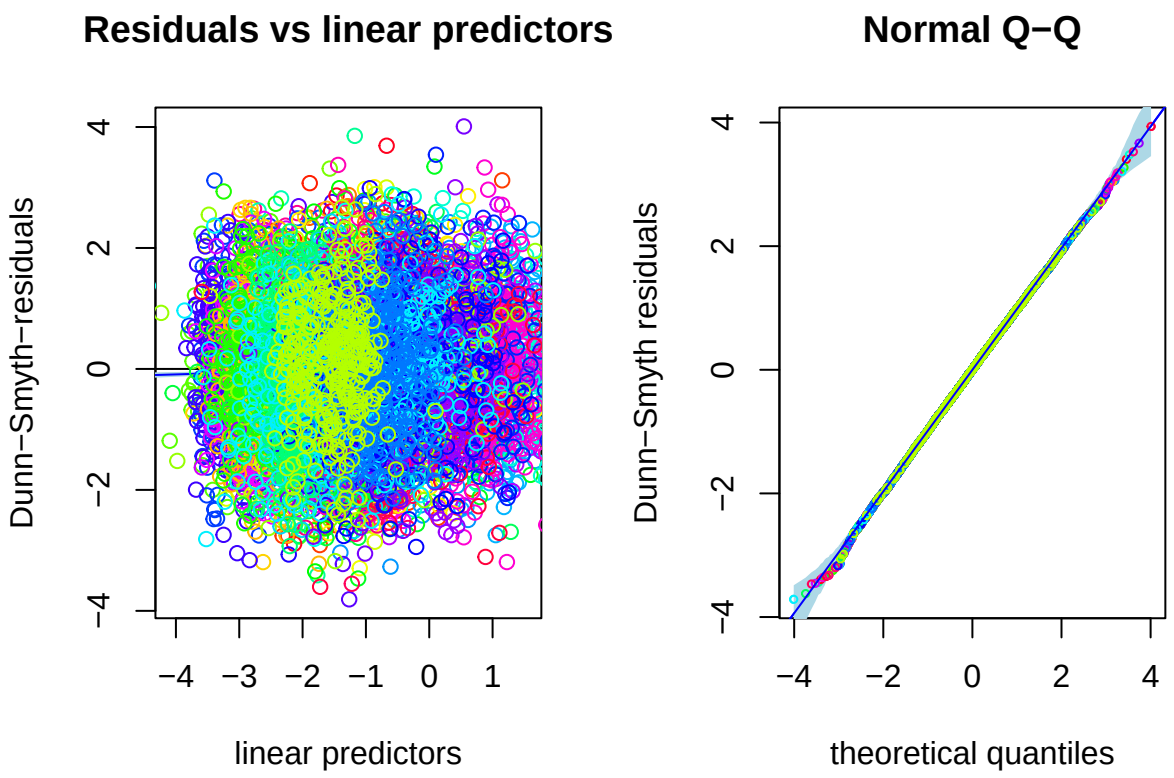

```
#conductivity, temperature and altitude most influential

# A posteriori projection of environmental variables in a CA
(spe.ca.env <- envfit(spe.ca, env))
```

```
##
## ***VECTORS
##
##                  CCA1     CCA2     r2 Pr(>r)
## cat_area     -0.42322 -0.90603 0.0327  0.007 **
## altitude      0.99313 -0.11706 0.4700  0.001 ***
## max_depth     0.90732 -0.42044 0.0057  0.343
## conductivity -0.90405 -0.42743 0.5050  0.001 ***
## pH            0.56958 -0.82194 0.0079  0.226
## temperature  -0.96475  0.26316 0.4440  0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

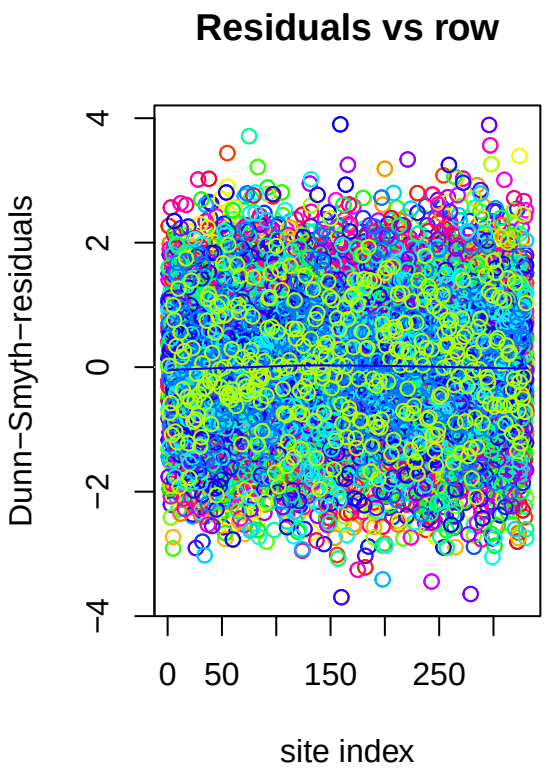
Multivariate analysis using gllvm package: We considered binomial approach as response variable in macroinvertebrates data is binary. Distribution was estimated using variational approximation method (method = "VA"), link function probit.

```
#Model based ordination

library(mvabund)
```

```
library(gllvm)

y <- as.matrix(df_list$macroinv)
x <- scale(as.matrix(df_list$env))
TR <-df_list$traits


fit_bin <- gllvm(y, family="binomial", method = "VA",link = "probit")

##plotting residuals for the Binomial model
par(mfrow = c(1,2))
plot(fit_bin, which = 1)
plot(fit_bin, which = 2)
```

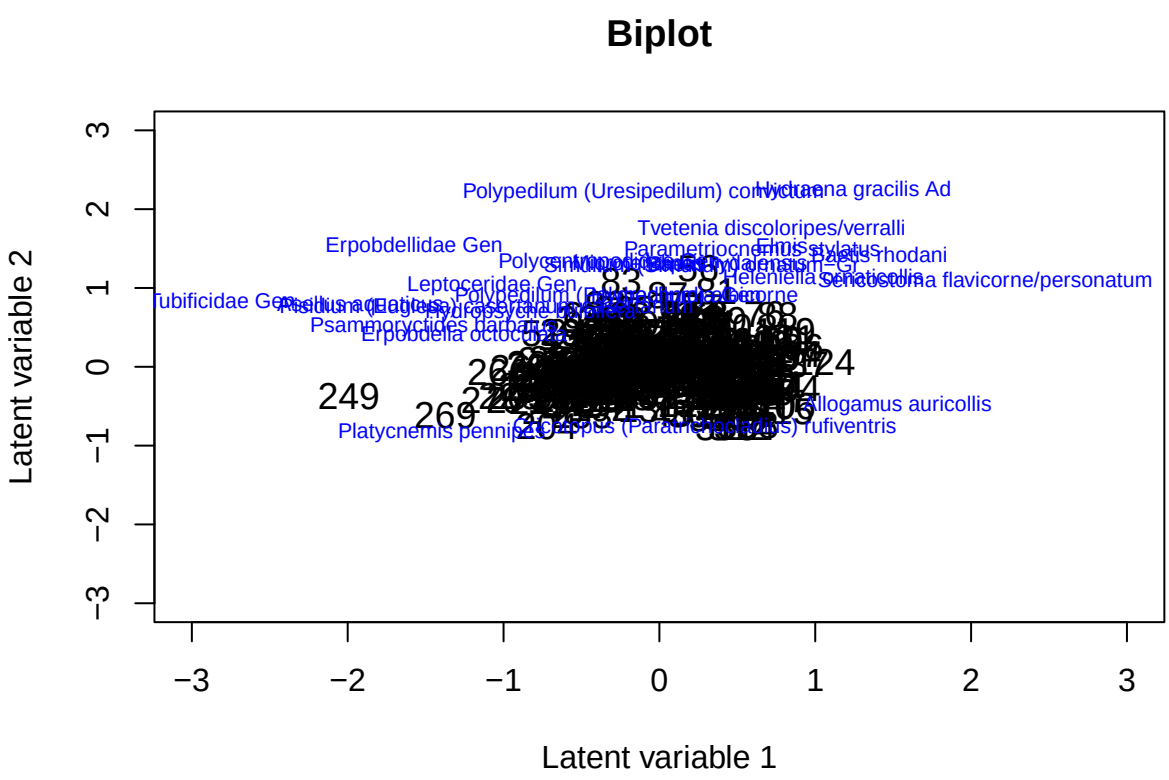



```
plot(fit_bin, which = 3) #information criteria suggests binomial as good fit
```

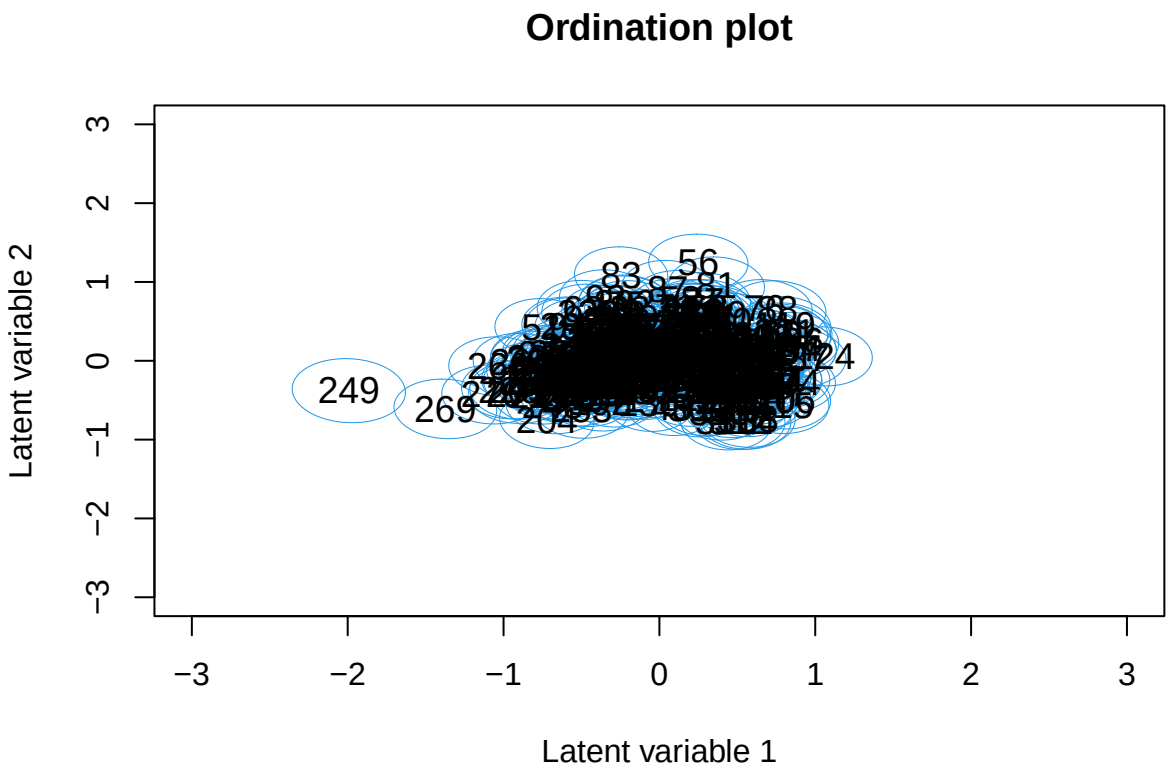


Using ordiplot() function to construct an ordination as a scatter plot of predicted latent variables.

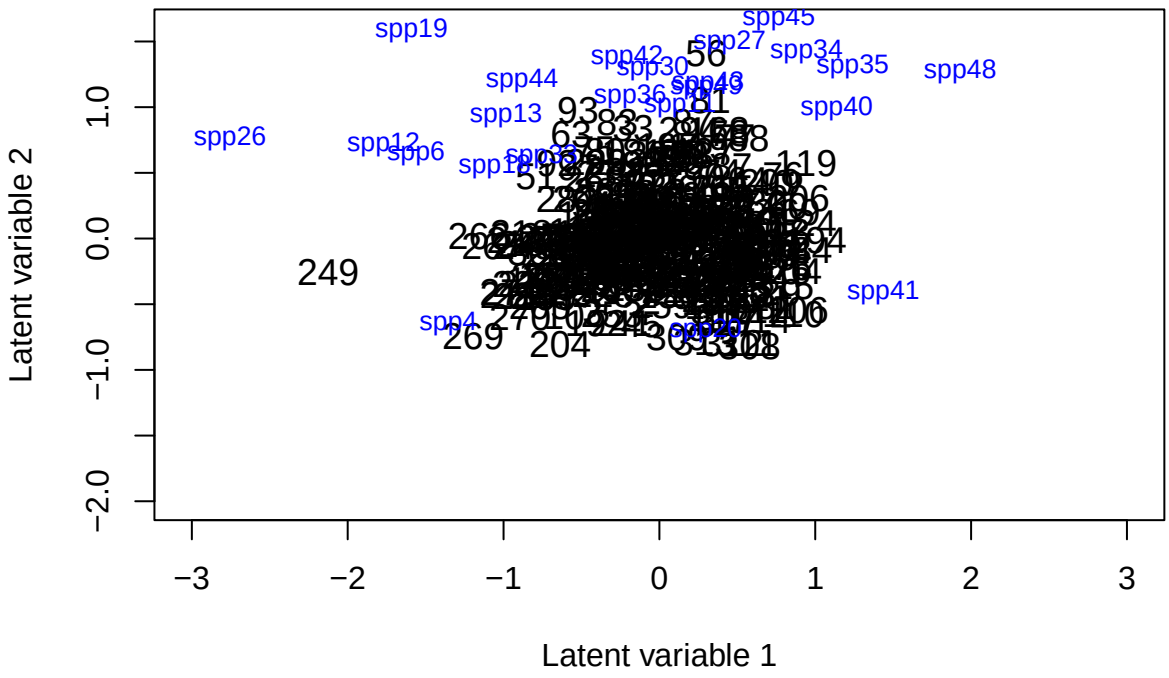
```
ordiplot.gllvm(fit_bin, biplot = TRUE,ind.spp= 25, xlim = c(-3, 3), ylim = c(-3, 3), arrow.scale = 0.8,
         main = "Biplot")
```




```
ordiplot.gllvm(fit_bin, biplot = FALSE, ind.spp = 25, xlim = c(-3, 3), ylim = c(-3, 3),
         main = "Ordination plot", predict.region = TRUE)
```




```
rownames(fit_bin$params$theta) <- paste("spp", 1:ncol(fit_bin$y), sep = "") #replacing long species nam
ordiplot.gllvm(fit_bin, biplot = TRUE, ind.spp = 25, xlim = c(-3, 3), ylim = c(-2, 1.6),
         main = "Biplot", jitter = TRUE, cex.spp = 0.8)
```



The graph shows larger cluster of sites in the center with very few indicator species whereas most indicator species are seen in the further side from larger cluster of sites.

#Modelling with environment variable Using residual analysis and information criteria to study which distribution offers the most suitable mean-variance relationship for the responses, and how many latent variables are needed.

```
criterias <- NULL
for(i in 0:5){
  fiti <- gllvm(y, x, family = "binomial", num.lv = i, sd.errors = FALSE,
                formula = ~ temperature + conductivity+ altitude, seed = 1234)
  criterias[i + 1] <- summary(fiti)$AICc
  names(criterias)[i + 1] = i
}

criterias
```

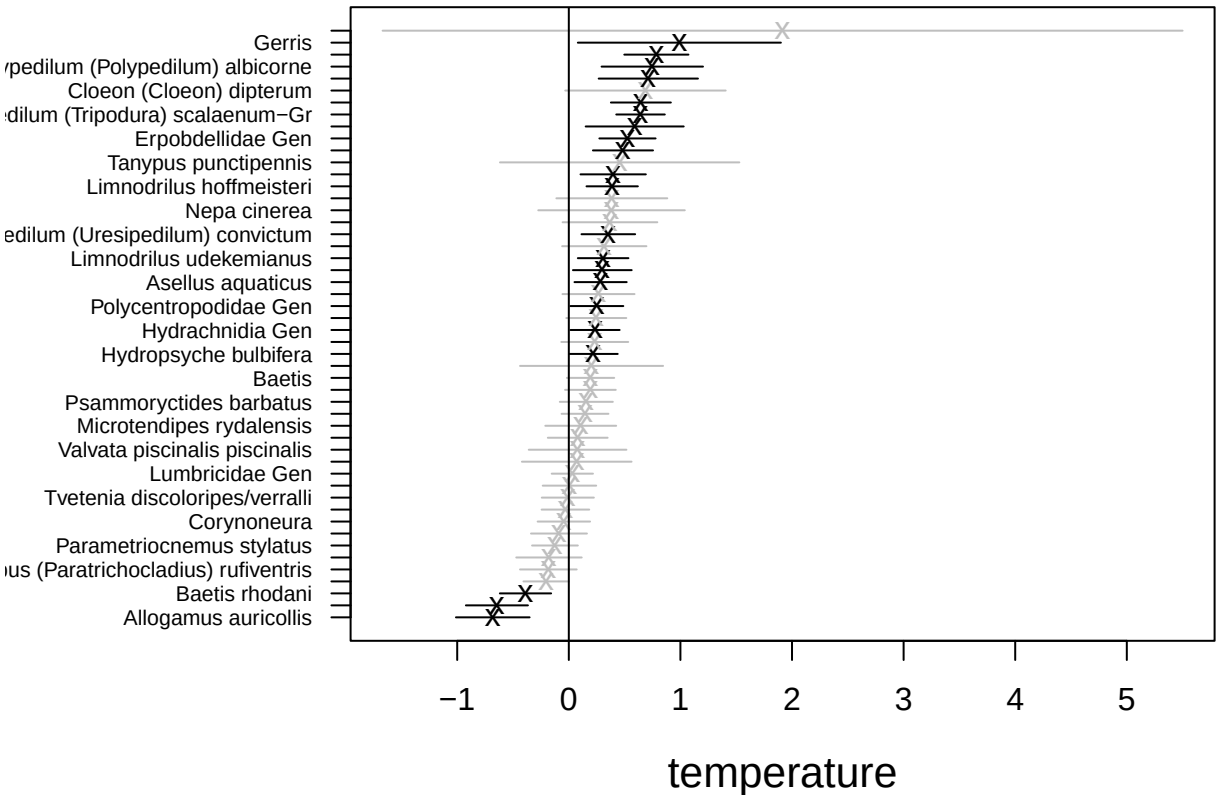
```
##        0        1        2        3        4        5
## 11642.72 11488.22 11444.42 11544.32 11642.67 11739.51
```

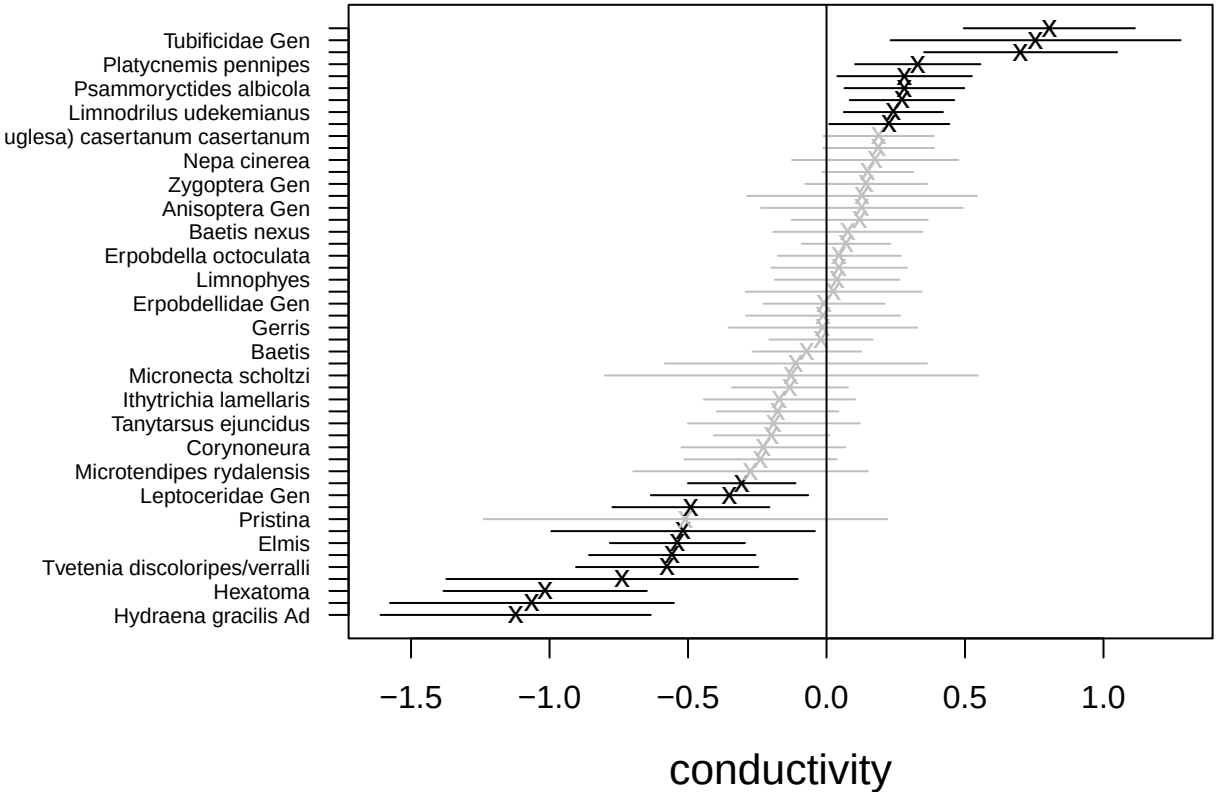
Based on AICc values, 2 Latent variables was chosen

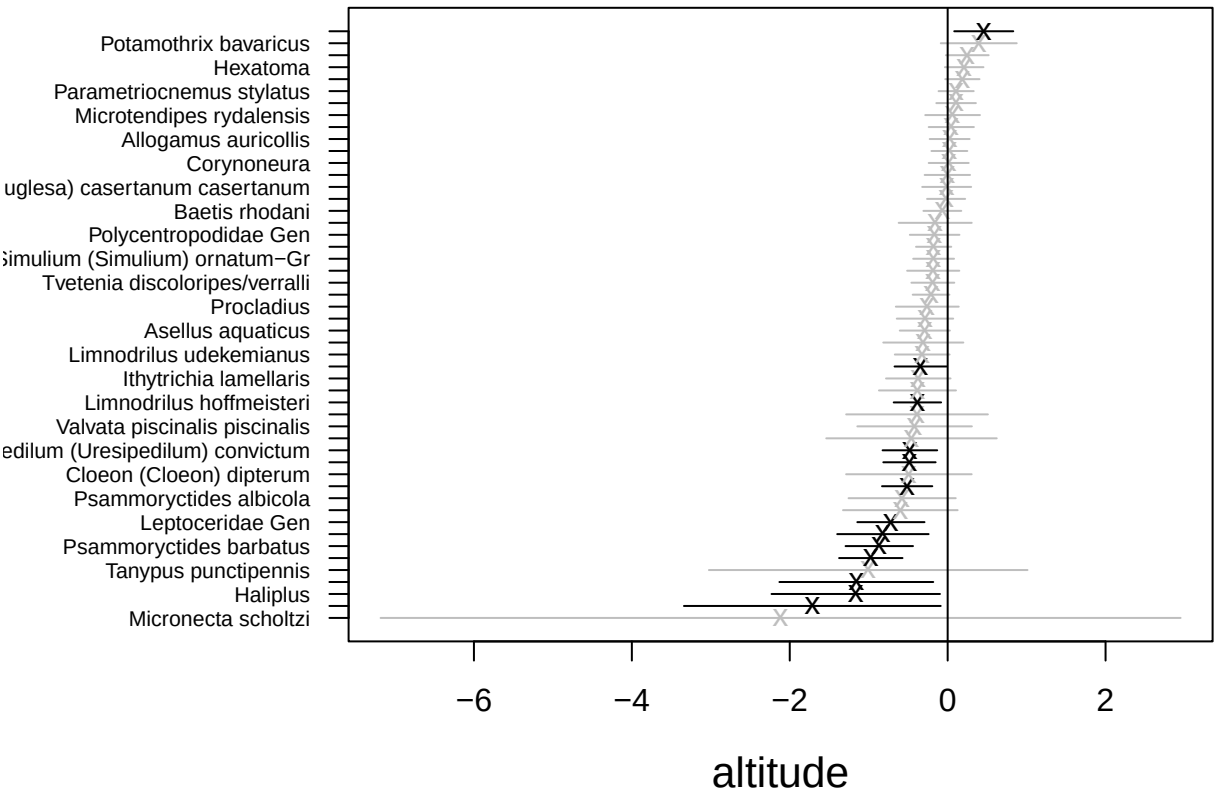
```
#gllvm with environmental variables
fit_bin_env <- gllvm(y, x,
                     formula = ~ temperature + conductivity+ altitude,
                     family="binomial")

AIC(fit_bin_env)
```

```
## [1] 11433.35
```

```
#coefplot plots
coefplot(fit_bin_env, cex.ylab = 0.7, mar = c(4, 9, 2, 1),
         xlim.list = NULL, mfrow=c(1,1))
```

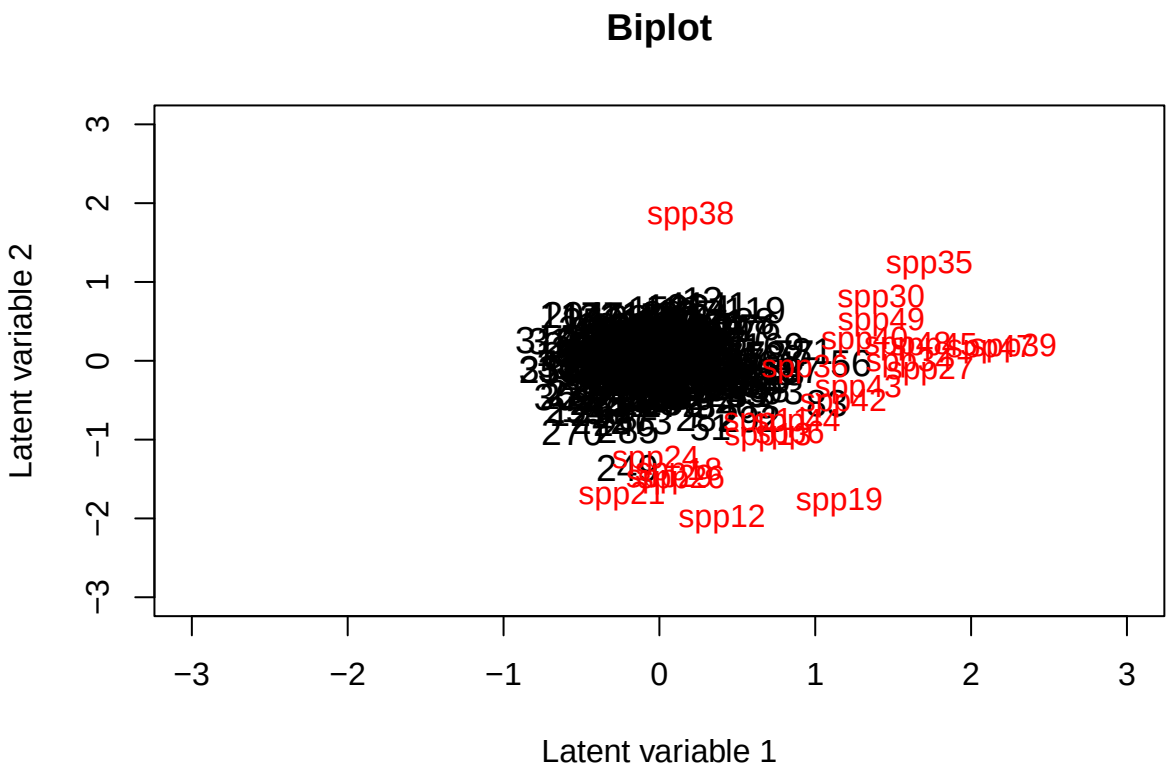



Tubificidae Gen
Platycnemis pennipes
Psammoryctides albicola
Limnodrilus udekemianus
uglesa) casertanum casertanum
Nepa cinerea
Zygoptera Gen
Anisoptera Gen
Baetis nexus
Erpobdella octoculata
Limnophyes
Erpobdellidae Gen
Gerris
Baetis
Micronecta scholtzi
Ithytrichia lamellaris
Tanytarsus ejuncidus
Corynoneura
Microtendipes rydalensis
Leptoceridae Gen
Pristina
Elmis
Tvetenia discoloripes/verralli
Hexatoma
Hydraena gracilis Ad
−1.5
−1.0
−0.5
0.0
0.5
1.0
conductivity



Plots of the point estimates for coefficients of the environmental variables and their 95% confidence intervals (lines) with those coloured in grey denotes intervals containing zero and black containing non zero values. In the resulting plot, 95% confidence intervals has less zero values for 3 environment variable indicating evidence of association between environment and species abundance.

```
#ordiplot
rownames(fit_bin_env$params$theta) <- paste("spp", 1:ncol(fit_bin_env$y), sep = "")

ordiplot.gllvm(fit_bin_env, biplot = TRUE,ind.spp= 25, xlim = c(-3, 3), ylim = c(-3, 3),
               spp.colors= "red", s.colors="black", cex.env= 0.007, cex.spp= 1,
             main = "Biplot")
```



In biplots we can see that Species 27 (Parametriocnemus stylatus), 43 (Microtendipes rydalensis), 45 (Tvetenia verralli) and 48(Sericostoma flavicorne) are close to each other. This can also be seen in the correlation plots as species correlations are positive.

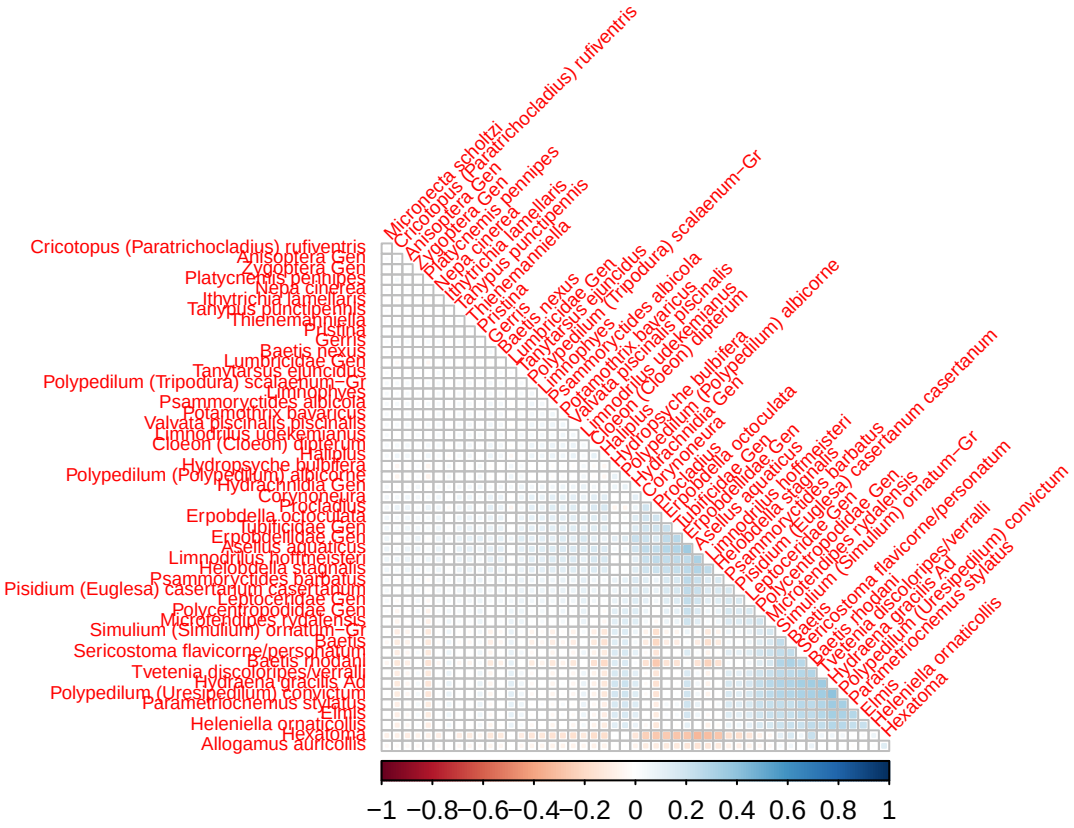
```
# Residual correlation matrix (using getResidulaCOr() to estimate correlation matrix of linear predicto
cr <- getResidualCor(fit_bin_env)
library(corrplot); library(gclus)
```

```
## corrplot 0.92 loaded
```

```
## Loading required package: cluster
```

```
## Registered S3 method overwritten by 'gclus':
##   method         from
##   reorder.hclust vegan
```

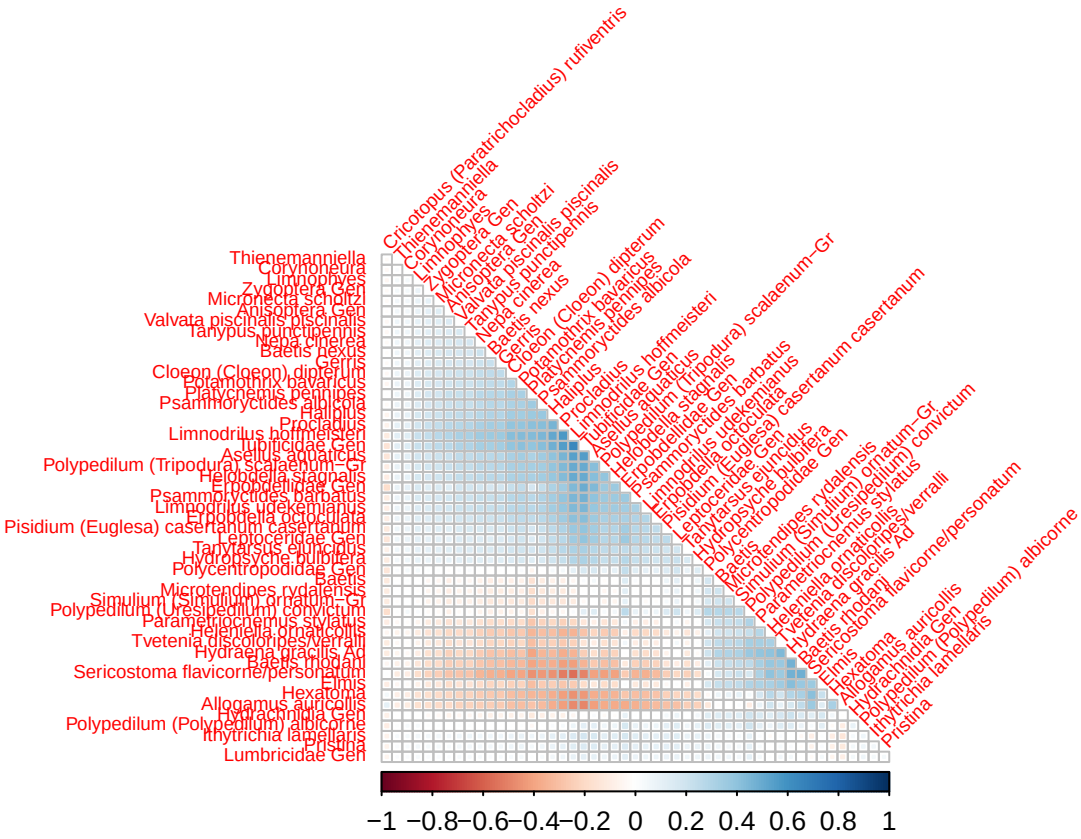
```
corrplot(cr[order.single(cr), order.single(cr)], diag = FALSE, type = "lower",
         method = "square", tl.cex = 0.6, tl.srt = 45, tl.col = "red")
```




```
#Regions coloured in blue in correlation plot indicate clusters of species that are positively correlat
```

The correlation is not seen as significant as indicated by the color of the boxes. There are only two regions colored in light blue, indicating positive correlation between pairs of species and just one regions colored in faint red indicating negative correlation between species.

```
# Using GLLVM without environmental variables and 2 latent variables
#refering to fit_bin
# Correlation matrix
cr0 <- getResidualCor(fit_bin)
corrplot(cr0[order.single(cr0), order.single(cr0)], diag = FALSE, type = "lower",
         method = "square", tl.cex = 0.6, tl.srt = 45, tl.col = "red")
```



The species correlations can be significantly observed with the correlation matrix without environment variables with darker blue and red observations indicating strong correlation.

Quantifying the amount of variation in the data that can be explained by environmental variables

```
res_env <- getResidualCov(fit_bin_env)
res_lv <- getResidualCov(fit_bin)
1- (res_env$trace/ res_lv$trace)
```

```
## [1] 0.1732914
```

Ratio of traces suggest environment variable explain 17.3% of covariation in macroinvetebrates species

Determining which species are in both the trait and plot datasets

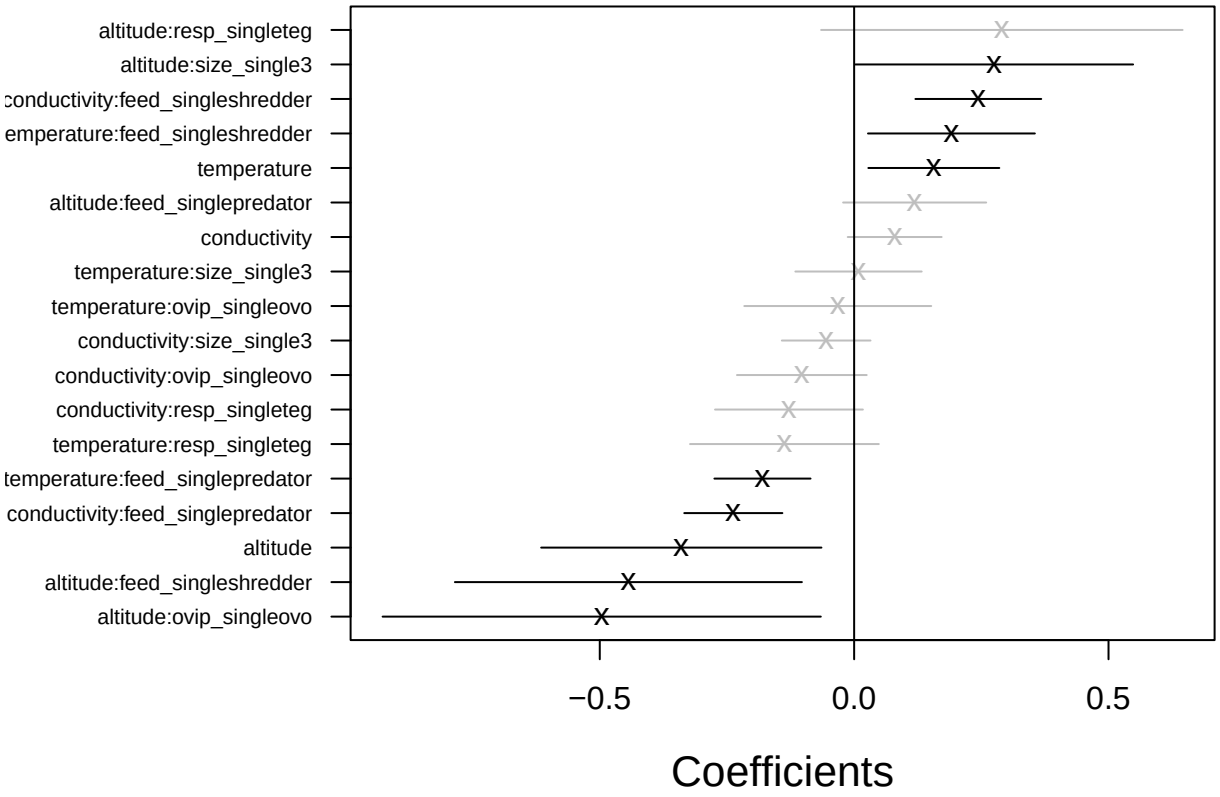
```
# difference in length of row names(species) in traits and column name(species) in macroinvertebrates
# adding column names to the data frame
df1_new<-as.data.frame(t(macro))
wordsFreq <- data.frame(species = rownames(traits), traits, row.names= NULL)
macro_1 <- data.frame(species= rownames(df1_new), df1_new=df1_new, row.names= NULL)

# finding the intersection of both column name vectors
cols_intersection <- intersect(wordsFreq$species, macro_1$species)
cols_intersection # 6 species identified in both data set
```

```
## [1] "Haliplus sp."          "Asellus aquaticus"     "Helobdella stagnalis"
## [4] "Baetis sp."            "Elmis sp."
```

```
## Extracting only those species with traits from data.frames
df_list$macroinv <- macro %>%
  select(`Asellus aquaticus`, `Haliplus sp.`, `Elmis sp.`, `Baetis sp.`, `Helobdella stagnalis`)
df_list$traits <- semi_join(wordsFreq, macro_1, by = "species")
df_list$traits %<>% select(-species)
```

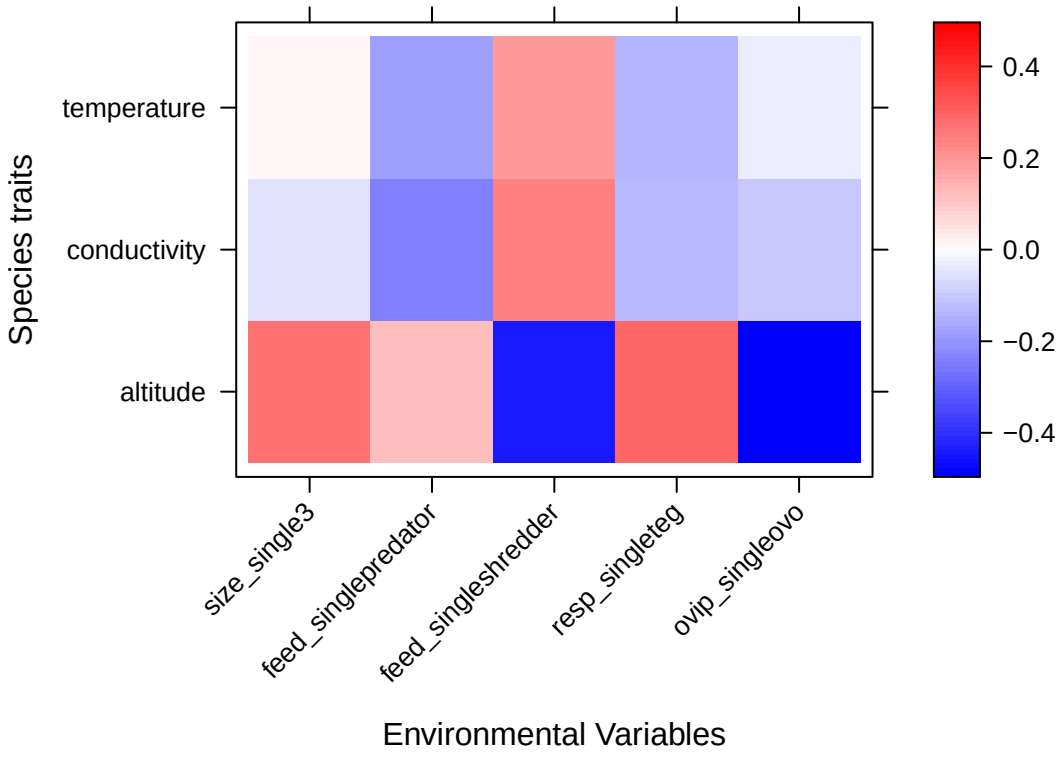
Incorporating traits into fourth corner models

```
y <- as.matrix(df_list$macroinv)
x <- scale(as.matrix(df_list$env))
TR <-df_list$traits

# Fitting fourth corner model with two latent variables
fit_4th <- gllvm( y = y, X= x, TR = TR,
  family = "binomial",
  num.lv = 2,
  formula = y ~
    (conductivity + temperature + altitude) +
                  (conductivity + temperature + altitude ) : (ovip_single + resp_single + feed_single +

#Plotting
coefplot(fit_4th, cex.ylab = 0.7, mar = c(4, 9, 2, 1),
         xlim.list = NULL, mfrow=c(1,1))
```




```
fourth = fit_4th$fourth.corner
#library(lattice)
a = max( abs(fourth) )
```

```
colort = colorRampPalette(c("blue","white","red"))
plot.4th = levelplot(t(as.matrix(fourth)), xlab = "Environmental Variables",
 ylab = "Species traits", col.regions = colort(100),
at = seq( -a, a, length = 100), scales = list( x = list(rot = 45)))
print(plot.4th)
```



The resulting plots indicate that interactions of the trait variable. The strongest negative interactions were observed between altitude and feeding stage(shredder), as well as between altitude and reproduction (ovivoparity) indicating higher altitude not supporting shredder species.The strongest positive effects occurred in interactions between altitude and size and respiration (tegument).It indicates that the species with tegument might be more resistant to higher altitude area(with lower CO2) and larger size species can thrive in higher altitude. Moreover, there is a positive correlation between conductivity and feeding where higher level of conductivity supports feeding of shredder species.

Using likelihood ratio test to see if traits vary with environment

```
fit_4th2 <- gllvm(y, x, TR, family = "binomial", num.lv = 2,
        formula = y ~ (conductivity + temperature + altitude))

# Test interactions using likelihood ratio test:
anova(fit_4th, fit_4th2)
```

```
## Model  1 :  y ~ (conductivity + temperature + altitude)
## Model  2 :  y ~ (conductivity + temperature + altitude) + (conductivity + temperature + altitude):(o
```

```
##   Resid.Df        D Df.diff     P.value
## 1     1633  0.00000       0
## 2     1618 89.44023      15 1.26132e-12
```

p-value suggests that the model where there is no strong evidence of traits mediating the environmental response of species.