Final Project - EEB590C

Devin Molnau and Elizabeth McMurchie

April 29, 2021

Homework 3

This assignment is due prior to the last day class. You are to self-select and work in groups: 2-3 in a group. For the assignment below submit one R-script. Annotations via comments are highly encouraged. The script should run!

Assignment:

1: Obtain a dataset. This may be one of your own, a dataset from DRYAD, or some other source. Identify hypotheses for this dataset and analyze patterns in the data. You may use any methods learned during the semester, but at least one analysis must come from material learned in weeks 11-13.

USE COMMENTS IN THE R CODE to describe what the patterns you find represent.

```r
setwd("D:/Documents/BoxSync/Classes_Spring2021/Advanced_Biostatistics_EEB590C/Homework/Final_EEB590")
#load appropriate libraries
library(knitr)
library(RRPP)
library(geomorph)
library(tidyverse)
library(readxl)
library(ade4)
library(vegan)
library(mice)
library(ggplot2)
```

Intro

Epidermal micromorphology in Neotropical bamboo foliage leaves has proven to be a useful tool for differentiating between different species. Certain unusual features have been noted in the epidermal micromorphology of one species of savanna bamboo, Guadua paniculata, including papillae on both leaf surfaces and saddle-shaped silica cells. However, it remains unknown whether the few other species of Guadua primarily found in forests, including G. virgata, have similar micromorphology. Additionally, there are several general trends in Guadua macromorphology, including the G. angustifolia type, which are tall, primarily forest-adapted species, the G. glomerata type, which are scandent species primarily found along rivers, the G. sarcocarpa type, which are similar to G. angustifolia but notable for their tree-killing ability, and species of both savanna and forest bamboos that closely resemble G. paniculata. Here, we analyzed whether micromorphological features such as shape of silica bodies and presence, placement, and shape of stomata and papillae on epidermal cells of foliage leaves were associated with different patterns in macromorphology, habitat type, and country of origin of specimens belonging to a selection of species of Guadua.

Preprocessing

Read in data

```
#READ IN DATA AS DATAFRAME
mydata<-read_excel(path="data/TransformGuaduaSet.xlsx", col_names = TRUE, na="x")
```

Remove Column V and Y

Remove column V and Y, because there is missing data and we cannot impute the data because it just doesn't make sense to attempt to predict these columns.

V. Adaxial: Frequency of stomates if present on the adaxial surface of foliage leaf blades: 0 = common; 1 = infrequent. Y. Adaxial: Papillae on long cells of the intercostal zone adjacent to the stomates: 0 = not overarching the stomates; 1 = overarching the stomates.

```
to_drop<-c("Adaxial: Papillae on long cells of the intercostal zone adjacent to the stomates:  0 = not 

df=data.frame(mydata[,!( names(mydata) %in% to_drop)],stringsAsFactors = TRUE)
```

Set columns to factors

Sets all columns to factors and reads them in as a dataframe.

```
df<-data.frame(lapply(df,as.factor))
```

Imputation of missing data

In the last three columns, in some species the papillae of the cells adjacent to the stomata obscured theshape of the subsidiary cells. Therefore, the shapes of these subsidiary cells resulted in missing data that we imputed using the MICE package's logistic regression method.

```
init = mice(df,maxit=0)
meth = init$method
predM=init$predictorMatrix
meth[c(colnames(df[,7:length(df)]))]<-"logreg"
set.seed(183)
imputed<-mice(df,method = meth, predictorMatrix = predM, m=5,printFlag = FALSE)
```

```
## Warning: Number of logged events: 90
```

```
imputed<-complete(imputed)
```

To quick visualize this and make sure that there are no NA left we can use sum(is.na()) in sapply().

```
check_is_boolean<-sapply(imputed,function(x) sum(is.na(x))) #double check to make sure there are no mor
```

Analysis

Correlation

```
Y<-imputed[,7:ncol(imputed)] # reads in the binary data only
Y<-data.frame(lapply(Y,function(x) as.numeric(levels(x))[x]))
correlation_y<-data.frame(cor(Y)) #calculated the correlation of the Y values
```

PCOA

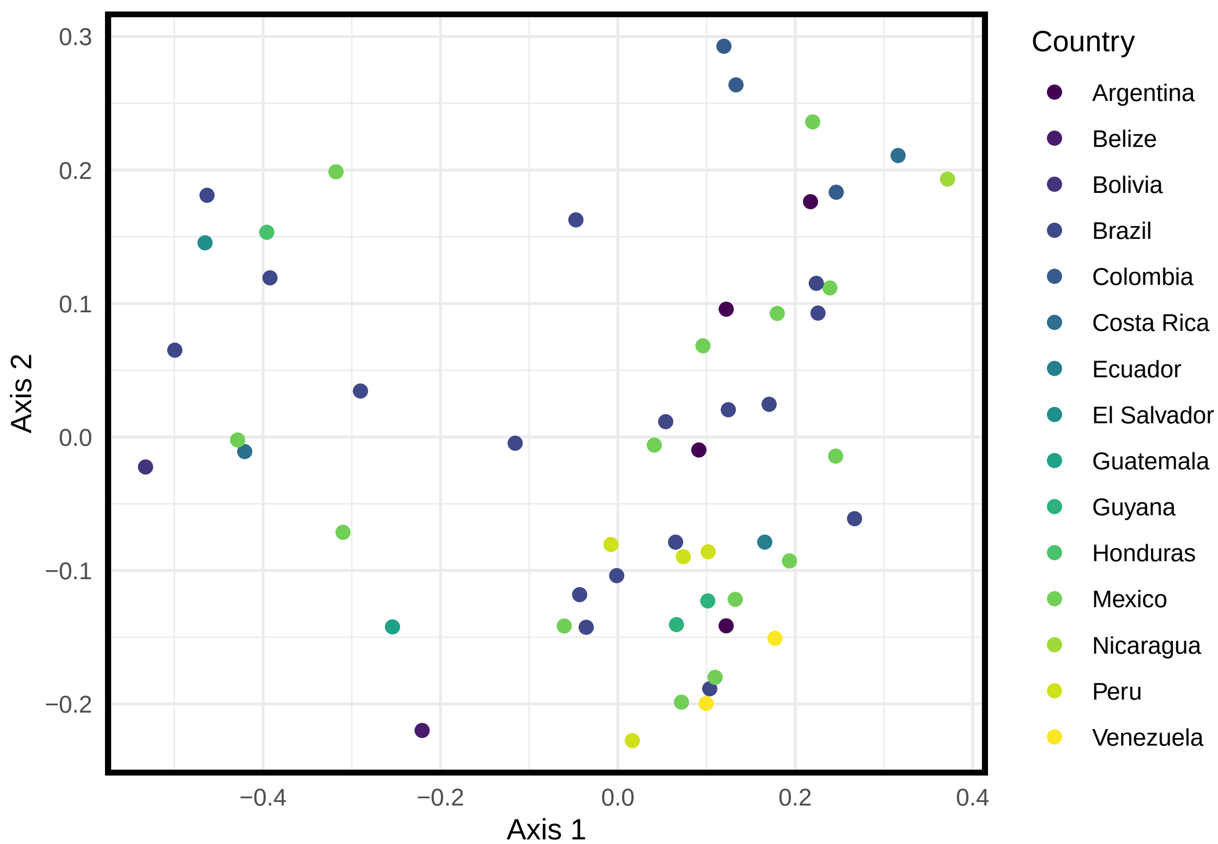
We get the distance matrix for our binary data using simple matching coefficient. Then we are able to calculate the PCoA on the distance matrix.

```r
#Distance matrix for binary is simple matching coefficient
Y.dist<-dist.binary(Y, method=2, diag = FALSE, upper = FALSE)
# Y.dist
Y.dist.matrix<-as.matrix(Y.dist)

PCoA<-cmdscale(Y.dist.matrix)    #from vegan
```

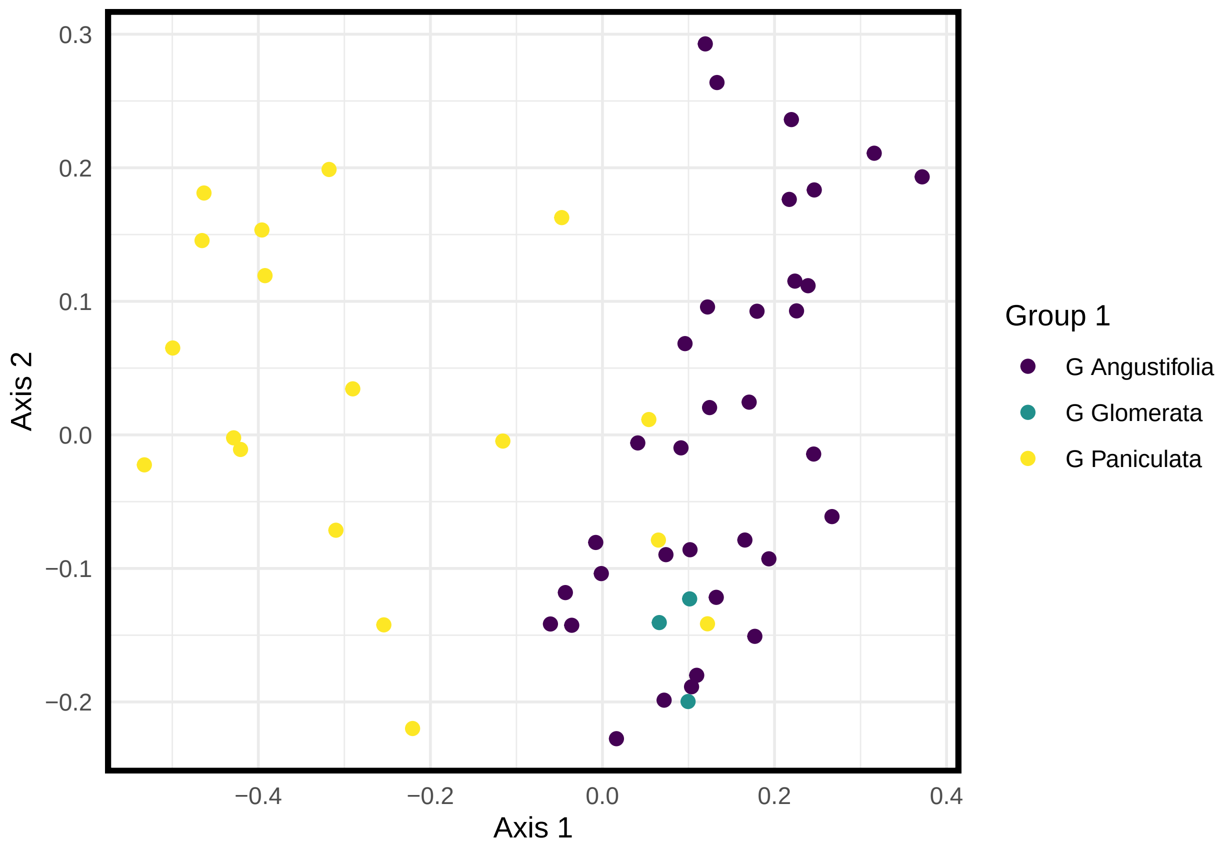
PCoA grouped by Country

```r
nice_df <- PCoA %>%
  data.frame
nice_df$country <- df$Country
nice_df %>%
  ggplot(aes(X1, X2)) +
  geom_point(aes(color = country), size = 2) +
  scale_color_viridis_d() +
  theme_minimal() +
  theme(
    panel.border = element_rect(size = 2, color = "black", fill = NA)
  ) +
  labs(
    color = "Country",
    x = "Axis 1",
    y = "Axis 2"
  )
```



PCoA grouped by Group1

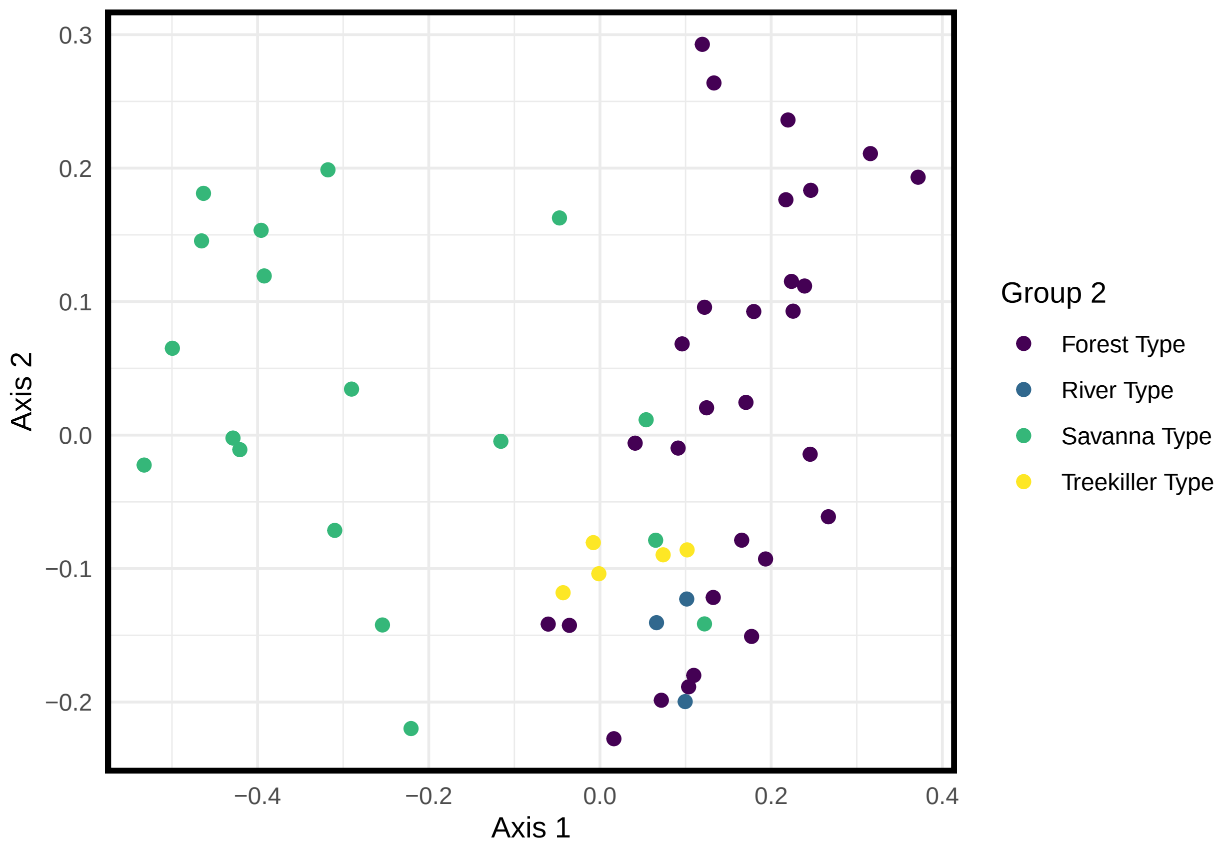
```
nice_df$Group_1<- df$Group_1
nice_df %>%
  mutate(
    Group_1 = str_to_title(str_replace_all(Group_1, "_", " "))
  ) %>%
  ggplot(aes(X1, X2)) +
  geom_point(aes(color = Group_1), size = 2) +
  scale_color_viridis_d() +
  theme_minimal() +
  theme(
    panel.border = element_rect(size = 2, color = "black", fill = NA)
  ) +
  labs(
    color = "Group 1",
    x = "Axis 1",
    y = "Axis 2"
  )
```



PCoA grouped by Group 2

```
nice_df$Group_2<- df$Group_2
nice_df %>%
  mutate(
    Group_2 = str_to_title(str_replace_all(Group_2, "_", " "))
  ) %>%
```

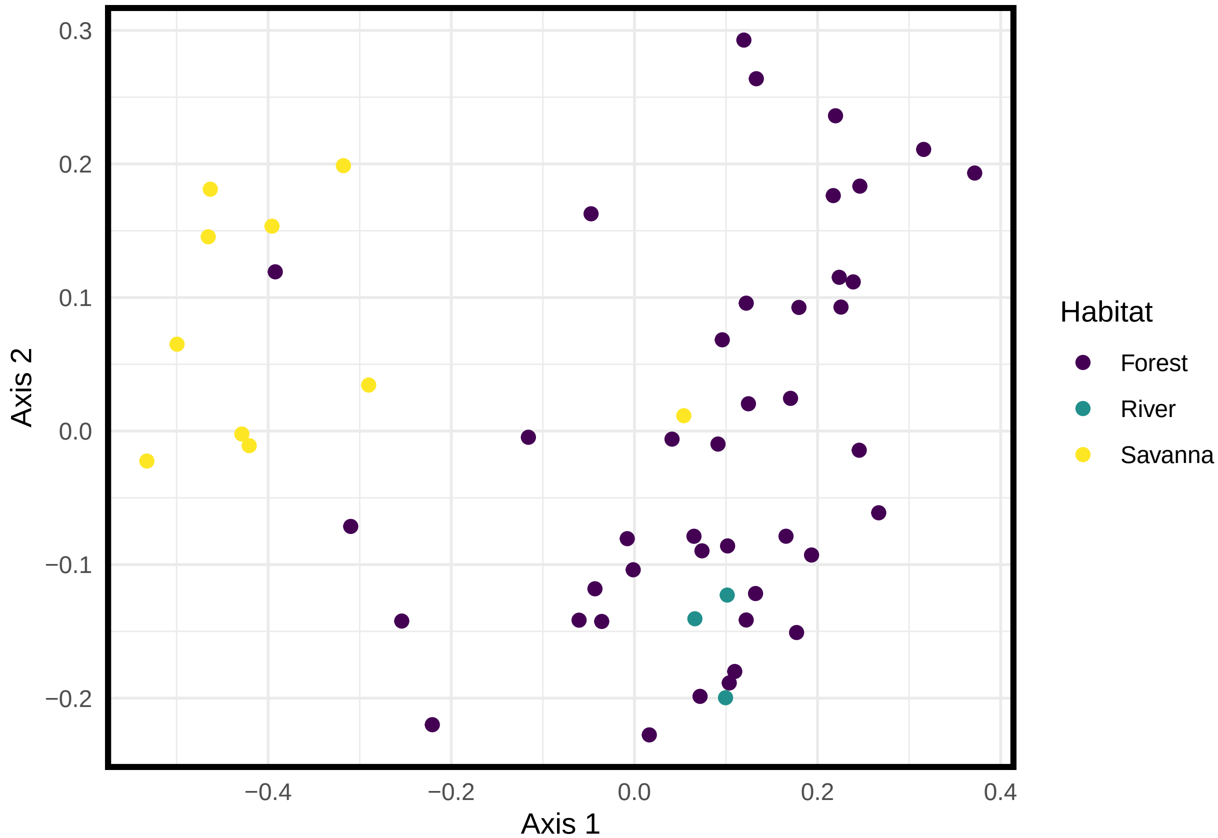
```r
  ggplot(aes(X1, X2)) +
  geom_point(aes(color = Group_2), size = 2) +
  scale_color_viridis_d() +
  theme_minimal() +
  theme(
    panel.border = element_rect(size = 2, color = "black", fill = NA)
  ) +
  labs(
    color = "Group 2",
    x = "Axis 1",
    y = "Axis 2"
  )
```



PCoA grouped by Habitat

```r
nice_df$Habitat<- df$Habitat
nice_df %>%
  mutate(
    Habitat = str_to_title(str_replace_all(Habitat, "_", " "))
  ) %>%
  ggplot(aes(X1, X2)) +
  geom_point(aes(color = Habitat), size = 2) +
  scale_color_viridis_d() +
  theme_minimal() +
  theme(
    panel.border = element_rect(size = 2, color = "black", fill = NA)
```

```
  ) +
  labs(
    color = "Habitat",
    x = "Axis 1",
    y = "Axis 2"
  )
```



Broken Stick Plot

To look at what compromises the PC1 variance we look at the rotation. The values that are farther away form 0 are more important for the PC.

CLUSTERING METHODS?

FACTORIAL MANOVA

```
#Factorial MANOVA via RRPP

# Y<-imputed[,7:ncol(imputed)]
# Y<-data.frame(lapply(Y,function(x) as.numeric(levels(x))[x]))

mydat<-rrpp.data.frame("Y"=Y,
                       "Group_2"= as.factor(imputed$Group_2),
                       "Habitat"=as.factor(imputed$Habitat),
                       "Country" = as.factor(imputed$Country))
```

```
model2.rrpp <- lm.rrpp(Y.dist.matrix ~ mydat$Group_2 * mydat$Habitat * mydat$Country, print.progress = F
```

```
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 180
## Final X columns (rank): 24
## Check coefficients or degrees of freedom in ANOVA to see changes.
```

```
anova(model2.rrpp)
```

```
##
## Analysis of Variance, using Residual Randomization
## Permutation procedure: Randomization of null model residuals
## Number of permutations: 1000
## Estimation method: Ordinary Least Squares
## Sums of Squares and Cross-products: Type I
## Effect sizes (Z) based on F distributions
##
##                               Df      SS      MS     Rsq      F       Z Pr(>F)
## mydat$Group_2                  3  2.9888 0.99626 0.28615 7.2974  6.2057  0.001 **
## mydat$Habitat                  1  0.5926 0.59259 0.05674 4.3406  4.5377  0.001 **
## mydat$Country                 14  1.9162 0.13687 0.18346 1.0026  0.0009  0.495
## mydat$Group_2:mydat$Country    4  0.6422 0.16054 0.06148 1.1760  0.7106  0.247
## mydat$Habitat:mydat$Country    1  0.0729 0.07292 0.00698 0.5341 -0.8921  0.797
## Residuals                     31  4.2322 0.13652 0.40519
## Total                         54 10.4448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call: lm.rrpp(f1 = Y.dist.matrix ~ mydat$Group_2 * mydat$Habitat *
##     mydat$Country, print.progress = FALSE)
```

POSTHOC ANALYSIS

Which groups as significantly different from one another?

```
# TODO
# RRPP does have a pairwise posthoc anaylsis to look at
#pairwise comparisons via RRPP
#posthoc <- pairwise(fit = model2.rrpp , groups = mydat$Group_2, print.progress = FALSE)
#summary(posthoc)
```

MODEL COMPARISON USING LIKELIHOOD RATIO TEST (LTR)

Setup

```
Y.group2<-lm.rrpp(Y.dist.matrix ~ mydat$Group_2,
                  print.progress=FALSE)
Y.habitat<-lm.rrpp(Y.dist.matrix ~ mydat$Habitat,
                   print.progress=FALSE)
Y.country<-lm.rrpp(Y.dist.matrix ~ mydat$Country,
                   print.progress=FALSE)
Y.group2.habitat<-lm.rrpp(Y.dist.matrix ~ mydat$Group_2 + mydat$Habitat,
                          print.progress=FALSE)
```

```
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 6
## Final X columns (rank): 5
## Check coefficients or degrees of freedom in ANOVA to see changes.
```

```
Y.group2.by.country<-lm.rrpp(Y.dist.matrix ~ mydat$Group_2 + mydat$Country,
                             print.progress=FALSE)
Y.habitat.country<-lm.rrpp(Y.dist.matrix ~ mydat$Habitat + mydat$Country,
                           print.progress=FALSE)
Y.habitat.by.country<-lm.rrpp(Y.dist.matrix ~mydat$Group_2 * mydat$Habitat,
                              print.progress=FALSE)
```

```
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 12
## Final X columns (rank): 5
## Check coefficients or degrees of freedom in ANOVA to see changes.
```

```
Y.mancova<-lm.rrpp(Y.dist.matrix ~mydat$Group_2 + mydat$Habitat*mydat$Country,
                   print.progress=FALSE)
```

```
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 48
## Final X columns (rank): 21
## Check coefficients or degrees of freedom in ANOVA to see changes.
```

```
Y.mancova_2<-lm.rrpp(Y.dist.matrix ~mydat$Habitat + mydat$Group_2*mydat$Country,
                     print.progress=FALSE)
```

```
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 62
## Final X columns (rank): 23
## Check coefficients or degrees of freedom in ANOVA to see changes.
```

```
Y.full<-lm.rrpp(Y.dist.matrix ~ mydat$Group_2 * mydat$Habitat * mydat$Country,
                print.progress=FALSE)
```

```
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 180
## Final X columns (rank): 24
## Check coefficients or degrees of freedom in ANOVA to see changes.
```

```
#anova(Y.full)
# anova(Y.group2.habitat,Y.group2)
# anova(Y.group2.habitat,Y.habitat)
# anova(Y.mancova,Y.group2.habitat)

#Error in anova.mlm(Y.full) : residuals have rank 31 < 55
```

Would AIC fix the problem we are having? Is this a dimensionality problem?

RRPP MODEL COMPARISON

```
?RRPP::model.comparison()
```

```
## starting httpd help server ... done
```

```
modelComp1<-model.comparison(Y.full,
                              Y.mancova,
                              Y.mancova_2,
                              Y.habitat.by.country,
                              Y.habitat.country, Y.habitat,
                              Y.group2.by.country,
                              Y.group2.habitat,
                              Y.group2,
                              Y.country,
                              type = "logLik", tol=0.01)
```

```
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 180
## Final X columns (rank): 24
## Check coefficients or degrees of freedom in ANOVA to see changes.
##
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 48
## Final X columns (rank): 21
## Check coefficients or degrees of freedom in ANOVA to see changes.
##
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 62
## Final X columns (rank): 23
## Check coefficients or degrees of freedom in ANOVA to see changes.
##
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 12
## Final X columns (rank): 5
## Check coefficients or degrees of freedom in ANOVA to see changes.
##
##
## Warning: Because variables in the linear model are redundant,
## the linear model design has been truncated (via QR decomposition).
## Original X columns: 6
## Final X columns (rank): 5
## Check coefficients or degrees of freedom in ANOVA to see changes.
```

```
modelComp1.summ<-as.data.frame(summary(modelComp1))
```

```
##
##
##  Summary statistics for model log-likelihoods
##
## 10 Models compared.
##
##
## mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Group_2:mydat$Habitat + mydat$Group_2:mydat$Cou
## mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Habitat:mydat$Country
## mydat$Habitat + mydat$Group_2 + mydat$Country + mydat$Group_2:mydat$Country
## mydat$Group_2 + mydat$Habitat + mydat$Group_2:mydat$Habitat
## mydat$Habitat + mydat$Country
## mydat$Habitat
## mydat$Group_2 + mydat$Country
## mydat$Group_2 + mydat$Habitat
## mydat$Group_2
## mydat$Country
##
## mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Group_2:mydat$Habitat + mydat$Group_2:mydat$Cou
## mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Habitat:mydat$Country
## mydat$Habitat + mydat$Group_2 + mydat$Country + mydat$Group_2:mydat$Country
## mydat$Group_2 + mydat$Habitat + mydat$Group_2:mydat$Habitat
## mydat$Habitat + mydat$Country
## mydat$Habitat
## mydat$Group_2 + mydat$Country
## mydat$Group_2 + mydat$Habitat
## mydat$Group_2
## mydat$Country
##
## mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Group_2:mydat$Habitat + mydat$Group_2:mydat$Cou
## mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Habitat:mydat$Country
## mydat$Habitat + mydat$Group_2 + mydat$Country + mydat$Group_2:mydat$Country
## mydat$Group_2 + mydat$Habitat + mydat$Group_2:mydat$Habitat
## mydat$Habitat + mydat$Country
## mydat$Habitat
## mydat$Group_2 + mydat$Country
## mydat$Group_2 + mydat$Habitat
## mydat$Group_2
## mydat$Country
##
## mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Group_2:mydat$Habitat + mydat$Group_2:mydat$Cou
## mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Habitat:mydat$Country
## mydat$Habitat + mydat$Group_2 + mydat$Country + mydat$Group_2:mydat$Country
## mydat$Group_2 + mydat$Habitat + mydat$Group_2:mydat$Habitat
## mydat$Habitat + mydat$Country
## mydat$Habitat
## mydat$Group_2 + mydat$Country
## mydat$Group_2 + mydat$Habitat
## mydat$Group_2
## mydat$Country
```