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.

```
#load appropriate libraries
library(knitr)
library(formatR)
library(RRPP)
library(geomorph)
library(tidyverse)
library(readxl)
library(ade4)
library(vegan)
library(mice)
library(ggplot2)
library(pander)
```

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.

We hypothesized that we would see a significant difference in the epidermal micromorphology of Guadua foliage leaves when categorized by country of origin, macromorphological group, and habitat.

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

We removed column V and Y, because there is missing data and in the case of columns of V and Y it doesn't make sense to attempt to predict these columns.When stomates are not present on the adaxial surface, we cannot estimate stomatal frequency.Similarly, when papillae are not present on long cells of the intercostal zone adjacent to the papillae, we cannot impute whether they overarch the stomates or not.

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
           "Adaxial: Frequency of stomates if present on the adaxial surface of foliage leaf blades:  0

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 the shape 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,
              rintFlag = FALSE)
imputed<-complete(imputed)
```

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

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

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

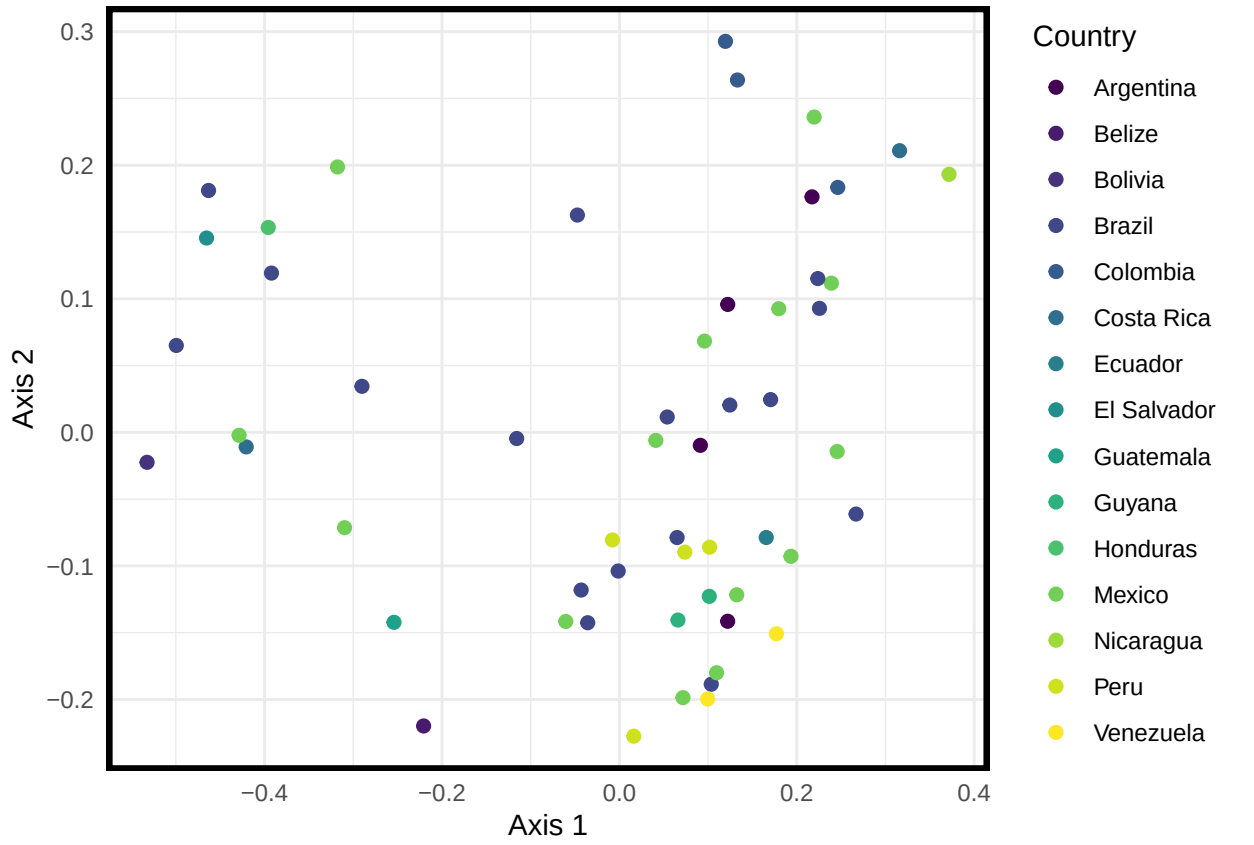
We get the distance matrix for our binary data using simple matching coefficient. Then we are able to calculate the PCoA on the distance matrix to visualize the pairwise distances between individual Guadua specimen epidermal micromorphology.

```
#Distance matrix for binary is simple matching coefficient
Y.dist<-dist.binary(Y, method=2, diag = FALSE, upper = FALSE)
Y.dist.matrix<-as.matrix(Y.dist)
PCoA<-cmdscale(Y.dist.matrix, eig= TRUE, x.ret=TRUE, list. = TRUE)    #from vegan
```

PCoA grouped by Country

To visualize the whether specimens of Guadua appear to group according to country of origin based on the pairwise differences seen between specimen epidermal micromorphology, we labeled individual specimen coordinates by the country in which the specimens were collected. As many of these specimens are old (sometimes upwards of 80 years old), locational information beyond country is often unavailable. However, if geographic distance, often related to allopatric speciation, is an important determinant of epidermal micromorphology, we suspected that we might see pairwise distances indicative of grouping by country. However, we did not observe this in our results.

```
nice_df <- PCoA$points %>%
  data.frame
nice_df$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"
  )
```

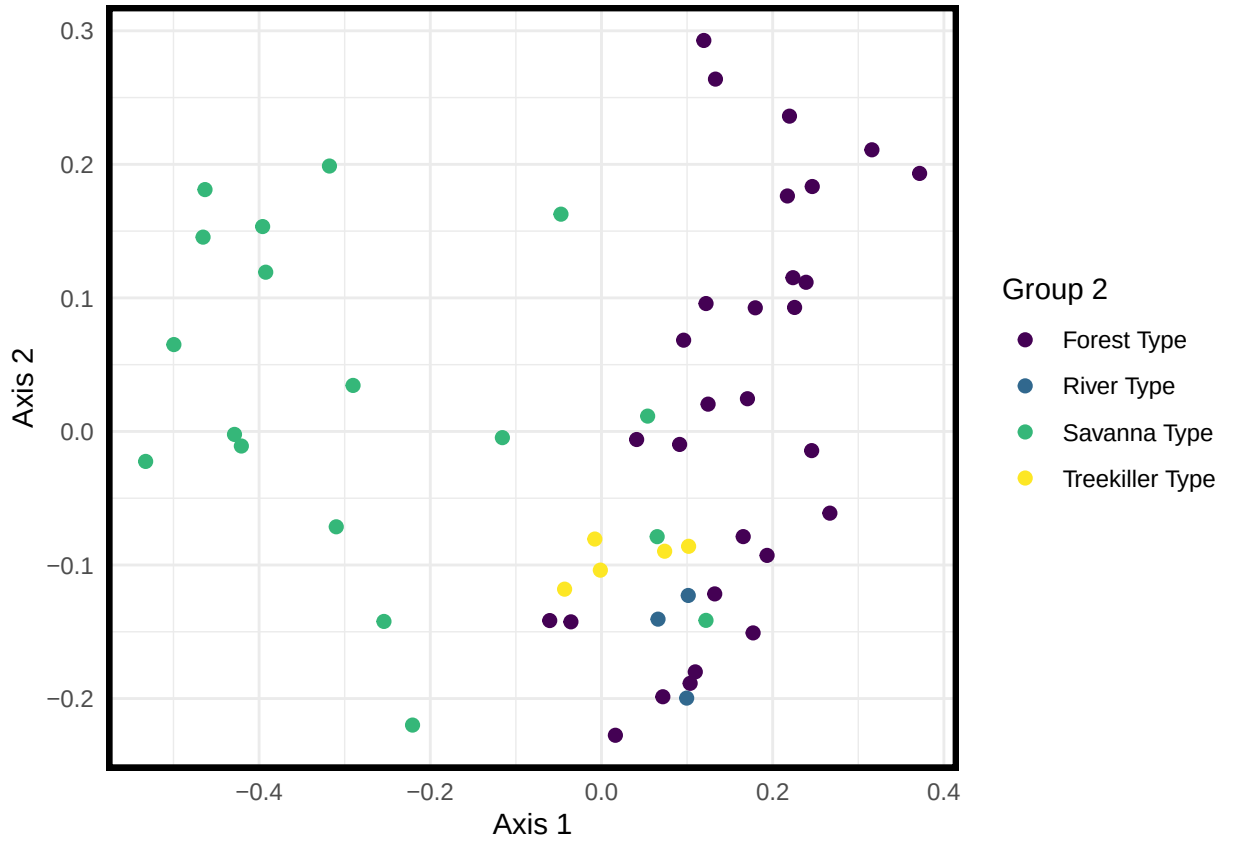


PCoA grouped by Group 2

Species of Guadua can be categorized by their macromorphology, which roughly aligns with habitat. Species of the "savanna type" tend to be under 4 m tall, have culms of less than 4 cm in diameter, and have narrow foliage leaves. Those of the "forest type" tend to be over 4 m tall, with erect culms 4 cm in diameter and wide foliage leaves. Guadua of the "treekiller type" are similar to those of the "forest type" but lean on trees, weighing them down and often breaking and killing them, while having longer rhizomes. "River type" Guadua live along rivers and have relatively small-diameter, solid, scandent culms that trail over surrounding plants. To visualize the whether specimens of Guadua appear to group according to these macromorphological types based on the pairwise differences seen between specimen epidermal micromorphology, we labeled individual specimen coordinates by macromorphological groups in which we categorized them. We expected that we might see grouping based on these different macromorphology types, and here observed that there appear to be roughly two groups, one including "savanna type" species and the other including the other three: "forest", "river', and "treekiller". Although the "treekiller" and "river" types appear to group closely with each other, they fall within the main group that includes these three types.

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

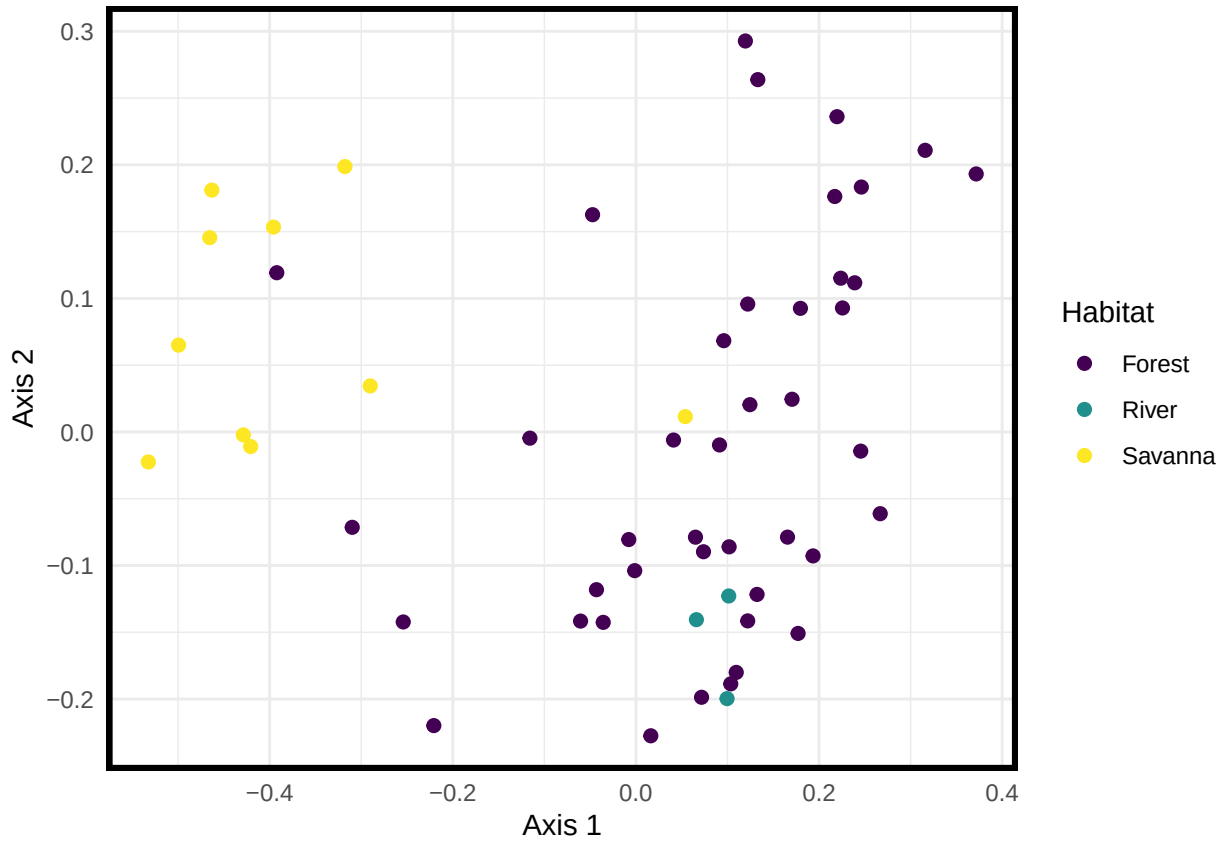


PCoA grouped by Habitat

Differences in epidermal micromorphology based on habitat have been observed in other genera of Neotropical woody bamboos. Here, we sorted Guadua into three primary habitat types: savanna, forest, and river. Note that some species with "savanna type" macromorphology are more commonly found in forests, and classified accordingly. To visualize the whether specimens of Guadua appear to group according to habitat based on the pairwise differences seen between specimen epidermal micromorphology, we labeled individual specimen coordinates by the habitat in which they are most commonly observed. We expected that we might see grouping based on these different habitat, and here observed that there appear to be roughly two groups: one of savanna species, and another of forest and river species. Although the river species appear to group together, they do not group separately from others.

```
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"
  )
```



Scree Plot

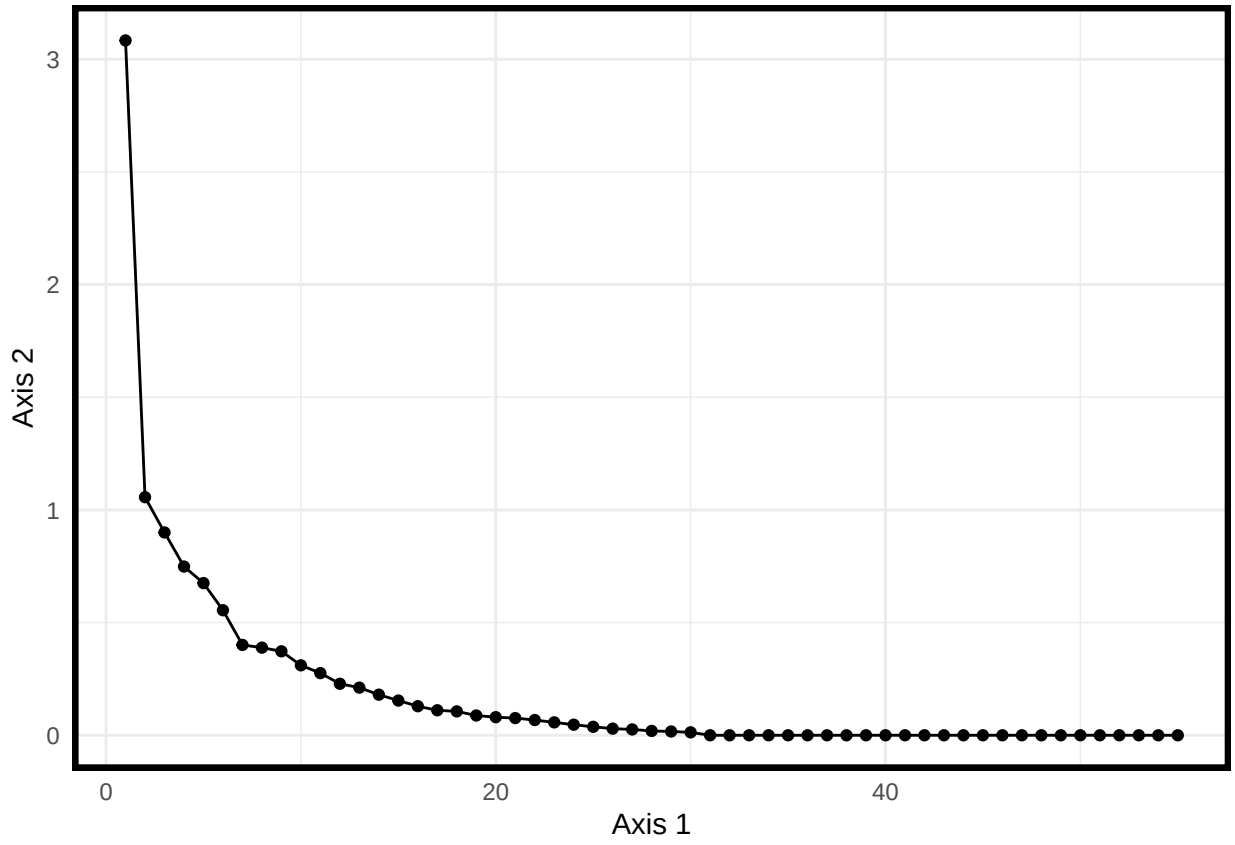
From the PCoA we are able to obtain eigenvalues and then plot them to show the percent variation explained along each PCoA axis. As seen in the plot below, PCoA axis 1 is responsible for the majority of variation in the PCoA, followed by a steep drop in variation. By the 20th axis, variation is arguably negligible.

```
eig_df <- data.frame( x_values = c(1:length(PCoA$eig)) , eig_value=c(PCoA$eig))



ggplot(data=eig_df, aes(x=x_values, y=eig_value))+
  geom_line() +
  geom_point() +
  theme_minimal() +
```

```
  theme(
    panel.border = element_rect(size = 2, color = "black", fill = NA)
  ) +
  labs(
      x = "Axis 1",
    y = "Axis 2"
  )
```



Clustering

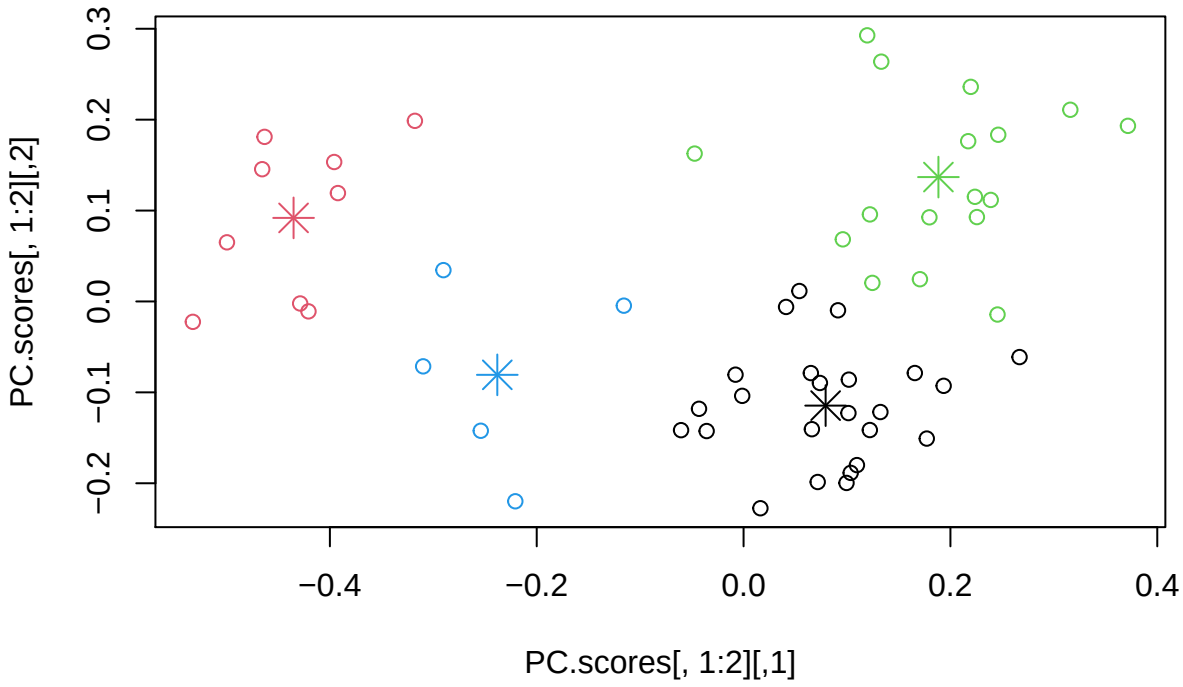
Although clustering by kmeans is somewhat qualitative, it does visualize the clustering we see in the PCoA above. When we cluster by 4, 3, and 2, kmeans=2 appears to explain the majority of the points.

```
PC.scores<-PCoA$points
```

Clustering by 4

Clustering under the assumption of 4 groups (k=4).

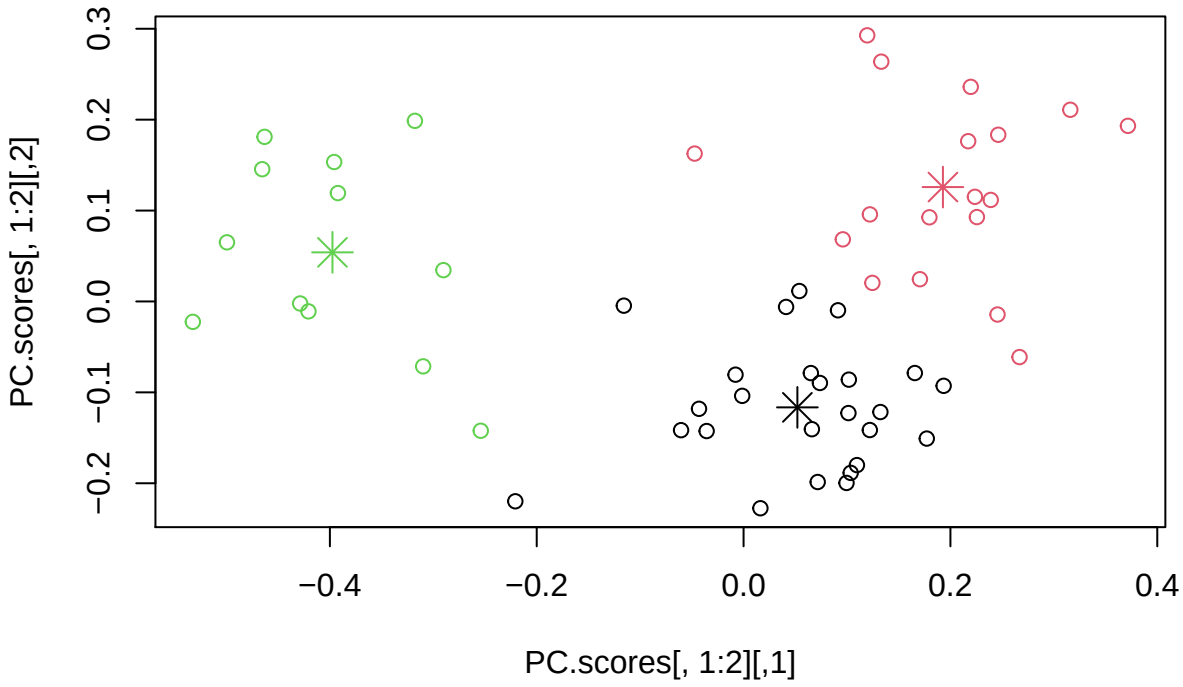
```
#K-means = 4
kclusters4<-kmeans(PC.scores,4)
plot(PC.scores[,1:2],col=kclusters4$cluster)
points(kclusters4$centers, col = 1:4, pch = 8, cex=2)
```



Clustering by 3

Clustering under the assumption of 3 groups (k=3).

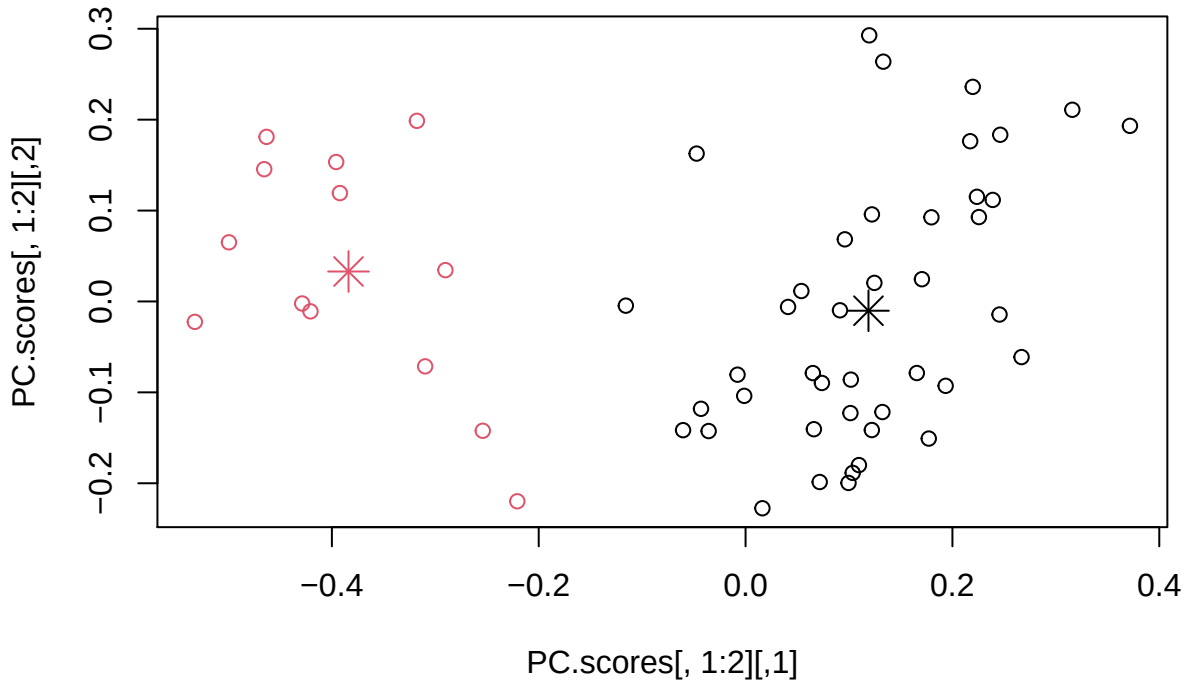
```
#K-means = 3
kclusters3<-kmeans(PC.scores,3)
plot(PC.scores[,1:2],col=kclusters3$cluster)
points(kclusters3$centers, col = 1:3, pch = 8, cex=2)
```



Clustering by 2

Clustering under the assumption of 2 groups (k=2).

```
#K-means = 2
kclusters2<-kmeans(PC.scores,2)
plot(PC.scores[,1:2],col=kclusters2$cluster)
points(kclusters2$centers, col = 1:2, pch = 8, cex=2)
```

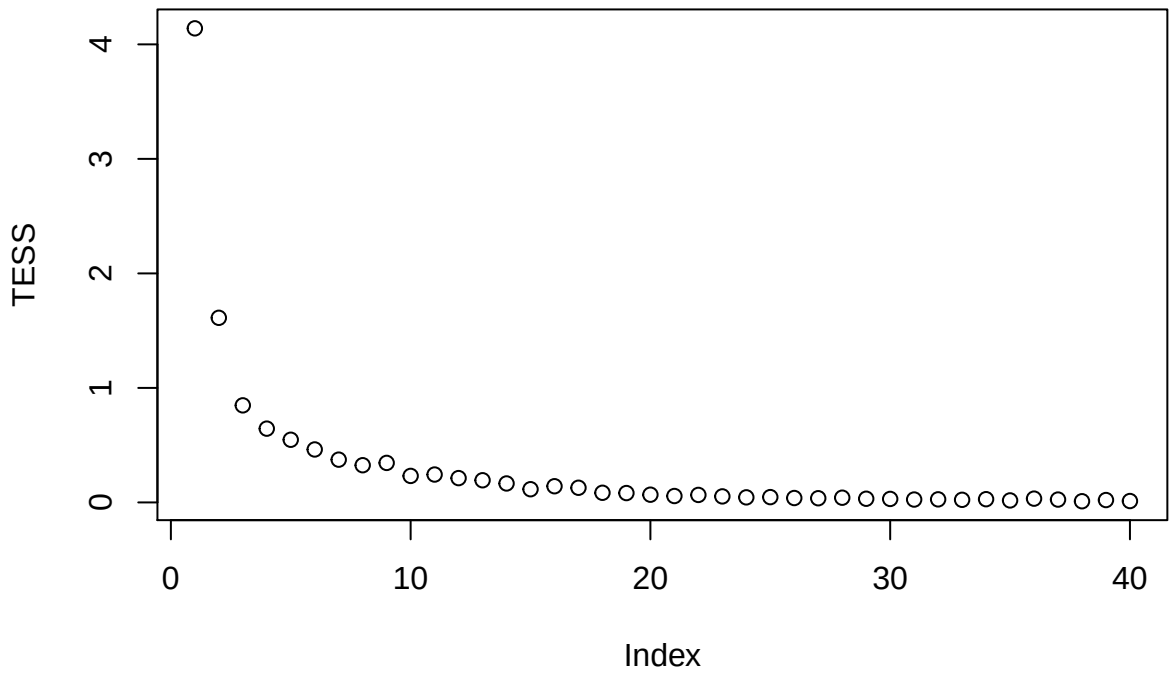



```
#NOTE: repeating k-means at a given level can lead to differing results
```

TESS: total error sums-of-squares

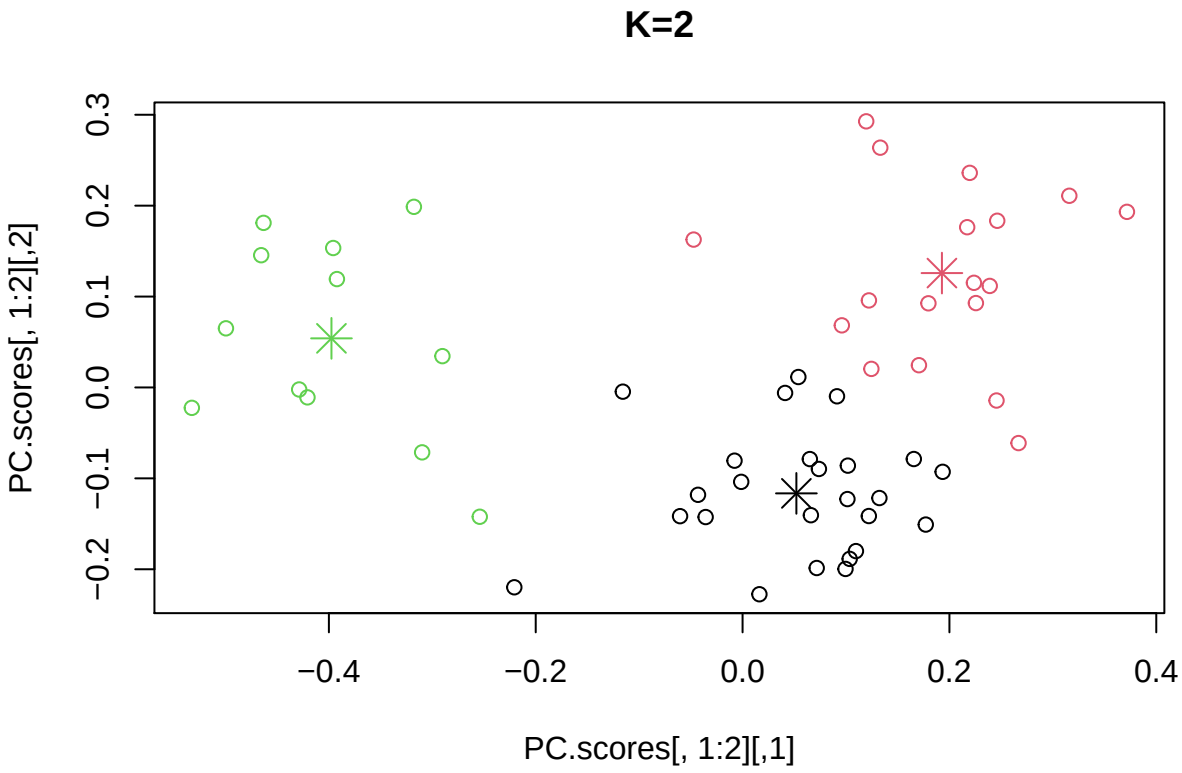
Compare the total error sums-of-squares to see which grouping results in a leveling off of the kmeans of PC scores.

```
#compare TESS
TESS<-array(NA,40)
for (i in 1:40){
  TESS[i]<-kmeans(PC.scores,i)$tot.withinss
}
plot( TESS)  #seems to bottom out at 3 groups
```



Based on the TESS results, it appears that the mean PC.scores level off at about a k grouping of 2 or 3. We would argue that there is 2 groups(k=2).

```
plot(PC.scores[,1:2],col=kclusters3$cluster, main="K=2")
points(kclusters3$centers, col = 1:3, pch = 8, cex=2)
```



FACTORIAL MANOVA

We conducted a factorial MANOVA to determine whether there were significant differences between the distances found for epidermal micromorphology by macromorphological type (group), habitat, and country. We found that there were significant differences in the distances for specimens based on macromorphological group and based on habitat. There was no significant difference between distances for specimens based on country, and no significant difference for the interaction between macromorphological group and country or habitat and country.

```
#Factorial MANOVA via RRPP
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 = 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(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)
```

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.
```

RRPP MODEL COMPARISON

For our model comparison, we used model.comparison() from the RRPP package using the log likelihood method. From this, we are able to determine that the Y.group2 model is the best fit based on it has the highest log-likelihood score and the lowest AIC score.

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

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

```
pandoc.table(modelComp1.summ,
             style = "grid",
               plain.ascii = TRUE)
```

```
##
##
```

	logLik	residual.pc.no
mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Group_2:mydat$Habitat + mydat$Group_2:mydat$Country + mydat$Habitat:mydat$Country + mydat$Group_2:mydat$Habitat:mydat$Country	2357	24
mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Habitat:mydat$Country	2399	25
mydat$Habitat + mydat$Group_2 + mydat$Country + mydat$Group_2:mydat$Country	2497	25
mydat$Group_2 + mydat$Habitat + mydat$Group_2:mydat$Habitat	2586	29
mydat$Habitat + mydat$Country	2256	25
mydat$Habitat	2345	28
mydat$Group_2 + mydat$Country	2294	25
mydat$Group_2 + mydat$Habitat	2586	29
mydat$Group_2	2690	30
mydat$Country	2041	24

Table: Table continues below

	penalty	AIC
mydat$Group_2 + mydat$Habitat + mydat$Country + mydat$Group_2:mydat$Habitat + mydat$Group_2:mydat$Country + mydat$Habitat:mydat$Country +	2370	-2345

```
## |  mydat$Group_2:mydat$Habitat:mydat$Country  |         |       |
## +---------------------------------------------+---------+-------+
## |               mydat$Group_2 +               |  2190   | -2608 |
## |        mydat$Habitat + mydat$Country        |         |       |
## |                       +                     |         |       |
## |         mydat$Habitat:mydat$Country         |         |       |
## +---------------------------------------------+---------+-------+
## |               mydat$Habitat +               |  2310   | -2684 |
## |        mydat$Group_2 + mydat$Country        |         |       |
## |                       +                     |         |       |
## |         mydat$Group_2:mydat$Country         |         |       |
## +---------------------------------------------+---------+-------+
## |               mydat$Group_2 +               |  1230   | -3942 |
## |               mydat$Habitat +               |         |       |
## |         mydat$Group_2:mydat$Habitat         |         |       |
## +---------------------------------------------+---------+-------+
## |               mydat$Habitat +               |  1950   | -2563 |
## |                mydat$Country                |         |       |
## +---------------------------------------------+---------+-------+
## |                mydat$Habitat                |  1110   | -3581 |
## +---------------------------------------------+---------+-------+
## |               mydat$Group_2 +               |  2010   | -2578 |
## |                mydat$Country                |         |       |
## +---------------------------------------------+---------+-------+
## |               mydat$Group_2 +               |  1230   | -3942 |
## |                mydat$Habitat                |         |       |
## +---------------------------------------------+---------+-------+
## |                mydat$Group_2                |  1170   | -4210 |
## +---------------------------------------------+---------+-------+
## |                mydat$Country                |  1830   | -2251 |
## +---------------------------------------------+---------+-------+
```

Based on the results of our analyses, it appears that there are significant differences in the epidermal micromorphology of Guadua based on macromorphological group (the Group_2 column in our data). Although there were also significant differences in epidermal micromorphology of Guadua based on habitat, classifying the species by macromorphological group appeared to better model the variation seen in our data.

The two clusters that come out of this break the groups into Savannah_type versus everything else. We were not able to do a Posthoc pairwise comparison of these types because of the nature of the data. For clarity we add a column purposefully breaking the specimens into Savannah type or other types and fit based off of the apparent clustering. After a factorial MANOVA, it appears that there is a significant difference between Savannah type and the other groups.

Group by Savannah or other:

```
# combine river group, tree killer and forest and compare against savannah
types<-c(as.character(unique(imputed$Group_2)))
Group_3<-rep(NA, 55)
savannah_index<-which(imputed$Group_2 == types[1])
not_sav_index<-which((imputed$Group_2 == types[1]) == FALSE)
Group_3[which(imputed$Group_2 == types[1])] <- c(as.character(imputed$Group_2[savannah_index]))
Group_3[not_sav_index]<-c("other")
new<-cbind(imputed[1:6], Group_3)
new<-cbind(new, imputed[7:ncol(imputed)])
new<-data.frame(lapply(new,as.factor))
#Factorial MANOVA via RRPP
```

```r
mydat2<-rrpp.data.frame("Y"=Y,
                        "Group_3"= as.factor(new$Group_3))

model2.rrpp <- lm.rrpp(Y.dist.matrix ~ mydat2$Group_3,
                       print.progress = FALSE)
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)
## mydat2$Group_3  1  2.0815 2.0815 0.19929 13.191 5.5948  0.001 **
## Residuals      53  8.3633 0.1578 0.80071
## Total          54 10.4448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call: lm.rrpp(f1 = Y.dist.matrix ~ mydat2$Group_3, print.progress = FALSE)
```