Introduction to basic multivariate analysis of community data using R

Introduction

In ecology in general and microbiome analysis in particular we use multivariate datasets. For instance, the species observed in several samples of a forest soil or environmental variables measured along a river. With this kind of data we could make the following questions, among others:

- What are the relationships among observed species?
- Are there differences between experimental groups, in terms of microbiome composition?
- etc.

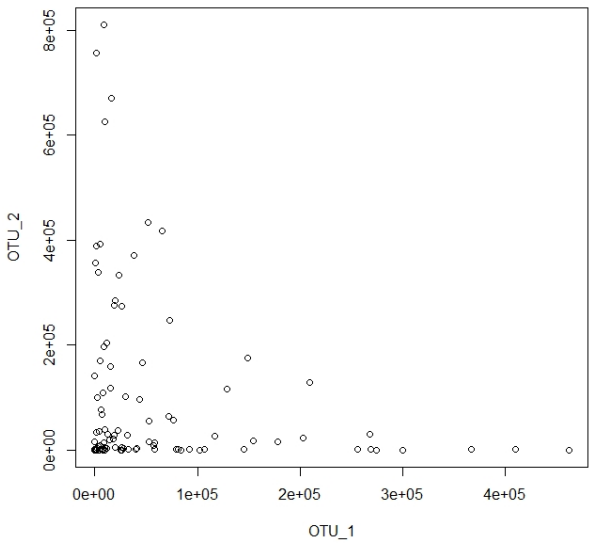
The most common statistical methods to respond the above questions start from computing some kind of association measure between species (in columns) or observations (in rows). In this introduction we will focus on relationships between observations (rows).

First of all, we need to download the datasets (here and here) we will analyse in this unit and place them in a folder. Then, we need to specify the working directory in R (i.e. the folder where we stored the datasets).

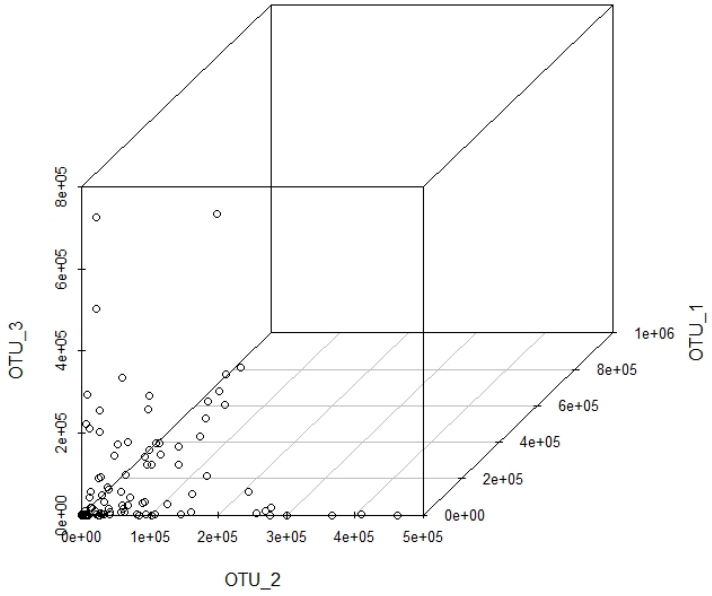
```
setwd("D:/APUNTEAK/Charles_University/Microbiome_course_2025/")
## Change this with your own working directory
```

The Multidimensional Space

We can interpret a multivariate dataset as a group of sites (rows of the dataset) located in the space, where each of the variables (columns of the dataset) represents a dimension. Therefore, a dataset contains as many dimensions as variables and, as many sites as observations located in those dimensions.



For example, in the above plot each site (point) takes a value for OTU_1 and other value for OTU_2, thus each point is located in two dimensions. What happens if we add a third OTU?



Now, each point takes a value for each of the three OTUs, hence the points are located in three dimensions. If we kept adding OTUs we would move to four, six, seven... dimensions. However, our brain is unable to interpret graphical representations of more than three dimensions.

Association Measures

Most of the analyses to study relationships between observations start from creating a distance matrix. Selecting among different normalization and transformation options as well as different association indices (i.e. distance or dissimilarity indices) is a whole world in itself. Although going into details is beyond the scope of this introductory course, it is good to mention some concepts.

When working species communities, an important aspect to consider when choosing between dissimilarity indices is, how does the index treat the double absences? If a species is present in two sites the interpretation is straightforward: both sites meet the basic requirements for the species to survive. However, the absence of an species can have several explanations: the niche of the species is occupied by another species; the species did not get to the site (e.g. due to dispersal limitation), even if the ecological conditions are suitable; the site does not meet the ecological requirements of the species; the species was there, but we could not detect it... The consequence of this is that the absence of an species from two sites may not give us information about the similarity between the sites, since both absences might have different reasons.

The most typical distance measure is the Euclidean distance. However, the Euclidean distance is not appropriate to work with species communities, since, when calculating the distance between two sites, it gives the same weight to double absences and double presences. Bray-Curtis dissimilarity solves this problem by ignoring double absences and has become the gold standard when making ordinations of species communities. However, rejecting the Euclidean distance means rejecting interesting methods like principal component analysis (PCA) or redundancy analysis (RDA), which are necessarily based on Euclidean distance. To overcome this problem, some data normalization and transformation procedures were proposed, after which,

the Euclidean distance can be safely used with species communities. The Hellinger transformation is one of these valid transformations. Combining Hellinger transformation with Euclidean distance generates the Hellinger distance matrix, which is a valid distance to use with species communities. You can refer to Legendre and Legendre 2012 to learn more about selection of association measures.

Although above approach have been widely applied by microbial ecologists, in recent years several publications have claimed that sequencing data is compositional, and should be treated as such. Microbiome data is compositional because sequencing machines provide relative counts of DNA fragments from different taxa, not absolute quantities. Since the total number of reads is limited and arbitrarily fixed by the sequencing depth, the resulting data reflects only proportions, making it inherently compositional. To combine compositional data with multivariate methods there are some specific transformations, known as ratio transformations. In this introduction we will use the centered log-ratio (CLR) transformation. A caveat of ratio transformations is that they can not deal with zeros in the data. Amplicon data is severely zero inflated, which represents a challenge when using compositional analyses. Here we will take a zero replacement approach to deal with it. However, there are other options too, each one with its own advantages and caveats.

We will load the packages vegan and zCompositions (install first, if it is not installed yet).

Lets load the OTU table. The table contains the OTUs identified in 142 samples of forest soils. The samples were collected in three countries: Panama, Costa Rica and Ecuador. The number of OTUs in the table is huge (17898). Usually, the first step in the analysis workflow is to remove global singletones, i.e. OTUs appearing in a single sample with just one count. These are likely to be sequencing artifacts rather than true species.

```
## Load OTU table (the dataset must be in the working directory)
otus_raw <- read.csv("Dunthorn_2017.csv",row.names = 1)
# remove global singletones from dataset
otus_raw <- otus_raw[,!colSums(otus_raw)<2]
```

Next, lets load the table containing the country of origin of the samples.

```
origin<-read.csv("origin.csv", stringsAsFactors = TRUE)
head(origin) # Display first rows of the dataset
```

```
##      Sample Origin
## 1      B010 Panama
## 2 B011_B012 Panama
## 3      B020 Panama
## 4 B029_B030 Panama
## 5      B030 Panama
## 6 B033_B034 Panama
```

```
# Sample = Sample names. The same as the row names of otus_raw.
# Origin = The country of origin of the samples.
```

```
library(vegan)
library(zCompositions)
```

Hellinger transformation first normalizes the table to relative abundances and, then, applies the square-root transformation.

```
# The row sums of otu_raw represent the sequencing depth: the sum of all the sequences
# observed in one site. Dividing each cell by its row sum, we get the relative abundance
# matrix.
otus_raw_rel<-otus_raw/rowSums(otus_raw)
# Applying the square root transformation (^0.5 = square root) to relative abundances,
# we get the Hellinger transformation.
otus_hel<-otus_raw_rel^0.5
# Rename the matrix rows to make the origin of the observations easier to identify
```

```r
rownames(otus_hel)<-paste0(origin$Origin,1:length(origin$Origin))
```

Next we will generate the CLR transformed table, which we will use to illustrate the analyses introduced in this document.

```r
# We apply the zero imputation and CLR transformations to raw counts.
# First, we input the zeros.
# Ideally, it would be better to use method = "GBM", but this method is sensitive to very sparse microb
otus_raw_zeroRepl <- suppressMessages(cmultRepl(otus_raw, method = "CZM",z.delete = FALSE,label = 0))
```

```
## No. adjusted imputations:  712948
```

```r
# Then, we define the CLR transformation
clr_transform <- function(x) {log(x) - mean(log(x), na.rm = TRUE)}
# Last, we apply the CLR transformation
otus_clr <- data.frame(t(apply(otus_raw_zeroRepl, 1, clr_transform)))
rownames(otus_clr)<-paste0(origin$Origin,1:length(origin$Origin))
```

It is also typical to apply some filter to reduce the number of OTUs. Many of these OTUs are probably too rare to provide any useful information. For instance, we can delete all the OTUs that are present in less than 5% of the samples. Filters based on abundance thresholds or combinations are also possible. Nevertheless, rare species affect distance/dissimilarity based multivariate methods very little, thus not removing any OTUs is also possible. In this case, we will apply a filter, to speed up the most computationally demanding steps.

```r
otus_clr_red<-otus_clr[,colSums(otus_raw>0)>(0.05*142)] # Filter OTUs < 0.05 prevalence
dim(otus_clr_red) # We retain 1053 OTUs
```

```
## [1]  142 1053
```

Now we will use the CLR-transformed matrix to illustrate a distance matrix. A Aitchison distance matrix in this case.

```r
as.matrix(vegdist(otus_clr_red, method="euclidean"))[1:6,1:6]
```

```
##          Panama1  Panama2  Panama3  Panama4  Panama5  Panama6
## Panama1  0.00000 33.79332 43.15132 43.29336 38.97730 44.76433
## Panama2 33.79332  0.00000 41.09400 39.66937 38.15631 40.91873
## Panama3 43.15132 41.09400  0.00000 49.14139 43.87320 48.58585
## Panama4 43.29336 39.66937 49.14139  0.00000 30.18608 45.14845
## Panama5 38.97730 38.15631 43.87320 30.18608  0.00000 45.05027
## Panama6 44.76433 40.91873 48.58585 45.14845 45.05027  0.00000
```

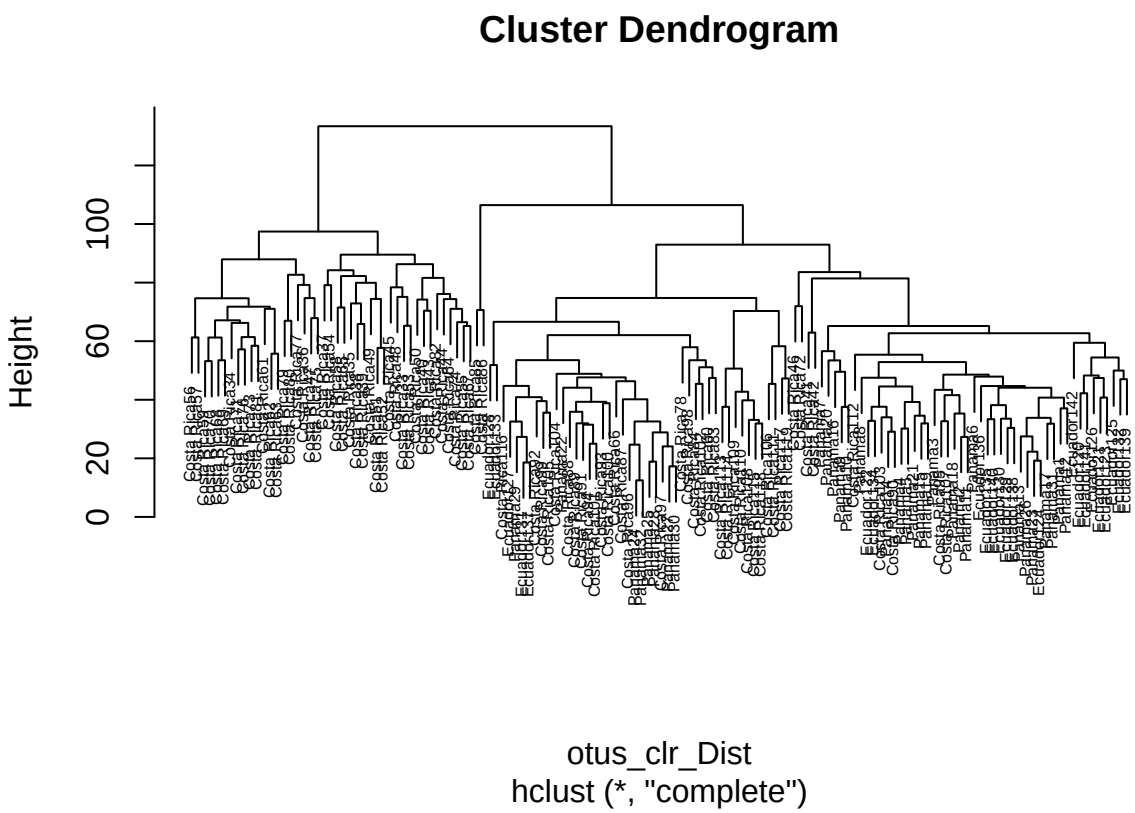
Note that we displayed the first 6 combinations of sites but the dimensions of the full matrix are 142 x 142. The bigger the distance value between two observations the more different those observations are (in terms of OTU composition). By contrast, if the distance between two observations is small, the observations are similar. Check the distance matrix above, the diagonal is composed by zeros (the distance of an observation with itself) and the matrix is symmetric above and below the diagonal (the distance from site 1 to 2 is the same as 2 to 1). The bottom half of the matrix is usually omitted.

Cluster analysis

Cluster analysis searches for discontinuous subsets in the data, which sometimes are discrete (as in taxonomy) but usually are continuous in ecology. The analysis is mostly exploratory and it can be useful to find general patterns in multivariate datasets. There are many ways to perform a cluster analysis, but here we will focus on hierarchical clustering, which, according to its name, organizes the data in a hierarchy. There are several ways of doing a hierarchical clustering too, but we will not delve into details.

As indicated above, a key step in any multivariate analysis is to choose an appropriate distance measure. Here we will use the Hellinger distance matrix generated above.

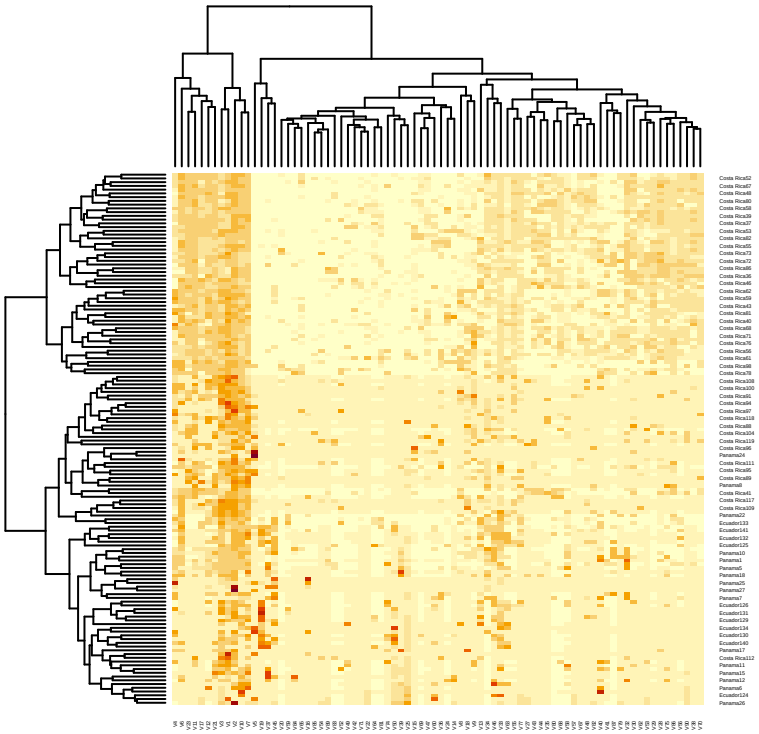
```
otus_clr_Dist <- vegdist(otus_clr_red, method="euclidean")
otus_clr_Clust <- hclust(otus_clr_Dist,method = "complete")
plot(otus_clr_Clust,cex=0.5)
```



Clusters are usually visualized in dendrograms. Above we see how the samples organize according to their OTU composition, the closest samples are the most similar. The vertical axis indicates the distance (as measured by the index used to generate the distance matrix) between samples and groups of samples. In our case, this raw dendrogram is not very useful, since the separations are not so obvious.

An interesting exploratory analysis, also based on cluster analysis, is the heatmap. A heatmap, by default, applies a hierarchical clustering of rows and columns, thus shows general patterns in both observations and variables (OTUs in our case).

```
heatmap(as.matrix(otus_clr_red[,1:80]),cexRow = 0.25,cexCol = 0.25)
```



These basic exploratory approaches not showing clear patterns does not mean that there are none. If the patterns are not very strong, we will need more sophisticated approaches to find them.

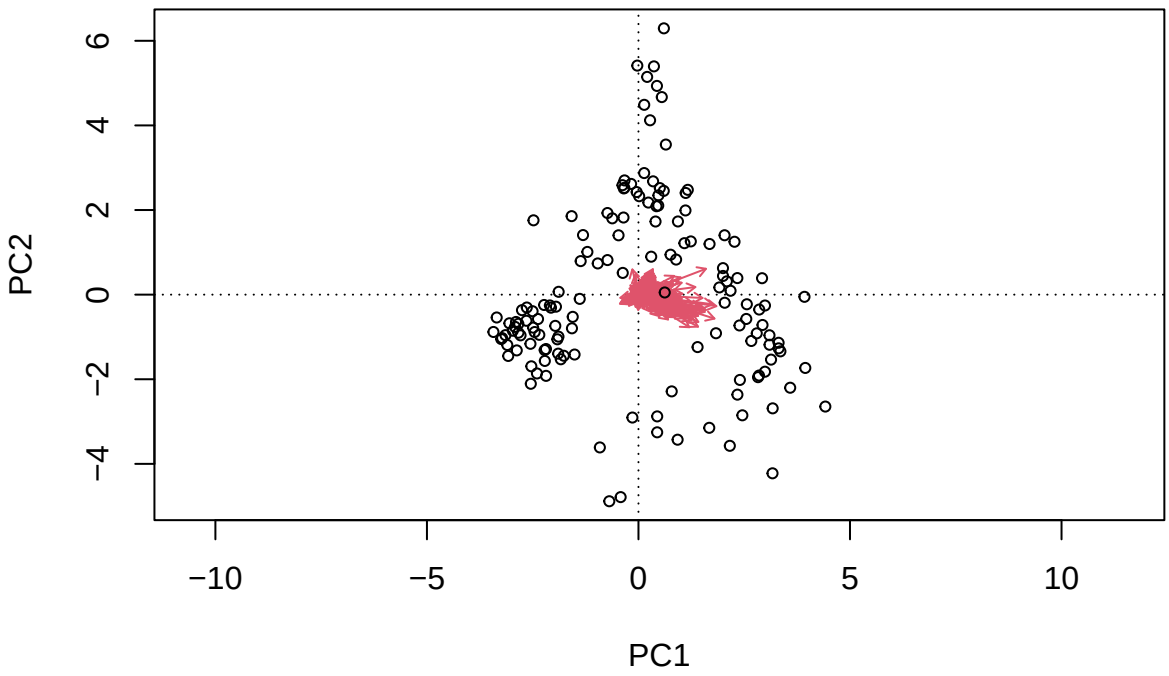
Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique. As mentioned above, a multivariate dataset is a set of n objects (as many as sites) and p dimensions (as many as variables). What PCA does is to represent this set of p dimensions in two dimensions (or as few as possible), by losing as little of the original variance as possible.

Imagine the above three dimensional figure (section **The Multidimensional Space**), and imagine a bubble containing all the points in the graph. The bubble would not be completely spherical but elongated in some direction. The principal component 1 (PC1) crosses the bubble along the longest axis of the bubble, which is the axis of maximum variance. Then, the bubble is crossed again with a second line, which represents the axis of second largest variance (PC2), and, which is always orthogonal to the first one. These new two lines are artificial variables that are linear combinations of the variables in the original dataset. By ploting these two variables (PC1 and PC2) we can draw the original 3D figure in 2D, while dismissing as little as the original variance as possible.

We will implement a PCA with the whole CLR-transformed OTU table.

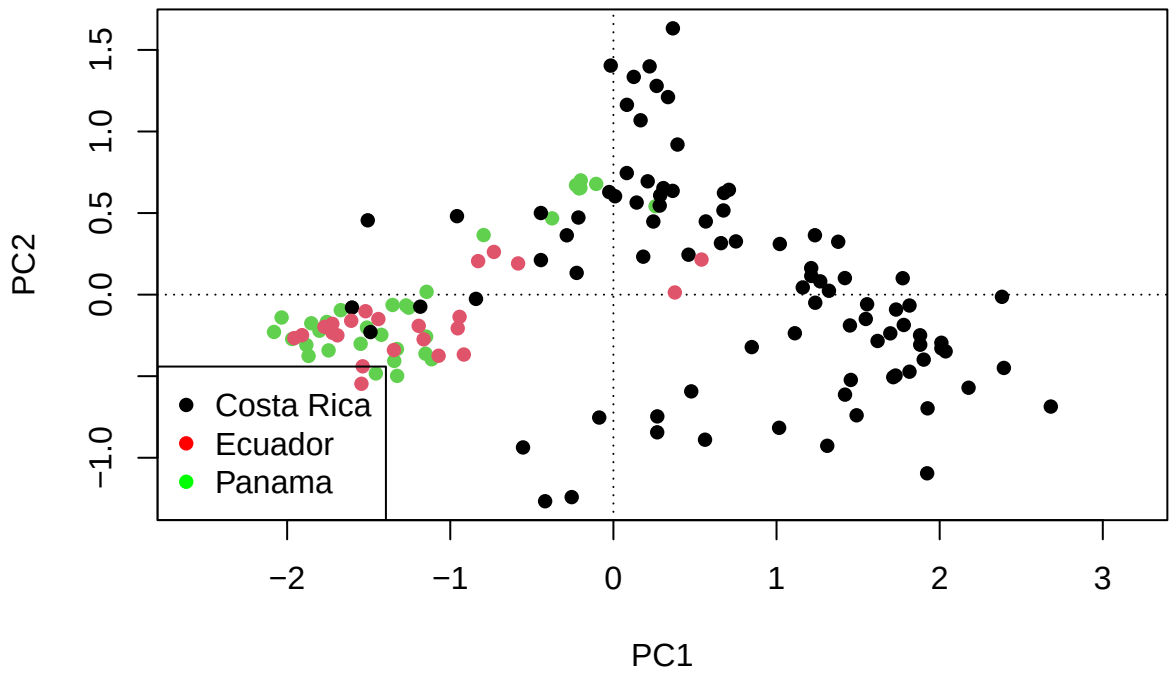
```
otu_clr_PCA <- rda(otus_clr_red, scale=FALSE)
biplot(otu_clr_PCA, scaling=2)
```



the *scale = FALSE* means that the PCA is performed on the covariance matrix. When using variables that are measured in different units, like chemical variables (pH, P...), we should use *scale = TRUE* and make the PCA on the correlation matrix. In our case, the units between variables are comparable (sequence counts), thus we will keep *scale = FALSE*. In the biplot, we specify *scaling = 2*, this means that the angles between the arrows (OTUs) are proportional to the correlations between them, but the distances between the points (the sites) are not proportional to the used distance (Aitchison distance in this case). If we set *scaling = 1* the opposite would happen, the angles would not reflect the correlations, whereas the distances between points would reflect actual distances.

Now, we can use PCA as an ordination technique by plotting the site scores and coloring them by countries. We will use *scaling = 1* to make the distances between points proportional to the real distances.

```
otus_clr_PCA <- rda(otus_clr_red, scale=FALSE)
biplot(otus_clr_PCA, scaling=1,display = "sites",type="n")
points(scores(otus_clr_PCA,display = "sites",scaling = 1),col=origin$Origin,pch=16)
legend("bottomleft", legend=c("Costa Rica","Ecuador","Panama"), pch=16,
       col=(c("black","red","green")))
```



As explained above, the first two components represent just a part of the whole picture. Next commands will give us the information behind the PCA.

```
head(summary(otus_clr_PCA))
```

```
##
## Call:
## rda(X = otus_clr_red, scale = FALSE)
##
## Partitioning of variance:
##               Inertia Proportion
## Total            2823          1
## Unconstrained    2823          1
##
## Eigenvalues, and their contribution to the variance
##
## Importance of components:
##                              PC1       PC2      PC3      PC4      PC5      PC6
## Eigenvalue            1040.3138 189.89189 83.71736 68.45898 54.28066 52.83459
## Proportion Explained     0.3684   0.06725  0.02965  0.02425  0.01922  0.01871
## Cumulative Proportion    0.3684   0.43570  0.46535  0.48960  0.50882  0.52754
##                             PC7      PC8      PC9     PC10     PC11     PC12
## Eigenvalue            47.11090 45.90039 38.84065 37.97273 36.04153 32.71955
## Proportion Explained   0.01669  0.01626  0.01376  0.01345  0.01276  0.01159
## Cumulative Proportion  0.54422  0.56048  0.57424  0.58768  0.60045  0.61204
##                            PC13     PC14     PC15      PC16     PC17      PC18
## Eigenvalue            30.65258 29.41936 28.44988 26.876641 26.34325 25.626906
```

```
## Proportion Explained   0.01086  0.01042  0.01008  0.009519  0.00933  0.009076
## Cumulative Proportion  0.62289  0.63331  0.64339  0.652908  0.66224  0.671314
##                            PC19     PC20      PC21      PC22      PC23
## Eigenvalue             24.856190 23.54836 22.930110 21.835006 21.287681
## Proportion Explained    0.008803  0.00834  0.008121  0.007733  0.007539
## Cumulative Proportion   0.680118  0.68846  0.696579  0.704312  0.711852
##                            PC24      PC25      PC26      PC27      PC28
## Eigenvalue             20.574096 20.280431 19.714081 19.382277 18.764681
## Proportion Explained    0.007287  0.007183  0.006982  0.006865  0.006646
## Cumulative Proportion   0.719139  0.726321  0.733304  0.740168  0.746814
##                            PC29      PC30      PC31      PC32      PC33
## Eigenvalue             18.649704 18.449127 18.191940 17.745353 16.624481
## Proportion Explained    0.006605  0.006534  0.006443  0.006285  0.005888
## Cumulative Proportion   0.753419  0.759953  0.766396  0.772681  0.778569
##                            PC34      PC35      PC36      PC37      PC38
## Eigenvalue             16.137450 15.850829 15.433970 15.382991 15.074916
## Proportion Explained    0.005715  0.005614  0.005466  0.005448  0.005339
## Cumulative Proportion   0.784285  0.789899  0.795365  0.800813  0.806152
##                            PC39      PC40      PC41      PC42      PC43
## Eigenvalue             14.612377 14.260791 13.868910 13.698898 13.314865
## Proportion Explained    0.005175  0.005051  0.004912  0.004852  0.004716
## Cumulative Proportion   0.811327  0.816378  0.821290  0.826142  0.830858
##                            PC44      PC45      PC46      PC47      PC48
## Eigenvalue             12.958391 12.760240 12.334598 12.187419 11.836638
## Proportion Explained    0.004589  0.004519  0.004369  0.004316  0.004192
## Cumulative Proportion   0.835447  0.839966  0.844335  0.848651  0.852844
##                            PC49      PC50      PC51      PC52      PC53
## Eigenvalue             11.730154 11.510731 11.371950 11.080489 10.802189
## Proportion Explained    0.004154  0.004077  0.004028  0.003924  0.003826
## Cumulative Proportion   0.856998  0.861075  0.865103  0.869027  0.872853
##                            PC54      PC55     PC56     PC57     PC58     PC59
## Eigenvalue             10.568046 10.351734 10.02339 9.870958 9.723244 9.608599
## Proportion Explained    0.003743  0.003666  0.00355 0.003496 0.003444 0.003403
## Cumulative Proportion   0.876596  0.880262  0.88381 0.887308 0.890752 0.894155
##                            PC60     PC61     PC62     PC63     PC64     PC65
## Eigenvalue             9.296325 9.110338 8.863245 8.268524 8.218683 8.116131
## Proportion Explained   0.003292 0.003227 0.003139 0.002928 0.002911 0.002874
## Cumulative Proportion  0.897447 0.900674 0.903813 0.906741 0.909652 0.912527
##                           PC66    PC67     PC68     PC69     PC70   PC71
## Eigenvalue             7.98924 7.93452 7.703074 7.471387 7.374445 7.0587
## Proportion Explained   0.00283 0.00281 0.002728 0.002646 0.002612 0.0025
## Cumulative Proportion 0.91536 0.91817 0.920895 0.923541 0.926153 0.9287
##                           PC72     PC73     PC74     PC75    PC76    PC77
## Eigenvalue             6.924387 6.737028 6.528853 6.343384 6.09873 5.98659
## Proportion Explained   0.002452 0.002386 0.002312 0.002247 0.00216 0.00212
## Cumulative Proportion 0.931105 0.933491 0.935803 0.938050 0.94021 0.94233
##                           PC78     PC79     PC80     PC81     PC82     PC83
## Eigenvalue             5.915449 5.784557 5.557210 5.371605 5.276477 5.175528
## Proportion Explained   0.002095 0.002049 0.001968 0.001902 0.001869 0.001833
## Cumulative Proportion 0.944425 0.946474 0.948442 0.950345 0.952214 0.954047
##                           PC84     PC85     PC86     PC87     PC88    PC89
## Eigenvalue             4.99888 4.839878 4.669467 4.584877 4.475454 4.43218
## Proportion Explained   0.00177 0.001714 0.001654 0.001624 0.001585 0.00157
## Cumulative Proportion 0.95582 0.957531 0.959185 0.960809 0.962394 0.96396
```

```
##                            PC90     PC91     PC92     PC93    PC94     PC95
## Eigenvalue              4.26346 4.131533 4.059333 3.970613 3.78460 3.666958
## Proportion Explained    0.00151 0.001463 0.001438 0.001406 0.00134 0.001299
## Cumulative Proportion 0.96547 0.966937 0.968375 0.969781 0.97112 0.972420
##                            PC96     PC97     PC98     PC99  PC100    PC101
## Eigenvalue              3.480286 3.461633 3.294953 3.255644 3.1068 3.022573
## Proportion Explained    0.001233 0.001226 0.001167 0.001153 0.0011 0.001071
## Cumulative Proportion 0.973653 0.974879 0.976046 0.977199 0.9783 0.979370
##                           PC102   PC103     PC104     PC105     PC106    PC107
## Eigenvalue              2.951902 2.88025 2.7768448 2.6470804 2.5775821 2.436734
## Proportion Explained    0.001045 0.00102 0.0009835 0.0009375 0.0009129 0.000863
## Cumulative Proportion 0.980415 0.98144 0.9824186 0.9833561 0.9842691 0.985132
##                            PC108     PC109     PC110    PC111     PC112
## Eigenvalue              2.3469856 2.2764939 2.2065391 2.111914 2.0101422
## Proportion Explained    0.0008312 0.0008063 0.0007815 0.000748 0.0007119
## Cumulative Proportion 0.9859633 0.9867696 0.9875511 0.988299 0.9890110
##                            PC113    PC114     PC115     PC116     PC117
## Eigenvalue              1.9466016 1.877647 1.8469581 1.7555647 1.7389381
## Proportion Explained    0.0006894 0.000665 0.0006541 0.0006218 0.0006159
## Cumulative Proportion 0.9897004 0.990365 0.9910196 0.9916413 0.9922572
##                            PC118     PC119     PC120     PC121     PC122
## Eigenvalue              1.6921616 1.6417483 1.5224765 1.4529872 1.3867912
## Proportion Explained    0.0005993 0.0005815 0.0005392 0.0005146 0.0004912
## Cumulative Proportion 0.9928565 0.9934380 0.9939772 0.9944918 0.9949830
##                            PC123     PC124     PC125     PC126     PC127
## Eigenvalue              1.3025181 1.2463555 1.1676101 1.1136411 1.0517874
## Proportion Explained    0.0004613 0.0004414 0.0004135 0.0003944 0.0003725
## Cumulative Proportion 0.9954443 0.9958857 0.9962992 0.9966937 0.9970662
##                            PC128     PC129     PC130     PC131    PC132
## Eigenvalue              1.0211445 0.9816981 0.9111264 0.8207663 0.816065
## Proportion Explained    0.0003617 0.0003477 0.0003227 0.0002907 0.000289
## Cumulative Proportion 0.9974278 0.9977755 0.9980982 0.9983889 0.998678
##                            PC133    PC134     PC135     PC136     PC137
## Eigenvalue              0.7121305 0.666302 0.5762188 0.4834034 0.3977452
## Proportion Explained    0.0002522 0.000236 0.0002041 0.0001712 0.0001409
## Cumulative Proportion 0.9989302 0.999166 0.9993702 0.9995414 0.9996823
##                            PC138     PC139     PC140     PC141
## Eigenvalue              0.3436405 0.2932286 1.805e-01 7.964e-02
## Proportion Explained    0.0001217 0.0001039 6.394e-05 2.821e-05
## Cumulative Proportion 0.9998040 0.9999079 1.000e+00 1.000e+00
##
## Scaling 2 for species and site scores
## * Species are scaled proportional to eigenvalues
## * Sites are unscaled: weighted dispersion equal on all dimensions
## * General scaling constant of scores:  25.11895
##
##
## Species scores
##
##          PC1     PC2     PC3      PC4     PC5      PC6
## V1    1.5957  0.6114 -0.4201  0.72190  0.2371 -0.30869
## V2    1.3449  0.1792  0.3481  0.43306 -0.1215  0.10568
## V3    1.0210  0.2809 -0.4938  0.41042 -0.2619 -0.12103
## V4    0.9975  0.4071 -0.6644  0.25724  0.1809  0.53405
```

```
## V5   -0.1451  0.5955  0.1806 -0.19357  0.2861  0.08427
## V6    1.5947 -0.3619 -0.3331  0.08237 -0.2347 -0.05409
## ....
##
##
## Site scores (weighted sums of species scores)
##
##            PC1     PC2      PC3     PC4    PC5      PC6
## Panama1 -2.553 -1.1628 -1.02031  0.5822 0.4512 -0.34181
## Panama2 -2.894 -0.6464 -1.13591  0.8365 2.9276 -0.05958
## Panama3 -2.399 -1.8633  0.00942  1.6703 3.3804 -0.37895
## Panama4 -2.230 -0.2450 -0.09977 -1.1521 0.2094 -0.16582
## Panama5 -2.343 -0.9501 -0.29705  0.3115 0.7570 -0.16925
## Panama6 -3.101 -1.1905 -0.60456 -0.3981 1.2014 -0.45129
## ....
```

The summary of the PCA gives a lot of information but we will focus just in the section **importance of components**. There, you can see that the proportion of variance explained by the PC1 and PC2 is ~44%. The other 56% is shared between the rest of PCs. Is not surprising that the largest fraction of the variance is not represented in the first two PCs, since we have reduced 1053 OTUs to two dimensions.

Non-Metric Multidimensional Scaling (NMDS)

The NMDS is a true ordination method. It is the choice when the purpose is to represent the relationships between objects as clearly as possible. As opposed to PCA, the aim of this method is not to represent as much variance as possible in two dimensions, but to represent relationships between objects as close to the reality as possible.

To understand the process in an intuitive way, we can imagine a hand (three dimensions) lighted with a torch. Then, we move the torch to try to project the hand in the wall (two dimensions), but we do it by trying to get the original shape as clearly as possible. NMDS makes something similar: it projects in two dimensions (or the specified dimensions) the points in the multidimensional space, keeping the distances between the points as close to the real distances as possible.

Another advantage of NMDS is that we can use any kind of distance/dissimilarity measure, as opposed to PCA, where we were bound to the Euclidean distance.

```
# Leaving default options the ordination is based on Bray-Curtis dissimilarity.
otu_clr_NMDS <- metaMDS(otus_clr_red,distance = "euclidean",maxit=1000, trymax=200,k=2,autotransform = F
```

```
## 'comm' has negative data: 'autotransform', 'noshare' and 'wascores' set to FALSE
```

Before interpreting the NMDS we need to have a look to the associated stress. The stress is the discrepancy between real distances between observations and represented distances in the NMDS. By default, the NMDS is performed in two axes (k=2). When the dataset is too complex to be represented in two axes, we can increase k to 3, 4... It is better to keep k as low as possible, if the stress is reasonably low. PRIMER manual gives the following rule of thumb for interpretation:

- STRESS smaller than 0.05. The configuration is excellent and allows for a detailed inspection.
- STRESS between 0.05 and 0.1. Good configuration and no need to increase k.
- STRESS between 0.1 and 0.2. Be careful with the interpretation.
- STRESS between 0.2 and 0.3. Problems start, especially in the upper range of this interval.
- STRESS larger than 0.3. Poor presentation and consider increasing k.

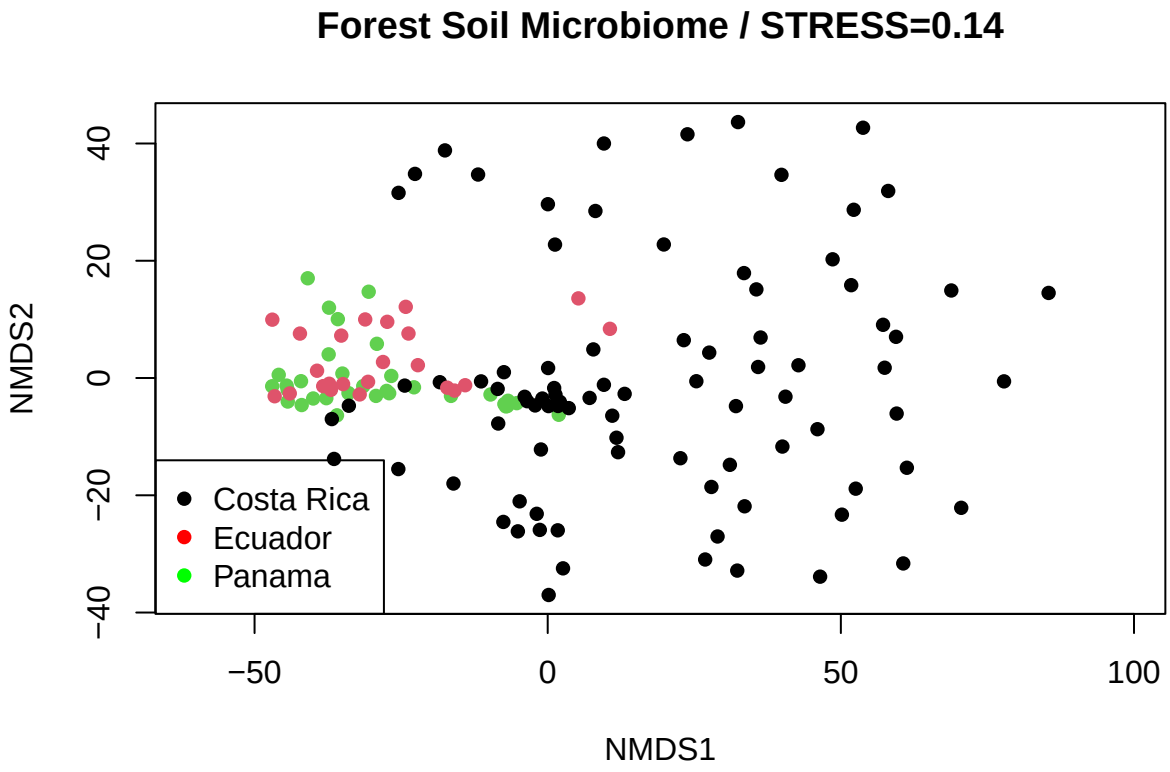
```
otu_clr_NMDS$stress
```

```
## [1] 0.1427393
```

In our example the stress is 0.14. It is common to have high stress values in microbiome analyses, since datasets are usually highly dimensional. In our case is acceptable.

Now we are ready to plot the results.

```
plot(otu_clr_NMDS, type="n", main="Forest Soil Microbiome / STRESS=0.14")
points(otu_clr_NMDS$points, col=origin$Origin, pch=16)
legend("bottomleft", legend=c("Costa Rica","Ecuador","Panama"), pch=16,
       col=(c("black","red","green")))
```



PERMANOVA

Lastly, NMDS ordiantions are usually accompanied by a PERMANOVA or some other formal statistical test. In this case, we will test the null hypothesis of no effect of country of origin on the microbiome composition of forest soils. PERMANOVA in R is implemented with the function adonis.

```
permutest(betadisper(dist(otus_clr_red),origin$Origin))
```

```
##
## Permutation test for homogeneity of multivariate dispersions
## Permutation: free
## Number of permutations: 999
##
## Response: Distances
##            Df Sum Sq Mean Sq      F N.Perm Pr(>F)
## Groups      2  11306  5653.3 45.107    999  0.001 ***
## Residuals 139  17421   125.3
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
adonis2(otus_clr_red~origin$Origin,method = "euclidean")
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = otus_clr_red ~ origin$Origin, method = "euclidean")
##                Df SumOfSqs      R2      F Pr(>F)
## origin$Origin   2    89301 0.22431 20.098  0.001 ***
## Residual      139   308812 0.77569
## Total         141   398113 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The effect of country is significant, but since betadisper is also significant, we can not conclude if the significant country effect is due to differences in communities or dispersions between countries.

Go in Depth

In this introduction we have seen a very shallow introduction to the basic multivariate analyses. Combining Legendre and Legendre (2012) and Borcard et al. (2018) you can have an in depth view of ordination based multivariate analysis of ecological data. Chapters 10-15 of Zuur et al. (2007) also explain the most common multivariate techniques. Details on compositional data analysis of microbiome data can be found in Gloor (2017) and other publications form the same author.

References

1. Borcard D., François G., Legendre P. 2018. *Numerical Ecology with R, 2nd Edition*. Springer. New York.
2. Dunthorn M., Kauserud H., Bass D., Mayor J., Mah H. 2017. Yeasts dominate soil fungal communities in three lowland Neotropical rainforests. *Environmental Microbiology Reports*. 9: 668-675.
3. Legendre P., Legendre L. 2012. *Numerical Ecology, 3rd Edition*. Elsevier. Montreál.
4. Zuur A.F., Ieno E.N., Smith G.M. 2007. *Analysisng Ecological Data*. Springer, New York.
5. Gloor G.B. , Macklaim J.M. , Pawlowsky-Glahn V. , Egozcue J.J. 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*. 8: 2224.