dispRity demo for ecologists

Thomas Guillerme
guillert@tcd.ie

November 26, 2015

This demo aims to give quick overview of the dispRity package (v.0.1.2) for ecological analysis. Please refer to GitHub page: github.com/TGuillerme/dispRity for other vignettes, namely the dispRity tutorial that explains the functions in more details.

To keep it short, this package allows to use all the dimensions of ordinated matrices (i.e. PCA, MDS, PCO) for statistical analysis rather than just a sub-set of dimensions. For example, one might want to know whether the some sort of water treatment alters invertebrate communities and composition in natural habitats.

Contents

1 Before starting **1**
1.1 Installing dispRity 1
1.2 Data 2

2 A classical two dimensional approach **2**

3 A multidimensional approach with dispRity **3**
3.1 Splitting the data 3
3.2 Calculating disparity 4
3.3 Bootstrapping the data 5
3.4 Summarising the data 6
3.5 Testing the hypothesis 9

1 Before starting

1.1 Installing dispRity

You can install this package easily if you are using the latest version of R and devtools.

```r
install.packages("devtools")
library(devtools)
install_github("TGuillerme/dispRity", ref = "release")
library(dispRity)
```

This is a quick demo for using the dispRity package (v.0.1.2) in ecological analysis. See the other dispRity demos for a general demo of the dispRity package.

1.2 Data

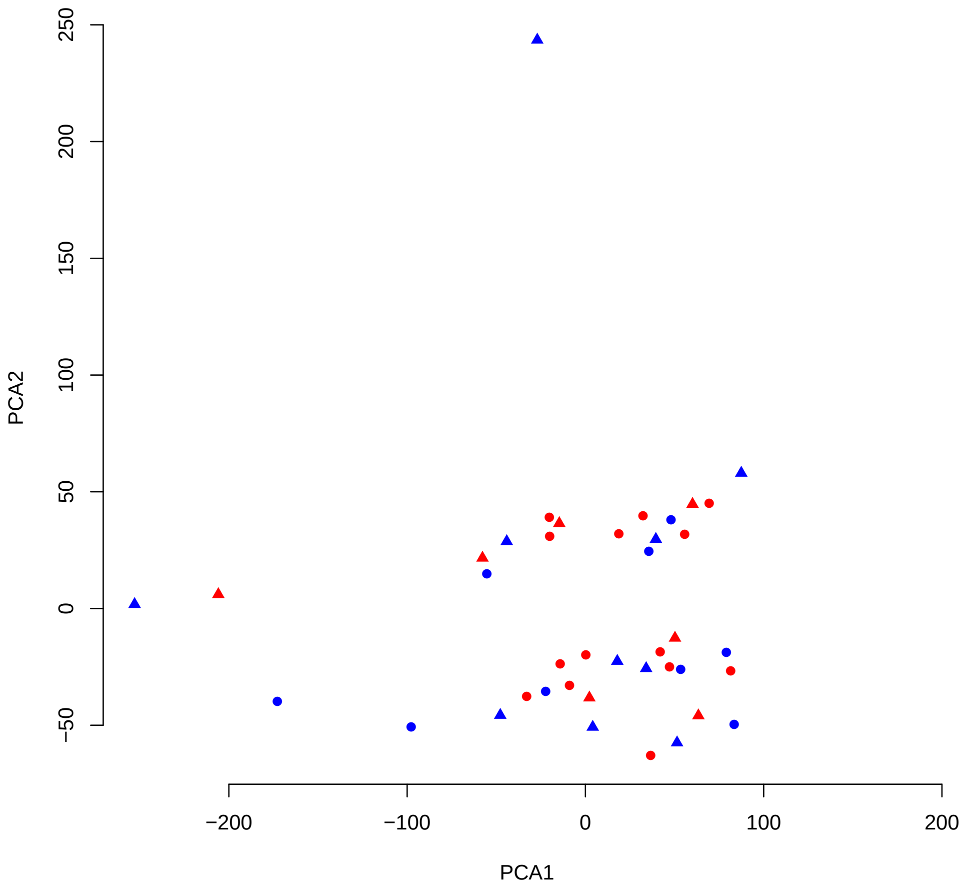
For this example with ecological data we are going to use data from McClean (unpubl.) that contains an ordination of a distance matrix between different experimental plots.

```
## Loading demo and the package data
library(dispRity)

## Loading required package:  paleotree

## let's use the matrix from McClean (unpubl.)
data(McClean_data)
## This dataset contains an ordinated matrix (in 20 dimensions) of the distance
## between experimental plots.
ord_matrix <- McClean_data$ordination
## As well as two list of different factors affecting each experimental plot:
## the treatment and the depth.
treatments <- McClean_data$treatment
depth <- McClean_data$depth
```

2 A classical two dimensional approach

A classical way to represent this ordinated data will be to use PCA plots.

```
## The x and y axis represent the two first dimensions of the PCA
x <- ord_matrix[, 1]
y <- ord_matrix[, 2]
## The colors will represent the treatments
cols <- sub("a", "red", treatments)
cols <- sub("b", "blue", cols)
## The symbols will represent the depth
pchs <- sub(1, 16, depth)
pchs <- as.numeric(sub(2, 17, pchs))
## Graphical option
par(bty = "n")
## A classic PCA plot
plot(x, y, col = cols, pch = pchs, xlab = "PCA1", ylab = "PCA2",
    xlim = range(x) + c(0, 100))
```



However, a common problem in this type of multivariate analysis is that often only a handful of dimensions are studied (usually the ones that bears the most variance of the ordinated matrix).

This shows the distribution of the experimental plots along the two first axis of variance of the ordinated distance matrix. However, the problem, is it ignores the 18 other axis of the ordination and the PCA axis 1 and 2 do not represent a biological reality *per se* but more some ordinations of correlations between the data and some factors. Therefore, one might want to approach this problem without getting stuck in only two dimensions and consider the whole dataset as a *n*-dimensional object. In practice, we might be interested in looking how some experimental treatment for example, will affect the position of our experimental plots in this *n*-dimensional object. For example: do the experimental plots shift in some specific space of the *n*-dimensional object when depth increases?

3 A multidimensional approach with `dispRity`

3.1 Splitting the data

As a first split, one might want to split the data (i.e. the *n*-dimensional object) into subsamples that we want to compare. First let's make a factor table:

```r
## Making the factor table
factors <- as.data.frame(matrix(data = c(treatments, depth), nrow = nrow(ord_matrix),
    ncol = 2, byrow = FALSE, dimnames = list(rownames(ord_matrix))))
names(factors)<-c("Treat", "Depth")
```

```
## And here is what it looks like
head(factors)

##     Treat Depth
## 1       a     1
## 1.1     a     2
## 2       b     1
## 2.1     b     2
## 3       a     1
## 3.1     a     1
```

Second, let's split the data according to these factors to create the subsamples of the ordinated space by using the `cust.series` function:

```
## Splitting the ordinated space into four subsamples
customised_series <- cust.series(ord_matrix, factors)
## Note that the output of dispRity functions are dispRity objects
class(customised_series)

## [1] "dispRity"

## These objects are automatically printed in a summary method (calling S3 print.dispRity)
## giving information about the object
customised_series

## 4 custom series for 40 elements
## Series:
## Treat.a, Treat.b, Depth.1, Depth.2.
```

For more details on the `dispRity` objects, see the other dispRity demos. Basically the idea is to avoid jamming the `R` console such as when using:

```
## Summarise the object
str(customised_series)
```

3.2 Calculating disparity

Now we're going to see the functionalities of the core function of this package: the `dispRity` function. This function is a modulable function that allow to simply (and quickly!) calculate disparity from a matrix. Disparity can be calculated in many ways, this function is a tool to measure disparity *as defined by the user* (and here's where the modulable part comes in). For more details on disparity, see the other dispRity demos.

One can usually decompose the disparity metrics into two elements:

1. the **class metric** that is a descriptor of the matrix. For example describing the ranges of each column in the matrix or the euclidean distances between each row and the centroid of the matrix.

2. the **summary metric** that is a summary of the class metric values. For example, the sum of the ranges or the median of the euclidean distances.

Basically the combination can be infinite between the class and summary metrics. For example, people might want to measure the median variances of the axis or the product of the distances from the centroid. However, it is probable that some metrics are better to reflect some biological aspects of the any-o-space than others...

In practice, the `dispRity` function intakes a pair of class and summary metrics as a definition of disparity. Several of these metrics are implemented in other packages (like `stats::median`, `base::sum`, etc.) and this package proposes several metrics listed in dispRity.metric (see ?dispRity.metric). But it is even possible to use your very own class and summary metrics! Or even any metric you decide makes sense (note that this is not yet implemented in v.0.1.2). This will be actually heavily encouraged and facilitate with the make.metric function in a future version.

To use these metrics pairs in the dispRity function, it's pretty easy:

```
## For example, let's calculate the median distance between each plot and the
## centroid of the ordinated space
disparity <- dispRity(customised_series, metric = c(median, centroids))
## Note that disparity is a dispRity object and printing it just gives details
## on the object, not the results. We need to use summary.dispRity (S3) to get
## the results.
summary(disparity)

##    series  n observed
## 1 Treat.a 21    82.38
## 2 Treat.b 19    94.68
## 3 Depth.1 23    83.98
## 4 Depth.2 17    89.72
```

Note that we calculated the median distance between plots and the centroid but the output displays mean values. This is because summary will, by default, summarise the data using the mean value. Here the mean represents the mean of the median distance for each series which is a bit useless (i.e. the mean of one value is that same value). We can display the median as well using:

```
summary(disparity, cent.tend=median)

##    series  n observed
## 1 Treat.a 21    82.38
## 2 Treat.b 19    94.68
## 3 Depth.1 23    83.98
## 4 Depth.2 17    89.72

## Or even the product! It won't affect the results
summary(disparity, cent.tend=prod)

##    series  n observed
## 1 Treat.a 21    82.38
## 2 Treat.b 19    94.68
## 3 Depth.1 23    83.98
## 4 Depth.2 17    89.72
```

This is because we did calculate the noise within our data. We can classically do so by bootstrapping the data!

3.3 Bootstrapping the data

The dispRity pacakge also provides easy way to bootstrap the data via the boot.matrix function. We can even rarefy the data to see the effect of the number of plots per series: For more details on the bootstrapping options, see the other dispRity demos.

```
bootstrapped_data <- boot.matrix(customised_series, bootstraps=100)
rarefied_data <- boot.matrix(customised_series, bootstraps=100, rarefaction=TRUE)
## Note that the output is a dispRity object giving some details on the series and the bootstraps
bootstrapped_data
```

```
## Bootstrapped ordinated matrix with 40 elements
## Series:
## Treat.a, Treat.b, Depth.1, Depth.2.
## Data was split using custom method.
## Data was bootstrapped 100 times, using the full bootstrap method.
```

We can now rerun a more robust disparity analysis:

```
disparity_BS <- dispRity(bootstrapped_data, metric = c(median, centroids))
disparity_rare <- dispRity(rarefied_data, metric = c(median, centroids))
## Note that calculation time is increased!
```

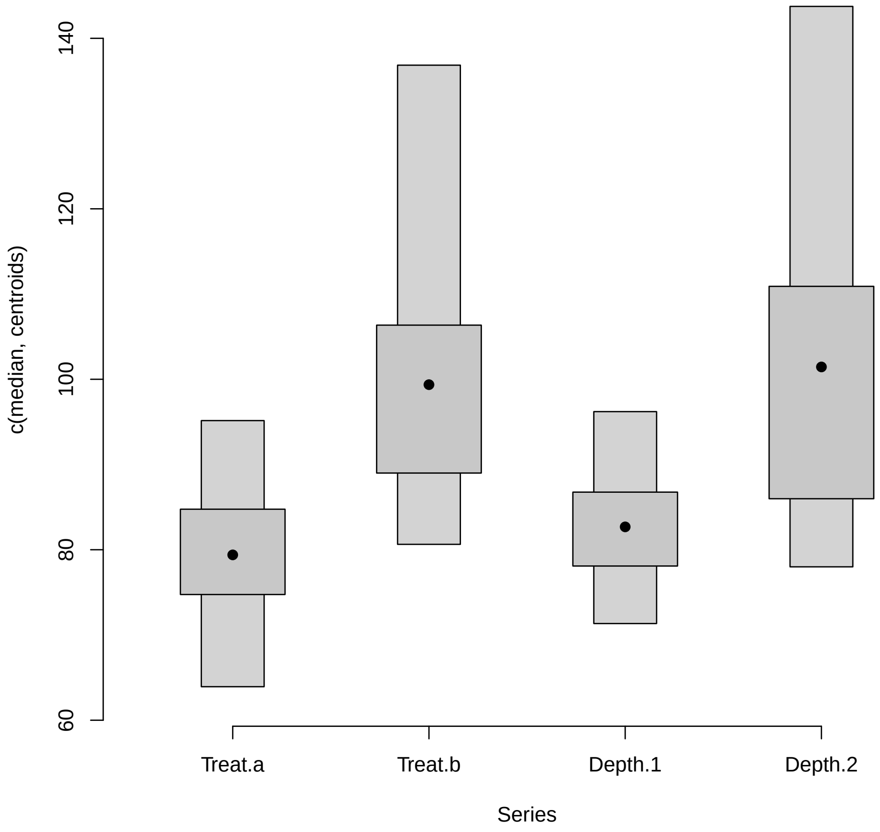
3.4 Summarising the data

We can now summarise the data using various options such as the confidence intervals levels and the central tendency.

```
## The default:
summary(disparity_BS)

##    series  n observed  mean  2.5%   25%   75% 97.5%
## 1 Treat.a 21    82.38  79.4 63.93 74.75  84.8  95.2
## 2 Treat.b 19    94.68  99.4 80.63 89.00 106.4 136.9
## 3 Depth.1 23    83.98  82.7 71.34 78.09  86.8  96.2
## 4 Depth.2 17    89.72 101.5 77.99 85.99 110.9 143.7

## The quantiles are calculated as 50 and 95 and the central tendency is the mean by default.
## But we can specify different options
summary(disparity_BS, quantile=90, cent.tend=median)

##    series  n observed median    5%   95%
## 1 Treat.a 21    82.38  79.04 65.21  91.7
## 2 Treat.b 19    94.68  96.17 82.21 129.5
## 3 Depth.1 23    83.98  82.30 72.47  94.3
## 4 Depth.2 17    89.72  97.85 79.01 140.9

## Finally we can see the results of the rarefaction analysis:
head(summary(disparity_rare))

##    series n observed mean  2.5%   25%  75% 97.5%
## 1 Treat.a 3       NA 69.3 28.73 53.56 84.4 119.5
## 2 Treat.a 4       NA 75.7 52.15 61.42 83.6 138.3
## 3 Treat.a 5       NA 74.5 45.69 61.52 84.1 137.5
## 4 Treat.a 6       NA 78.1 49.79 66.20 85.1 127.4
## 5 Treat.a 7       NA 75.8 53.80 67.57 80.1 119.9
## 6 Treat.a 8       NA 77.4 60.21 69.87 82.8 100.0

## This outputs a longer table with all the variations of plots down to 3 plots per series.
```

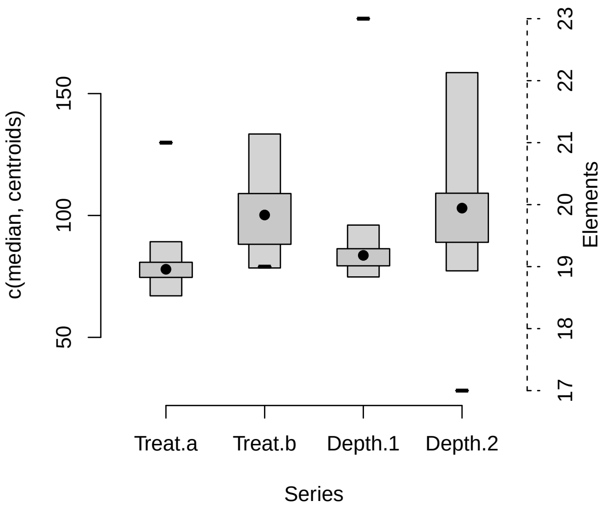
Finally we can also plot the results using the simple `plot.dispRity S3` method:

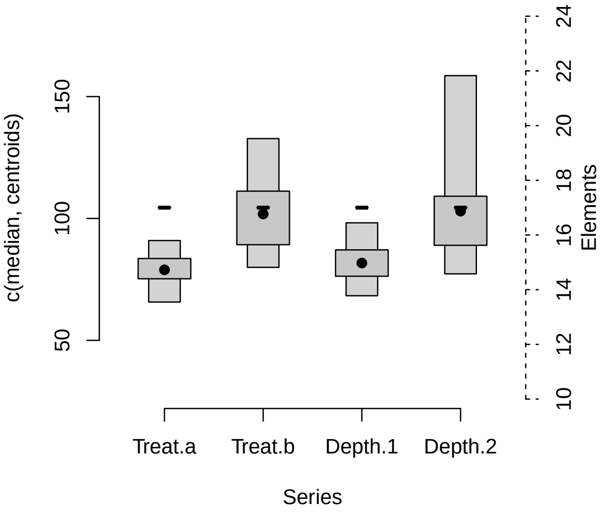
```
## Graphical option
par(bty = "n")
## Plotting the score for each groups
plot(disparity_BS)
```



Or have a look at the effect of the number of experimental plots:

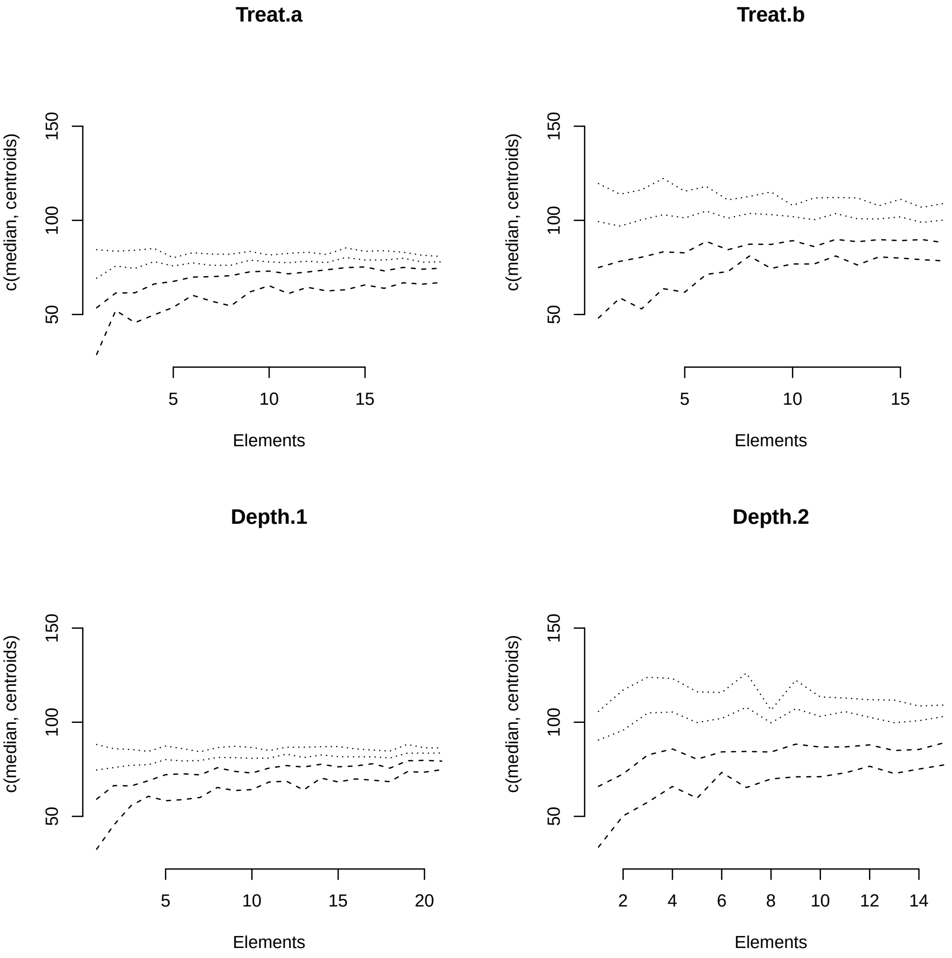
```
## Graphical options
quartz(width = 10, height = 5) ; par(mfrow = (c(1,2)), bty = "n")
## The same but looking at the number of plots
plot(disparity_rare, elements = TRUE)
## With the same number of plots per group
plot(disparity_rare, elements = TRUE, rarefaction = 17)
```





Or event have a look at the rarefaction curves:

```
## Graphical option
par(bty = "n")
## Plotting the rarefaction curves
plot(disparity_rare, rarefaction = "plot")
```



3.5 Testing the hypothesis

References