

Raw-data and distance based methods

in practice

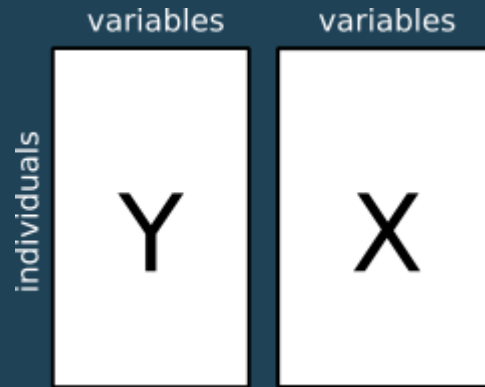
Stéphane Dray

2022-01-07

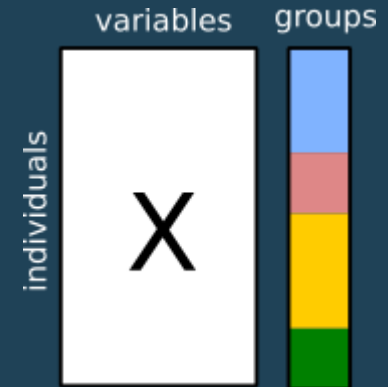
Raw data



**Principal Component
Analysis**

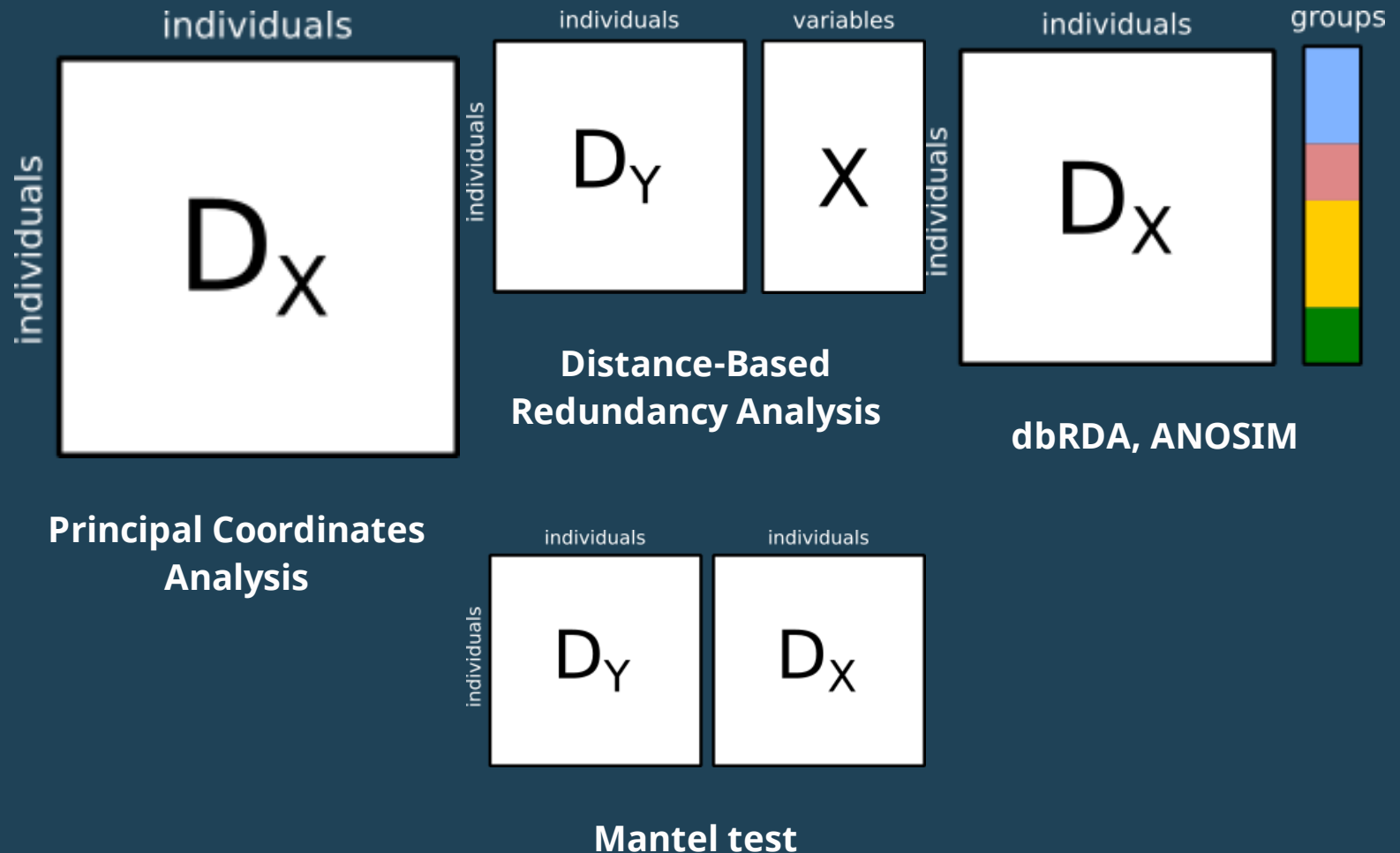


Co-Inertia Analysis



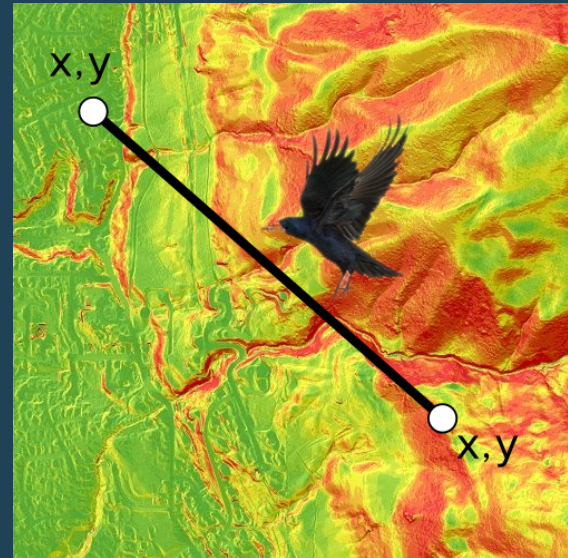
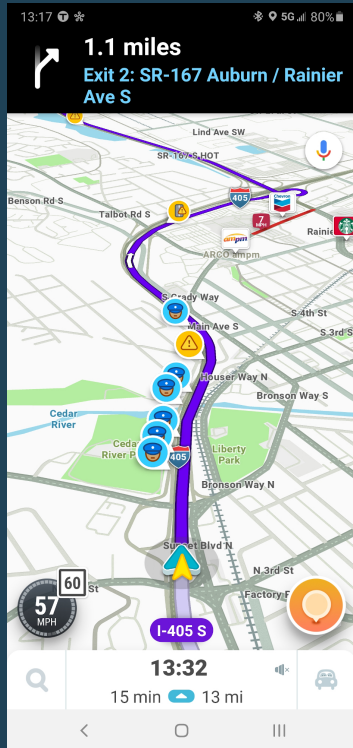
Between-Class Analysis

Distances



Distance vs Raw

- Distances can be directly measured or inherited from raw-data



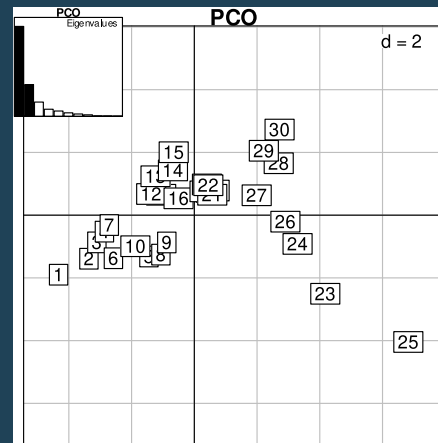
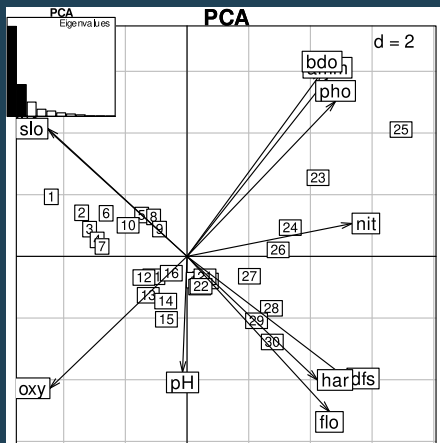
Distance vs Raw

- Raw-data methods produce information on both individuals and variables when distance-based methods focus only on individuals

```
library(ade4)
library(adegraphics)
data(doubs)
pca <- dudi.pca(doubs$env, scannf = FALSE)
pco <- dudi.pco(dist(scale(doubs$env)), scannf = FALSE)
```

```
scatter(pca, main = "PCA")
```

```
scatter(pco, main = "PCO")
```



Distance vs Raw

- Distance-based methods allow for more flexibility and can be more suitable in some contexts

```
x
```

```
##           species 1 species 2 species 3 species 4
## site 1           1           1           0           1
## site 2           0           0           0           1
## site 3           1           1           1           1
## site 4           0           0           1           1
```

```
## Euclidean
round(dist(x), 2)
```

```
##           site 1 site 2 site 3
## site 2       1.41
## site 3       1.00    1.73
## site 4       1.73    1.00    1.41
```

```
## Jaccard
round(dist.binary(x, method = 1),
```

```
##           site 1 site 2 site 3
## site 2       0.82
## site 3       0.50    0.87
## site 4       0.87    0.71    0.71
```

Distance vs Raw

- In some cases, both approaches can be equivalent (Euclidean distances)

For the univariate case, Euclidean distance is $d_{ij} = \sqrt{(x_i - x_j)^2}$ and we have:

$$\text{var}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2$$

```
x <- rnorm(10)
var(x) * 9/10
```

```
## [1] 0.4947034
```

```
sum(as.matrix(dist(x))^2)/(2 * 10^2)
```

```
## [1] 0.4947034
```

When using Euclidean distance, several raw- and distance-based analysis would produce the same results.

Your turn

1. Create an Rmd file
2. Load the `meaudret` data set from `ade4`
3. Perform the principal component analysis (`dudi.pca`) and principal coordinates analysis (`dudi.pco`) using Euclidean distances (`dist`) on faunistic data. Compare the outputs and conclude.
4. Transform the data into presence-absence (`ifelse(meaudret$spe>0, 1, 0)`). Perform principal coordinates analysis (`dudi.pco`) using Euclidean (`dist`) and Jaccard distances (`dist.binary`). Compare the results.
5. Perform the between-class analysis (`bca`) and distance-based RDA (`vegan::dbRda`) using Euclidean distances using the factor `meaudret$design$season` as an exploratory variable. Compare the results (be aware that `vegan` use $\frac{1}{n-1}$ to compute variances while `ade4` uses $\frac{1}{n}$).
6. Look at the percentage of variation explained by the between-class analysis (stored in the object). Perform permutational multivariate analysis of variance with the function `vegan::adonis` using the Euclidean distances. Compare

Summary

When using Euclidean distances, we have:

- Principal Coordinates Analysis \Leftrightarrow Principal Component Analysis
- Distance-based Redundancy Analysis \Leftrightarrow Redundancy Analysis (Between-Class Analysis when a single factor is used as explanatory variable)
- Permutational Distance-based Multivariate Analysis of Variance \Leftrightarrow Permutation test of Between-Class Analysis

