Multivariate Analysis: Redundancy Analysis (RDA)

BIOL 680 – Alex Engler

Guest lecture – 11/04/2023

Outline

Introduction

1. What is a RDA?

2. Constrained and Unconstrained Variances

3. Plotting and interpreting the RDA

4. Variance partitioning

5. Exercice (We will go through a RDA together)

7. Question session

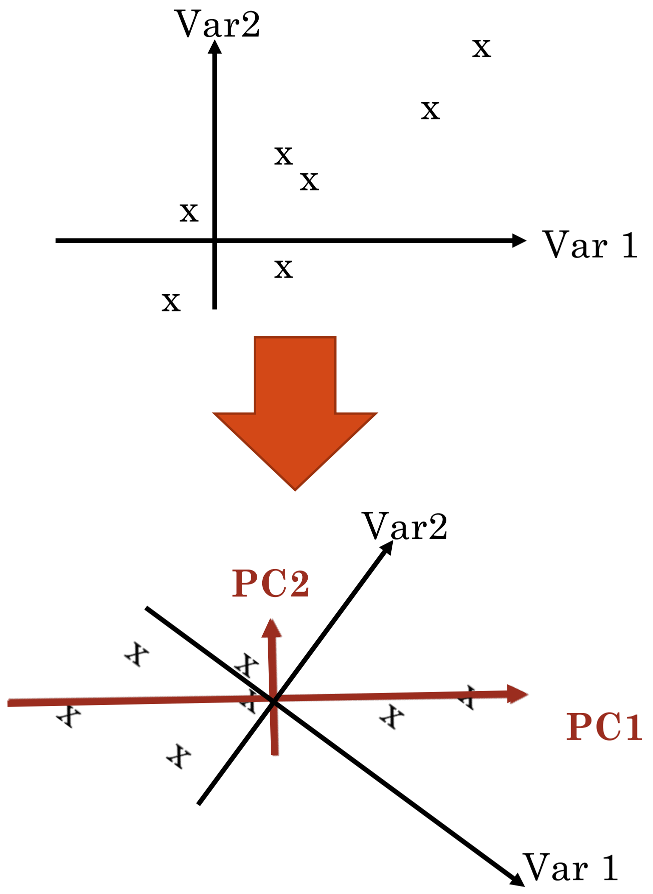
Introduction

Multivariate analysis

- The most common multivariate analysis is the Principal Component Analysis (PCA)

➔Summarise the colinearity of the variables

➔Reduce a n-dimension table to a smaller one



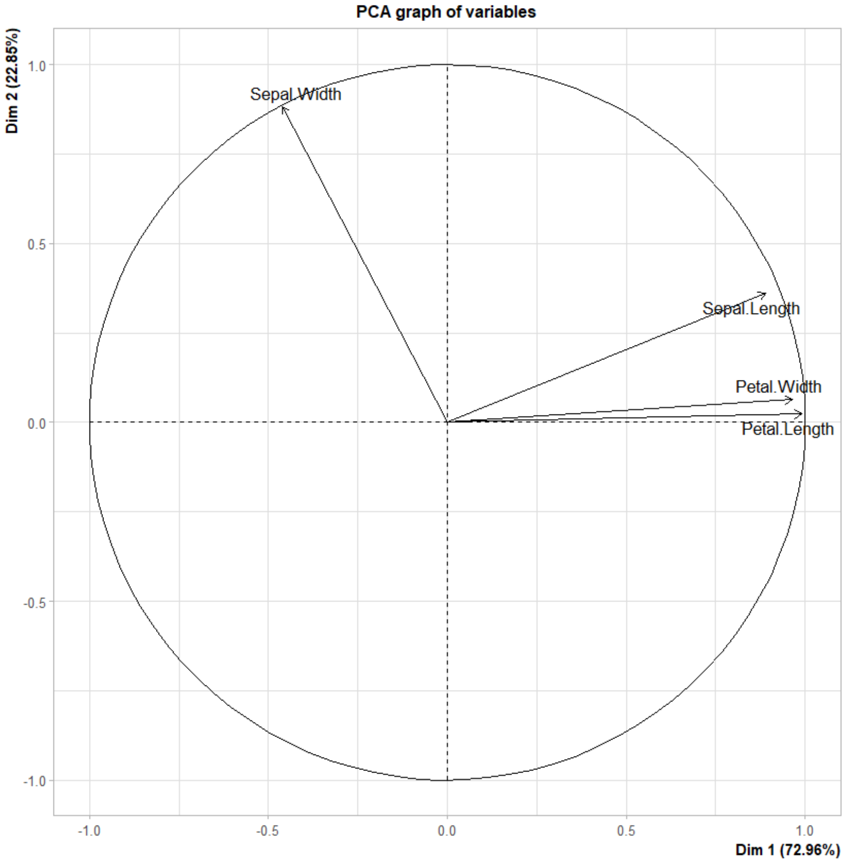
Multivariate analysis

- Multivariate analysis: when you have a table with many variables:

➔Want to understand their potential correlations

➔The similarities between individuals

➔Summarise complex information into fewer variables

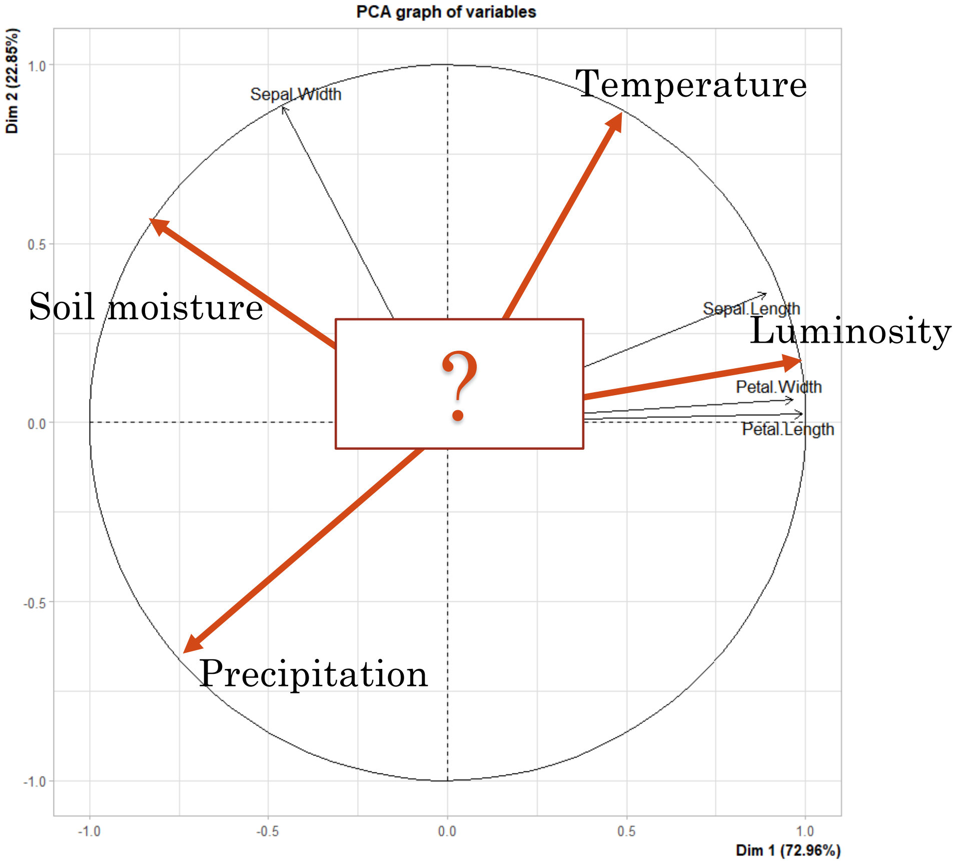


Multivariate analysis

- What to do when we want to explain a set of variables with another set of variables?

(example: explaining the plant's morphology by environmental variables)

➔ We run a Redundancy Analysis (RDA)



What is an RDA?

How do we build an RDA

Multiple regressions

- Multiple linear regression

$$Y \sim a + b_1X + b_2X + . . . + b_mX + \varepsilon$$

In this case:

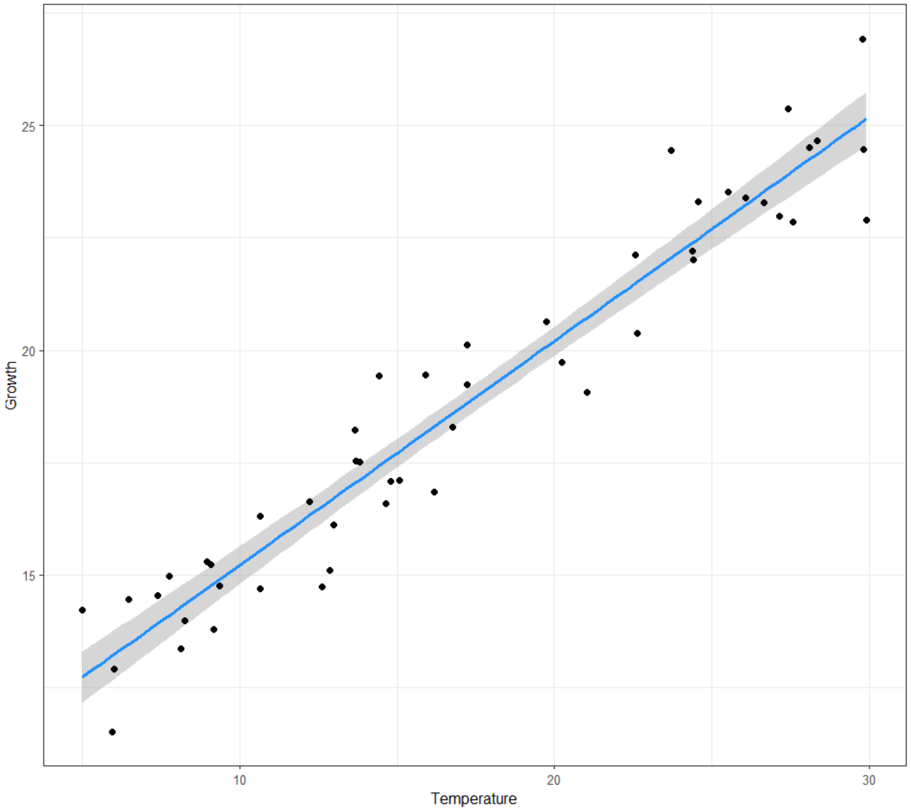
Y : response variable

X_m : explanatory variables

a: the intercept

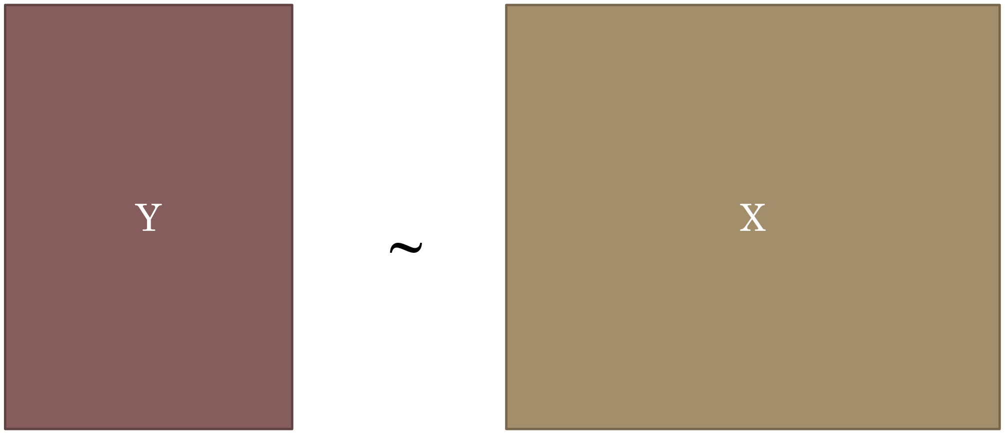
b_m: the slope related to Y_m

ε: the residuals



Multiple regressions

- An RDA: a direct extension of the multiple linear regression with multivariate data



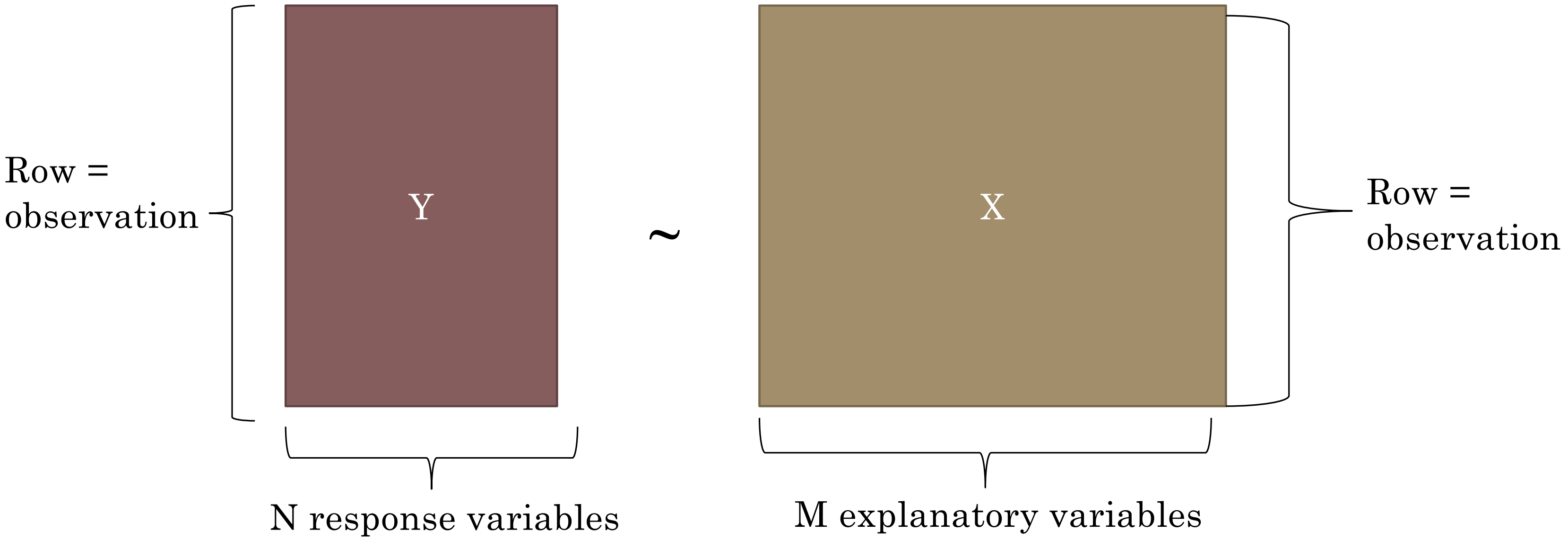
Set of response variables

Set of explanatory variables

BE CAREFUL YOU SHOULD HAVE THE SAME NUMBER OF OBSERVATIONS (ROWS) IN BOTH TABLES

Multiple regressions

- An RDA: a direct extension of the multiple linear regression with multivariate data



Row = observation

Y

~

X

Row = observation

N response variables

M explanatory variables

BE CAREFUL YOU SHOULD HAVE THE SAME NUMBER OF OBSERVATIONS (ROWS) IN BOTH TABLES

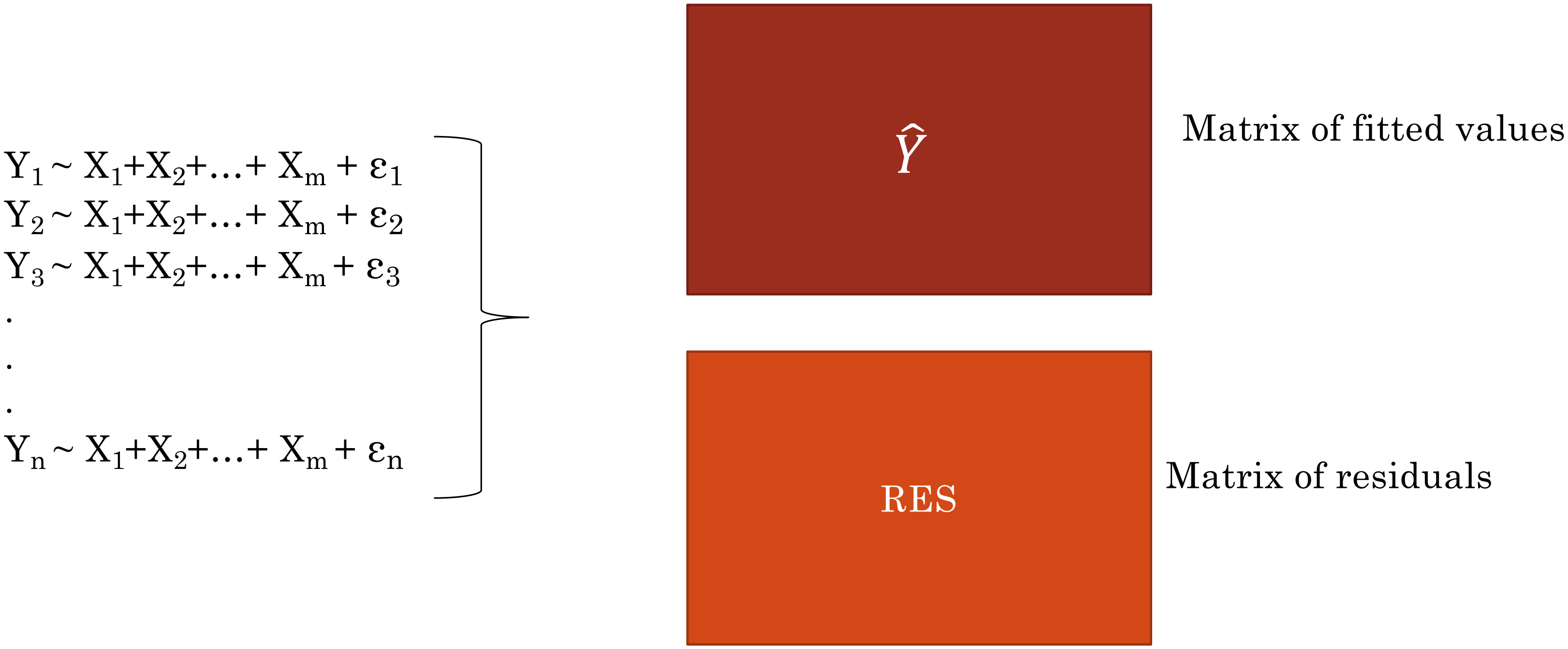
Multiple regressions

- FIRST STEP : regress each response variable by the explanatory variables

$$
\begin{aligned}
&Y_1 \sim X_1+X_2+\ldots+ X_m + \varepsilon_1\\
&Y_2 \sim X_1+X_2+\ldots+ X_m + \varepsilon_2\\
&Y_3 \sim X_1+X_2+\ldots+ X_m + \varepsilon_3\\
&.\\
&.\\
&.\\
&Y_n \sim X_1+X_2+\ldots+ X_m + \varepsilon_n
\end{aligned}
$$

Multiple regressions

- STEP 2: extract the fitted values and the residuals for each linear model



$Y_1 \sim X_1+X_2+\ldots+X_m+\varepsilon_1$

$Y_2 \sim X_1+X_2+\ldots+X_m+\varepsilon_2$

$Y_3 \sim X_1+X_2+\ldots+X_m+\varepsilon_3$

.

.

.

$Y_n \sim X_1+X_2+\ldots+X_m+\varepsilon_n$

$\hat{Y}$

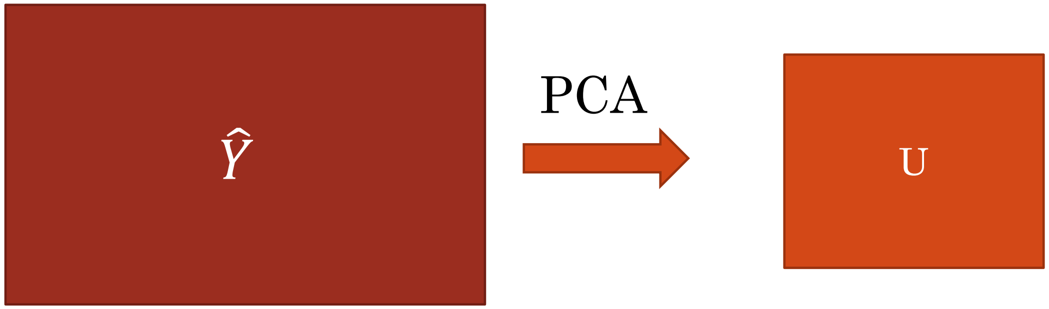
Matrix of fitted values

RES

Matrix of residuals

Multiple regressions

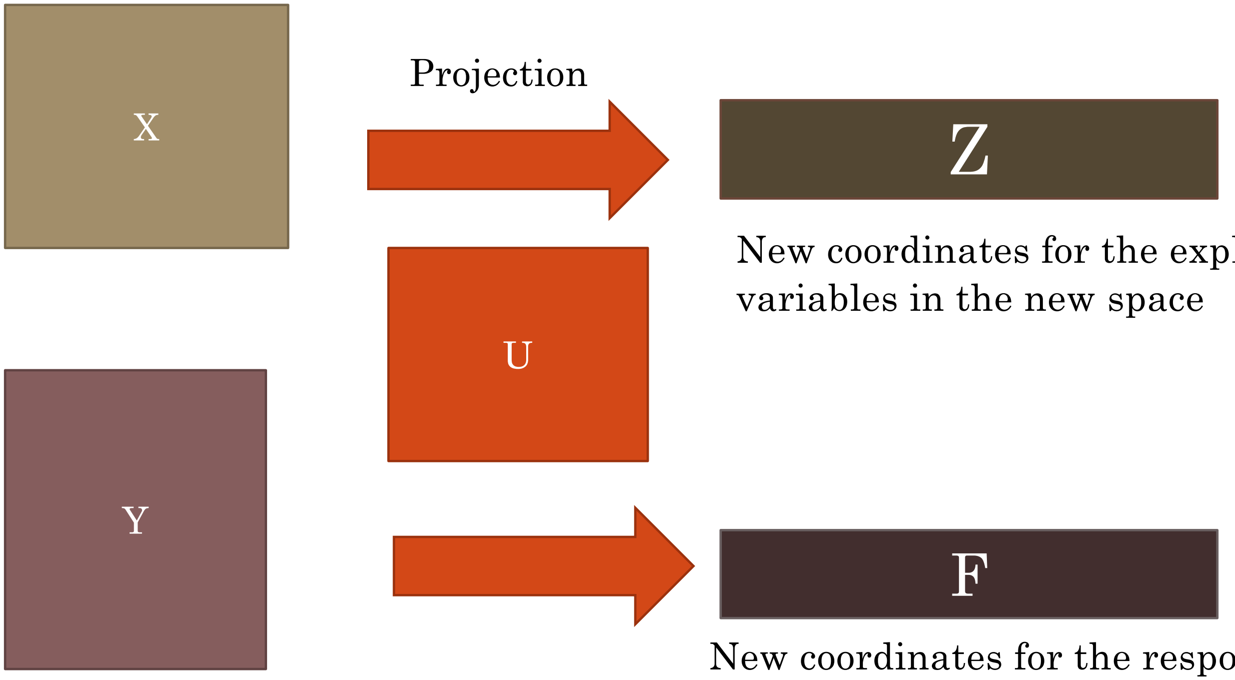
- STEP 3: pca on the fitted values



Matrix of eigen-values
(creation of the new
multivariate space)

Multiple regressions

- STEP 4: projection of the raw data and fitted values in the new space

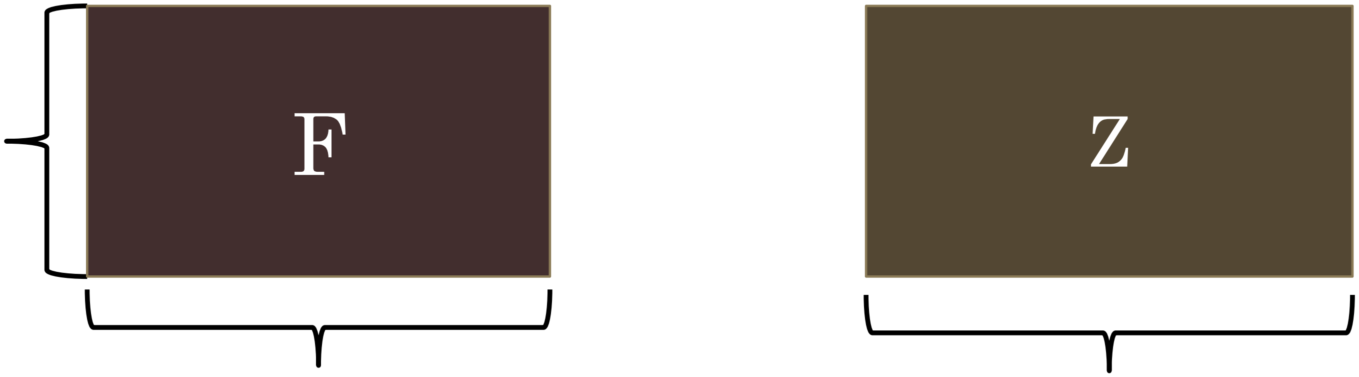


New coordinates for the explanatory variables in the new space

New coordinates for the response variables in the new space

Multiple regressions

- STEP 4: projection of the raw data and fitted values in the new space



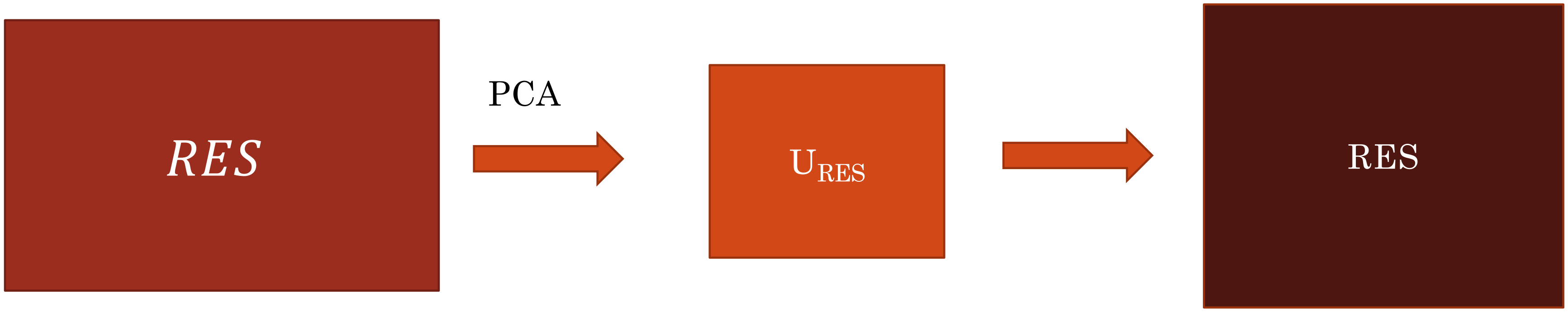
Rows:
sites

New coordinates for the response variables in the new space

New coordinates for the explanatory variables in the new space

Multiple regressions

- STEP 5: pca on the residuals



RES

PCA

U_{RES}

RES

Matrix of eigen-values (creation of the new space for residuals)

Projection of the residuals in the new space

How do we build an RDA

An RDA is two PCAs in a trenchcoat:

- One on the fitted values (summarising the results of multiple linear régressions)
- One on the residuals (summarising the variability that was not catch by the linear régressions)

The results of the RDA

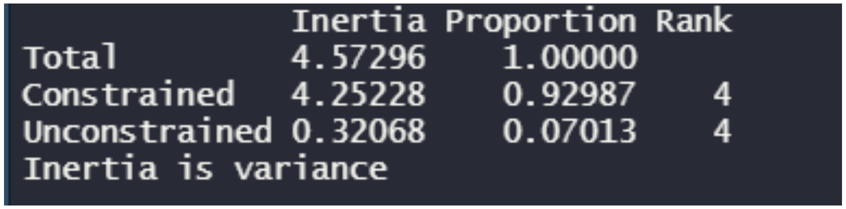
Constrained and unconstrained parts of the RDAs

An RDA is two PCAs in a trenchcoat:

- The fitted model (constrained)
- The residuals (unconstrained)

Constrained and unconstrained variance

When you run a RDA, you will have this table :

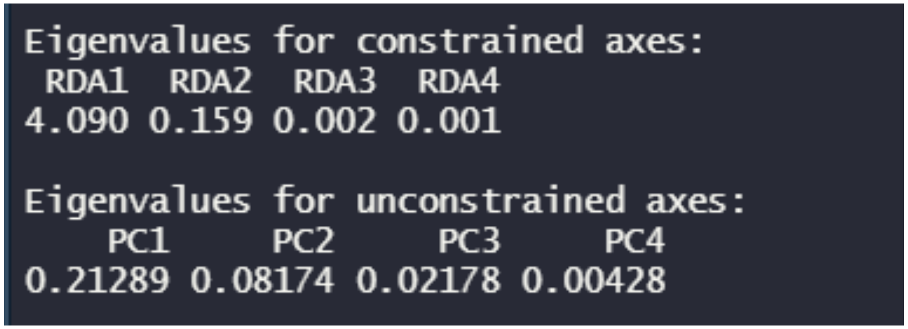


```
              Inertia Proportion Rank
Total         4.57296    1.00000
Constrained   4.25228    0.92987    4
Unconstrained 0.32068    0.07013    4
Inertia is variance
```

Inertia is synonym to variance

Total is the total variance of your response variables
Constrained is the variance explained by your linear model
Unconstrained is the remaining variance that is not explained by the model

Constrained and unconstrained axis



```
Eigenvalues for constrained axes:
 RDA1  RDA2  RDA3  RDA4
4.090 0.159 0.002 0.001

Eigenvalues for unconstrained axes:
    PC1     PC2     PC3     PC4
0.21289 0.08174 0.02178 0.00428
```

Similarly to the PCA, you can look at the variance explained by each axis:

The constrained axis are the variance for the fitted model: the sum of the « eigenvalues » should equal to the inertia of the constrained model

The unconstrained axis is variance of the residuals. We don’t usually interpret them.

The performance of the RDA

- Similarly to a regular linear regression, you have a R^2, that represents the overall fit of the models (the adjusted R^2 take in account the number of explanatory variables)

- There is F-statistics that compare the model with a null model

➔ H_0 : the strength of the linear relationship, measured by the canonical R2, is not larger than the value that would be obtained for unrelated Y and X matrices of the same sizes

➔ We can test for the whole model or for each axis of the model or for each explanatory variables !

Plotting an RDA

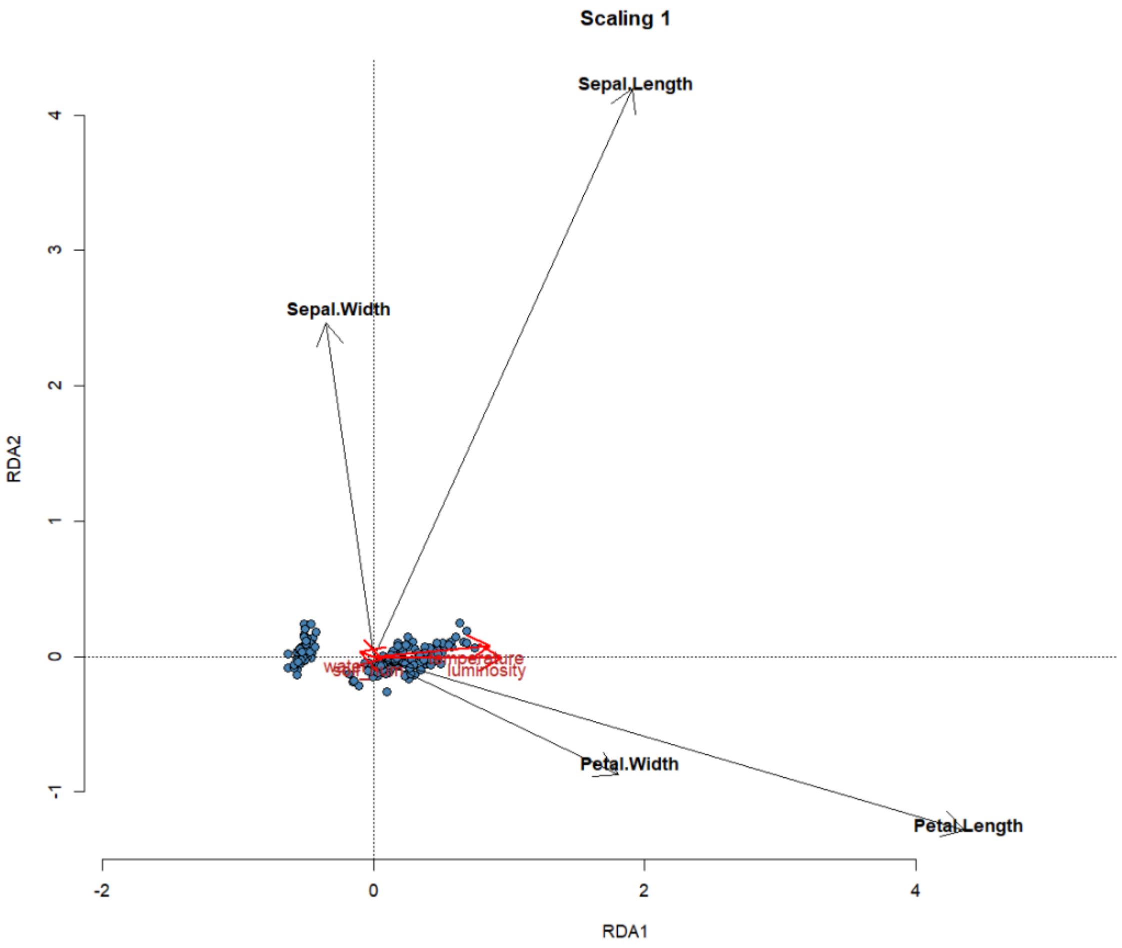
2 types of plot in RDA

- **Scaling 1 or distance triplot**

➔The distance between sites (points) is an approximation of the euclidean distance

➔The relationship between explanatory variables is preserved

➔The angles between explanatory variables (red arrows) are meaningless

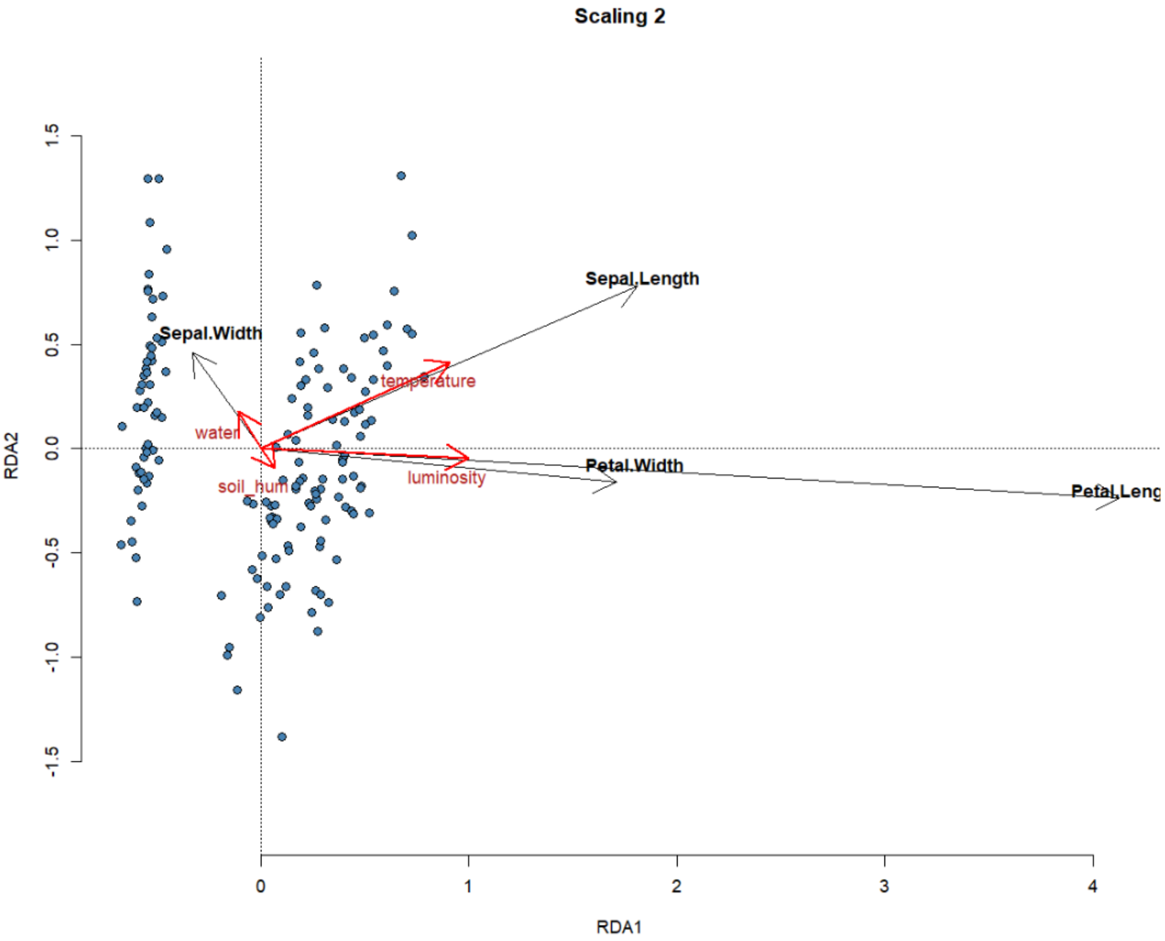


2 types of plot in RDA

- **Scaling 2 or correlation triplot**

➔The angle of two variables is an approximation of the correlations between the two variables

➔The distance between two objects does not reflect their Euclidean distance



Summary

- An RDA is just a linear regression with multiple response variables and multiple explanatory variables

$$Y \sim X + RES$$

- An RDA is two PCA: one that will capture the variance of the linear models and the other the residuals

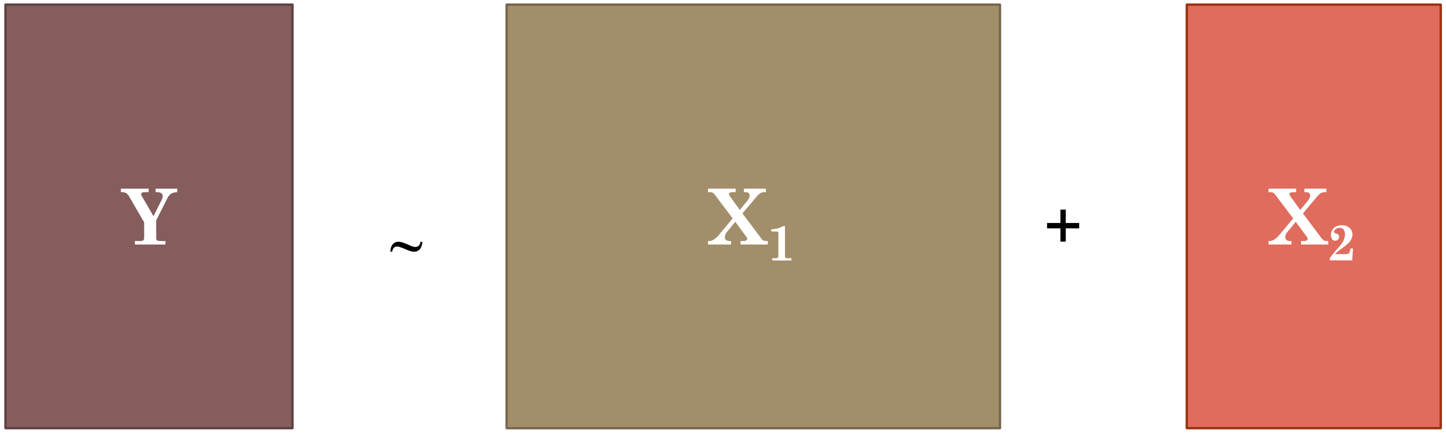
- You can either plot accurately the distance between observations or the relationships between the response variables. In both case, the correlations between response and explanatory variables is conserved

Variance partitioning in RDAs

Variance partition

- Type of analysis for RDAs when we separate the explanatory variables in several sets of explanatory variables
- Goal: Understand what is the variance explained by a specific set of variables

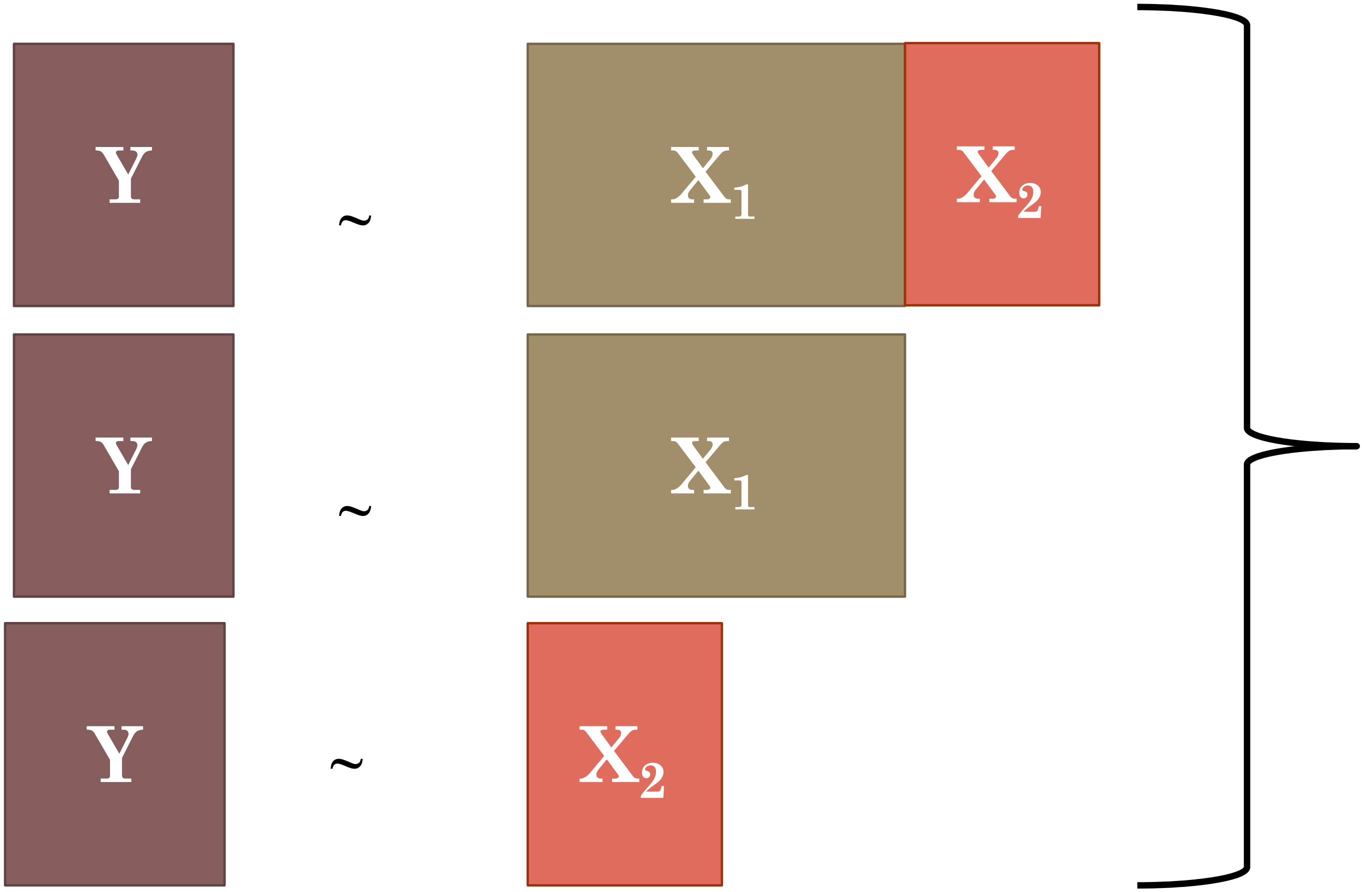
➔ Very useful to distinguish the effects of variables when there are confounding variables



$Y \sim X_1 + X_2$

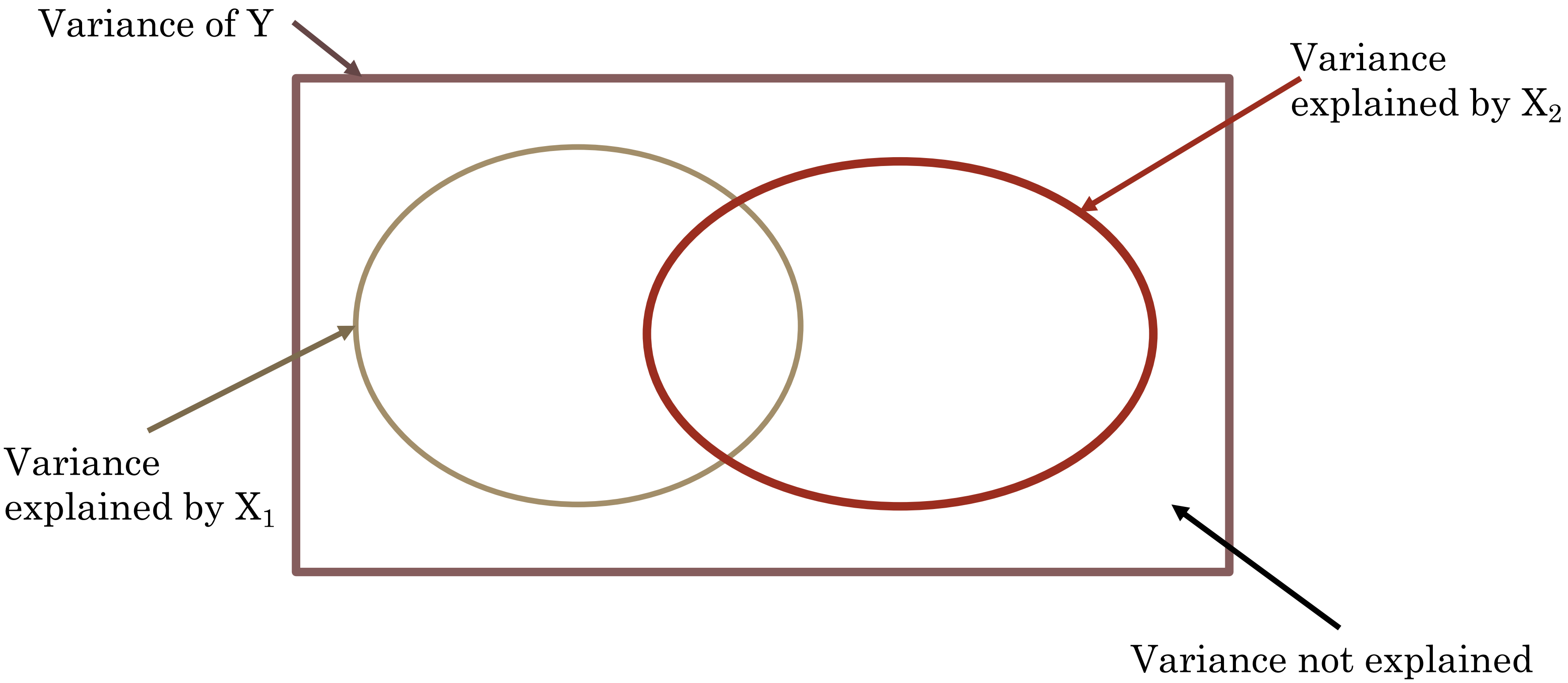
Variance partition

- It consists in three consecutive RDAs

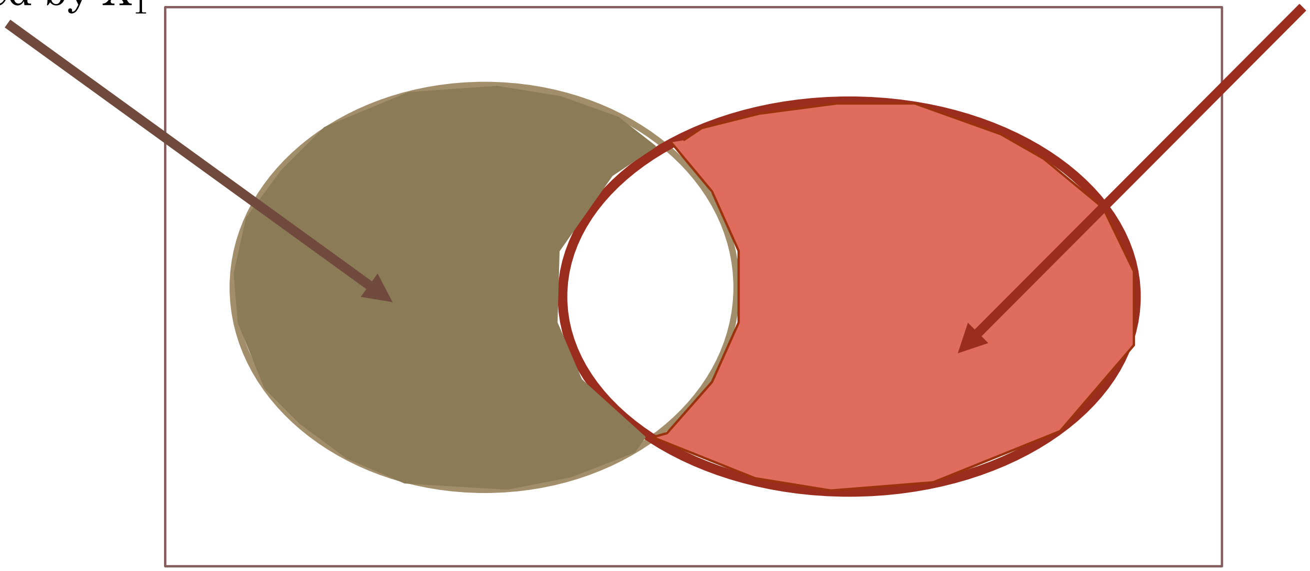


We extract the variance explained by each model

Variance partition



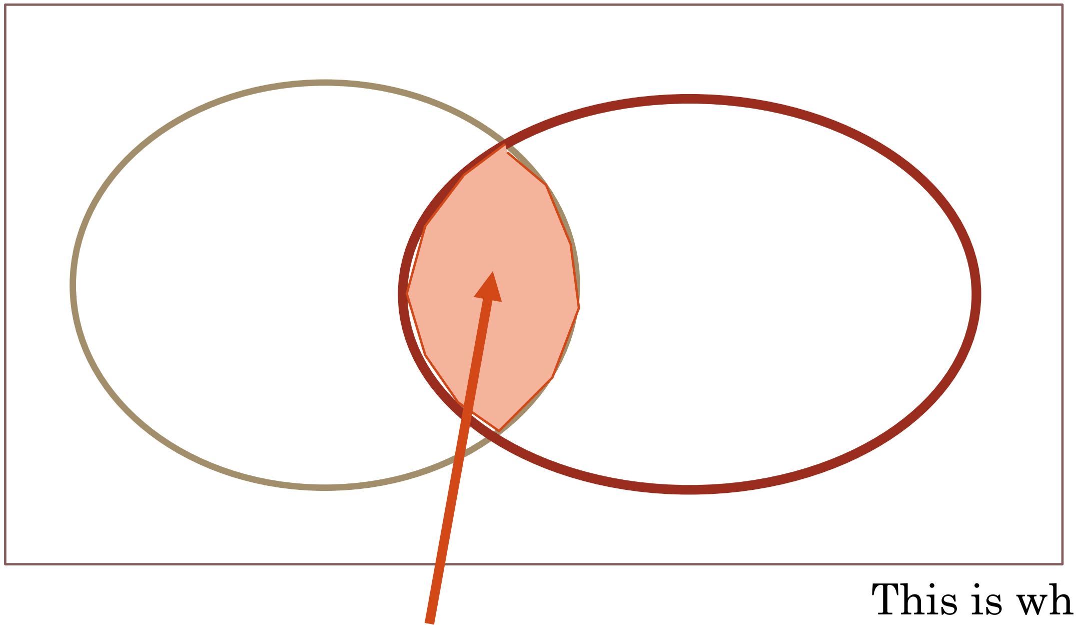
Variance partition



Variance only explained by X_1

Variance only explained by X_2

Variance partition



Variance explained by both X_1 and X_2 → This is where the two models are confounded Should be as small as possible

Summary of the variance partitioning

- It is a method to understand where confounding effects are in our models : How much of the variance could have been explained by different variables ?

➔Can be very useful when you have geographically structured data

- What to do when you have a lot of shared variance between two variables:

➔You can not separate the effect of one or the other

➔You have to redesign your experiment/data collection to be able to separate the effects of the two variables

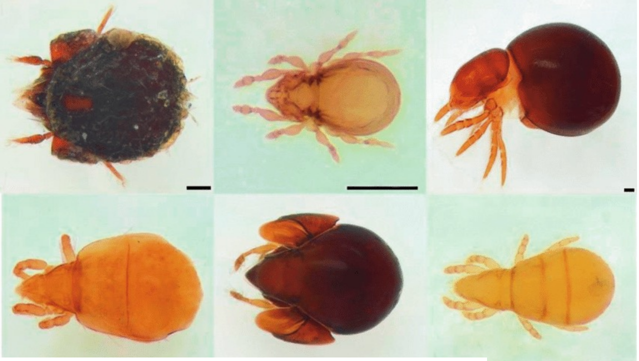
Little Break

5 minutes

Exercice

Let’s do an RDA together!

Exercice : How the environmental variables impact the communities composition ?

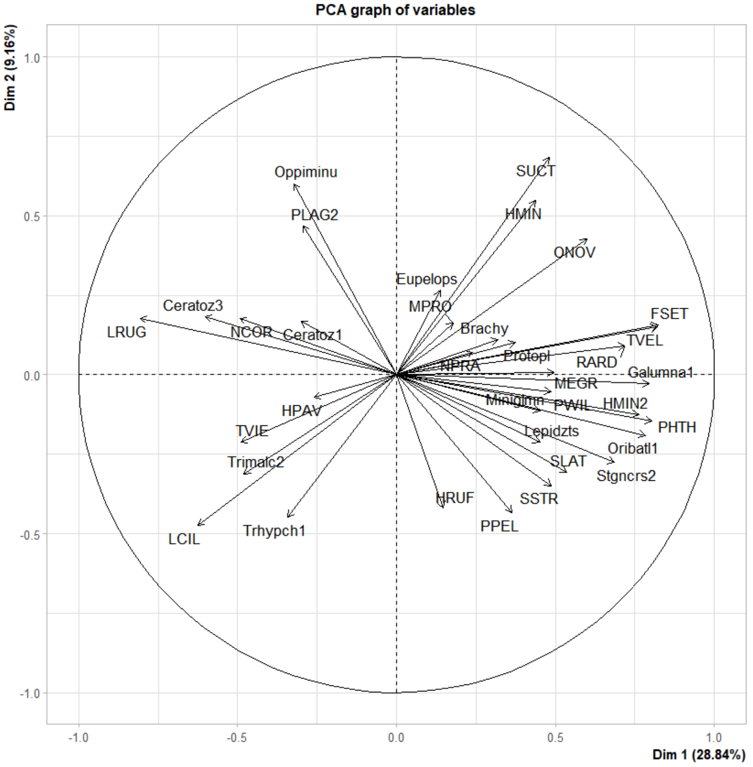


In 1989: Daniel Brocard sampled 75 sites and described the mites communities on those sites

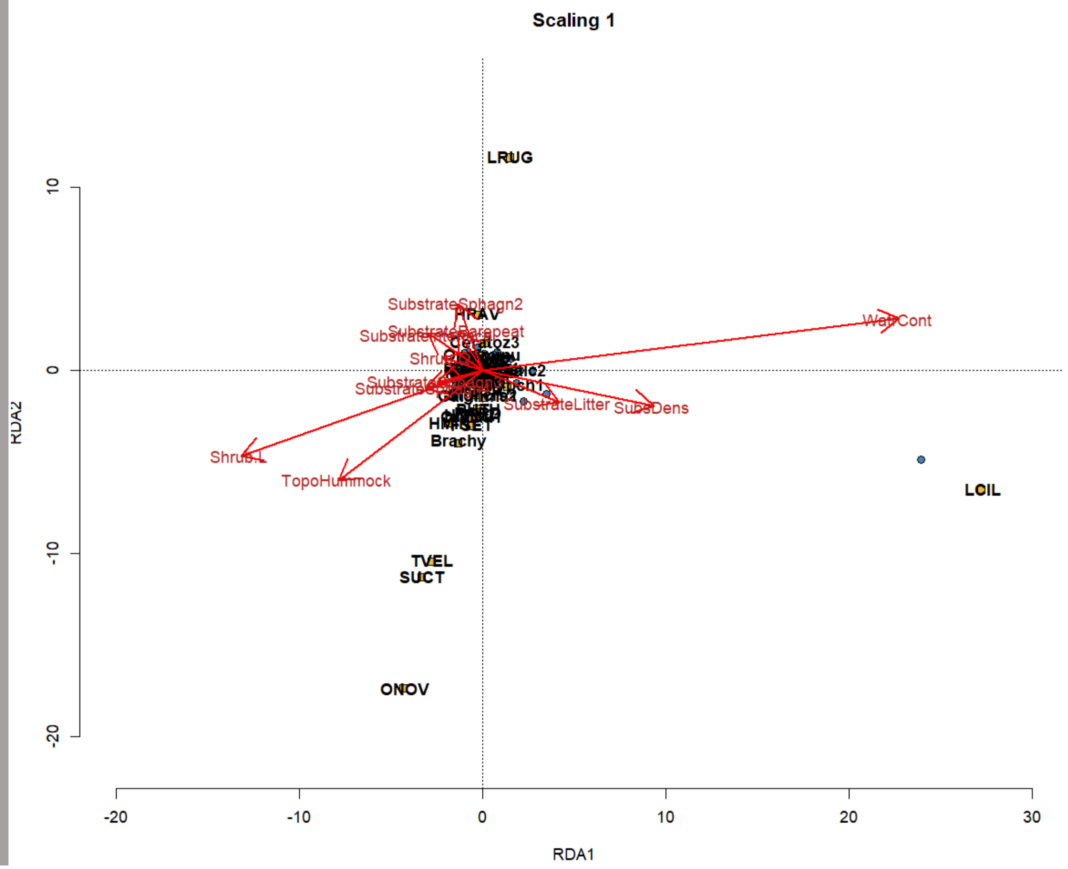
There were 35 species recorded in total.

He measured 5 environmental variables

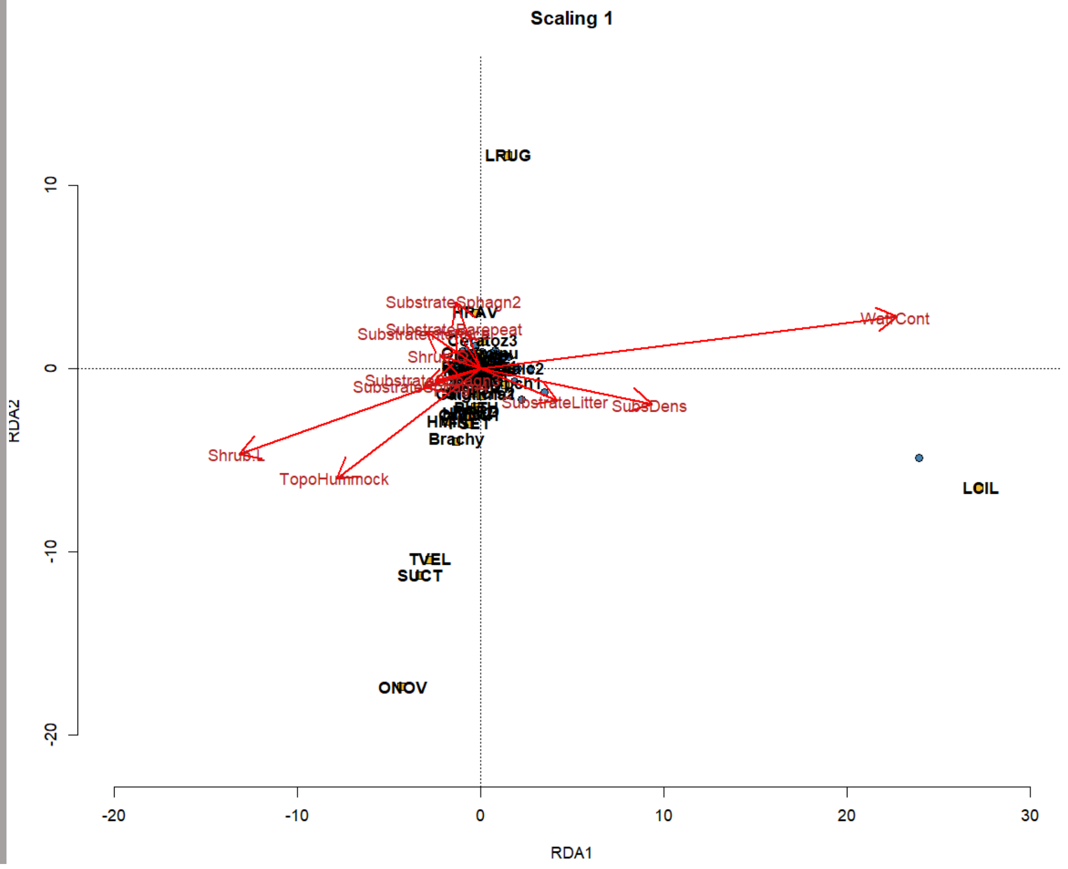
Goal: Which environmental variables drive the species composition of the mites communities



What can you tell me about the correlations between the response and explanatory variables?



What can you tell me about the correlations between the response and explanatory variables?

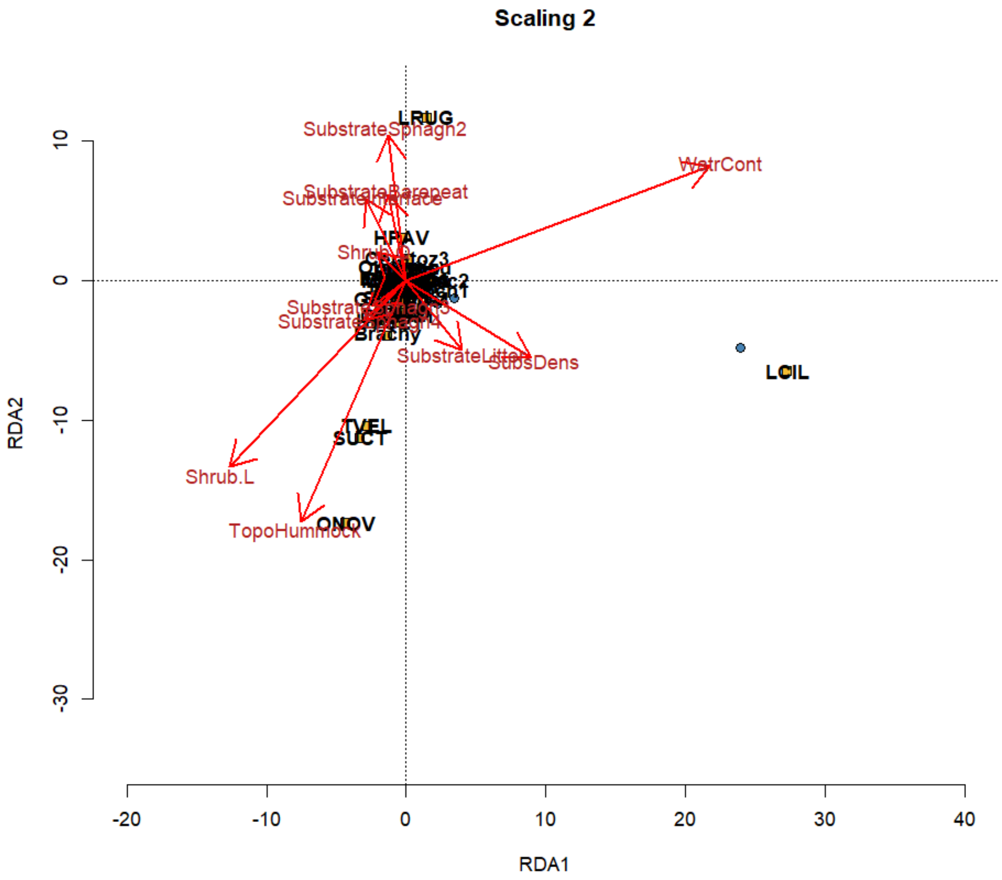


We are in Scaling 1 :

- The distances between sites are preserved

➔ One site is very different from the others and it probably due to the presence of one species

What can you tell me about the correlations between the response and explanatory variables?

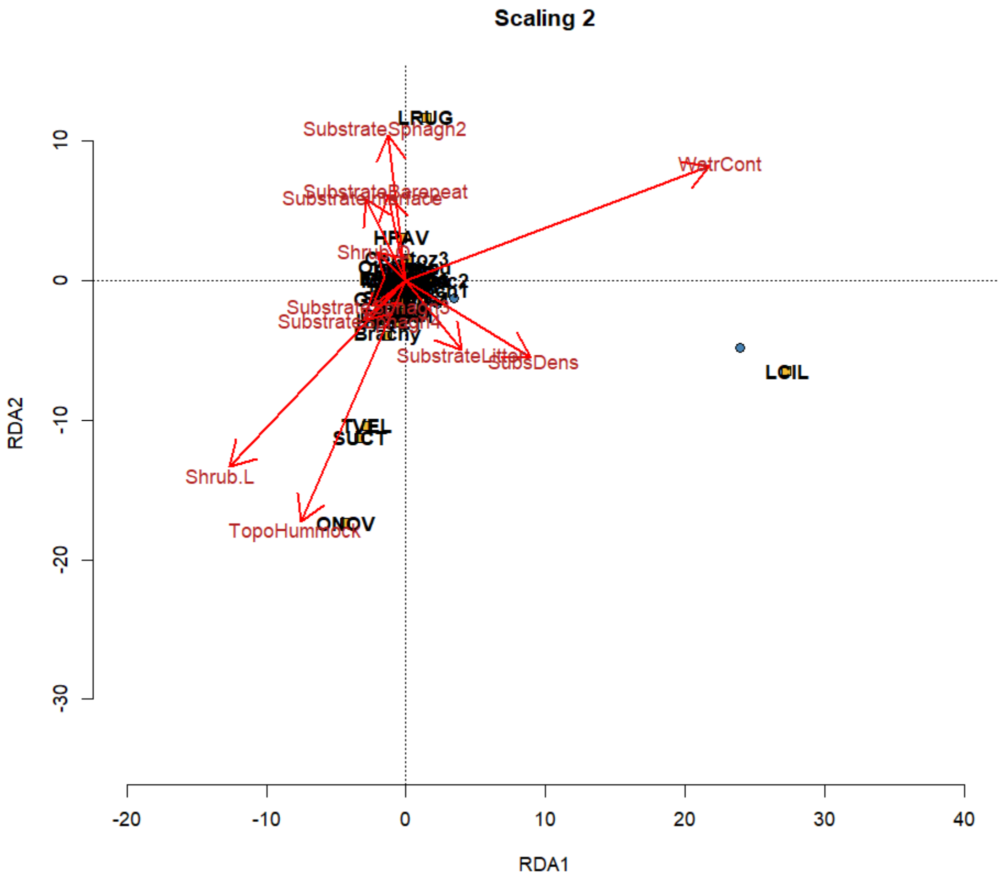


What can you tell me about the correlations between the response and explanatory variables?

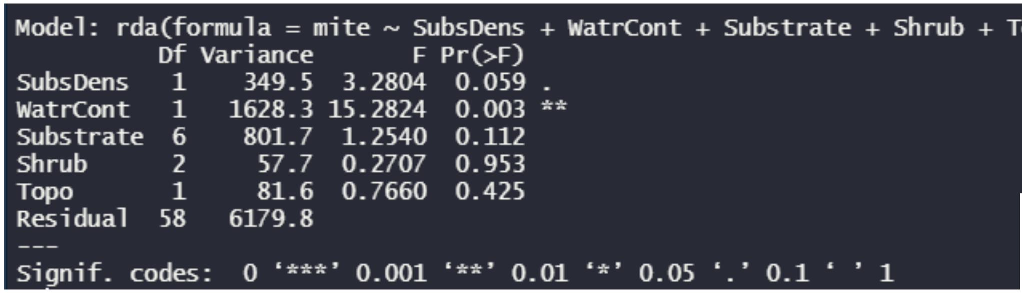
We are in scaling 2: We can only interpret the angles between variables

➔Water Content drive the first axis

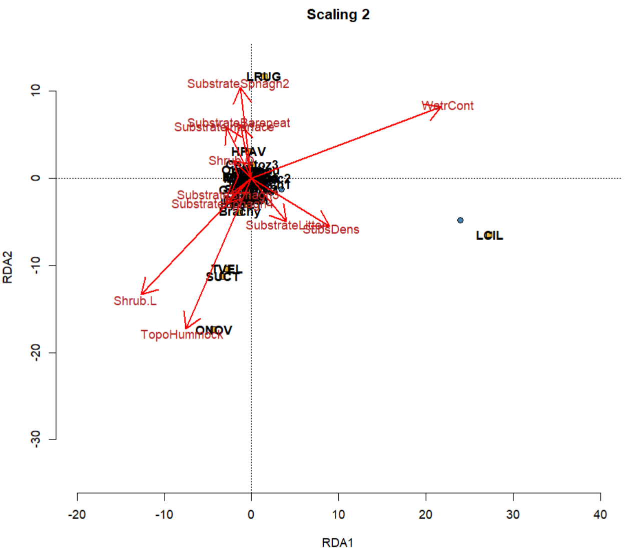
➔Topology, and different substrate driving the second axis.



Are there significant environmental variables that drive the mite communities?

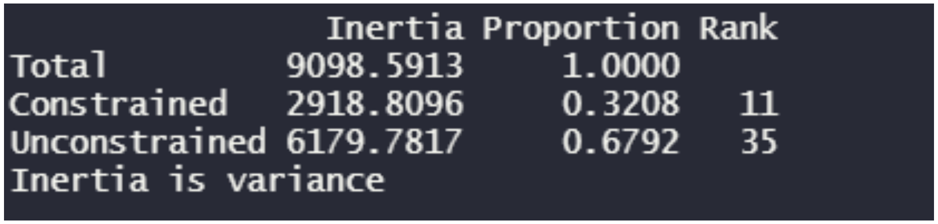


```
Model: rda(formula = mite ~ SubsDens + WatrCont + Substrate + Shrub + Topo, data = mite.env)
          Df Variance       F Pr(>F)
SubsDens   1    349.5  3.2804  0.059 .
WatrCont   1   1628.3 15.2824  0.003 **
Substrate  6    801.7  1.2540  0.112
Shrub      2     57.7  0.2707  0.953
Topo       1     81.6  0.7660  0.425
Residual  58   6179.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



What are the performance of the RDA

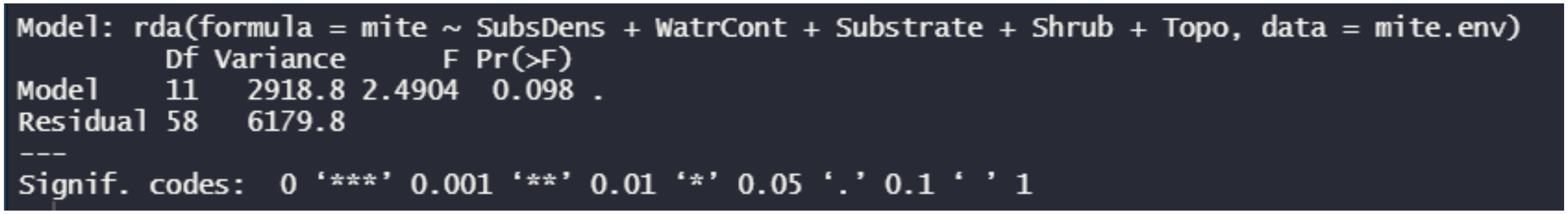
How much variance is explained by our model? How much is left unexplained ?



```
                 Inertia Proportion Rank
Total          9098.5913     1.0000
Constrained    2918.8096     0.3208   11
Unconstrained 6179.7817      0.6792   35
Inertia is variance
```

Are the results of our model significant?

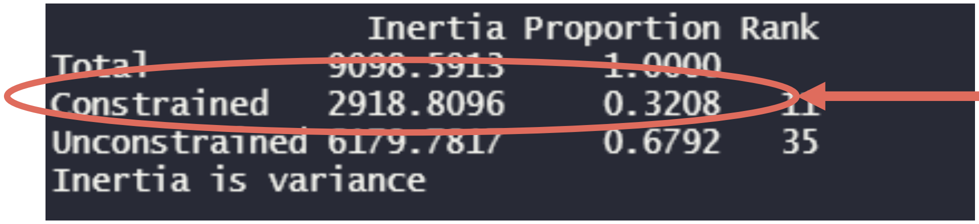
Test for the model significance



```
Model: rda(formula = mite ~ SubsDens + WatrCont + Substrate + Shrub + Topo, data = mite.env)
         Df Variance      F Pr(>F)
Model    11   2918.8 2.4904  0.098 .
Residual 58   6179.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What are the performance of the RDA

How much variance is explained by our model? How much is left unexplained ?

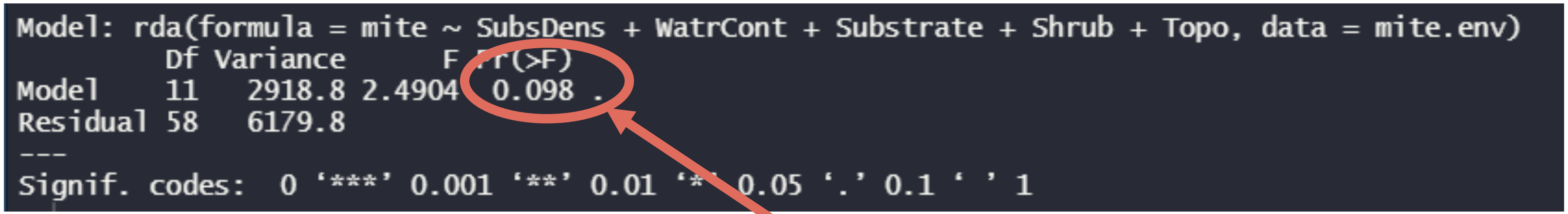


```
                Inertia Proportion Rank
Total          9098.5913     1.0000
Constrained    2918.8096     0.3208   11
Unconstrained  6179.7817     0.6792   35
Inertia is variance
```

The model explains 32% of the variance

Are the results of our model significant?

Test for the model significance



```
Model: rda(formula = mite ~ SubsDens + WatrCont + Substrate + Shrub + Topo, data = mite.env)
         Df Variance      F Pr(>F)
Model    11   2918.8 2.4904  0.098 .
Residual 58   6179.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Not significant

TIPS TO RUN A RDA

- SCALE YOUR EXPLANATORY AND RESPONSE VARIABLES (avoid to detect effect solely due to difference of units)
- Try to reduce as possible collinearity between explanatory variables as much as possible before running the rda (with a PCA, for instance)
- Make sure you have the good number of observations in both set of variables
- Don't forget to check the percentage of explained variance
- You can do a variance partition to desentangle the specific effect of one set of variables

Bibliography

Legendre, P., & Legendre, L. *(2012). Numerical ecology. Elsevier*

Iris dataset: Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society, 59, 2–5 (found in "datasets" package)

Mite dataset: Borcard, D., P. Legendre and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045-1055. (found in "vegan" package)

Thank you for listening

TIME FOR QUESTIONS !