MarMic Statistic II

Tutotrial

Emiliano Pereira

Eigen decomposition

Eigen decomposition is the factorization of a matrix, where that matrix is represented in terms of its **eigenvalues** and **eigenvectors**.

Eigendecomposition

A vector **v** of dimension N is an eigenvector of a square (N×N) matrix **A** if it satisfies the linear equation:

$$\mathbf{Av} = \lambda\mathbf{v}$$

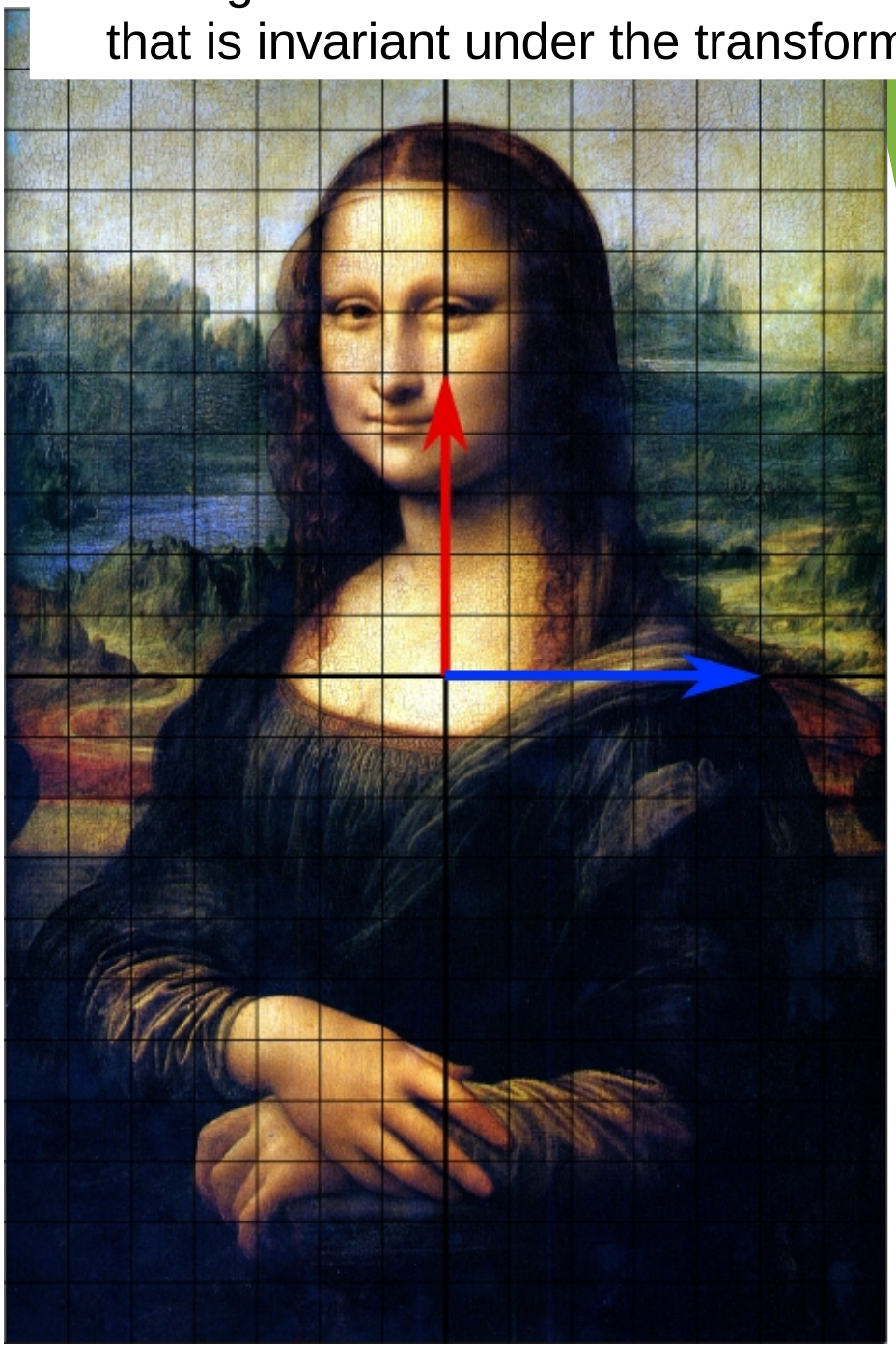
where **λ** is a scalar, termed the eigenvalue corresponding to **v**.

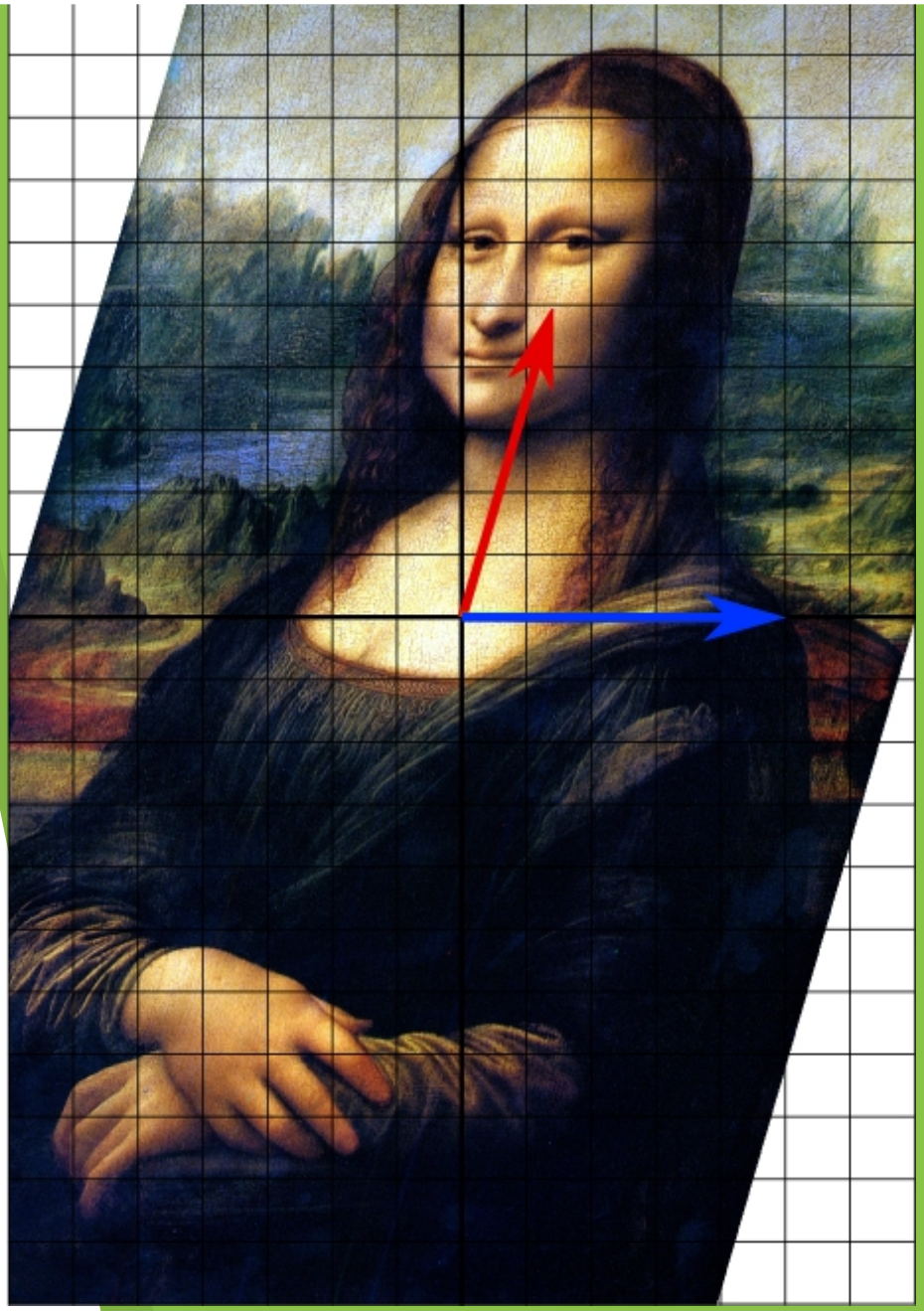
Eigendecomposition

$$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \times \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 2 \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \times \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} 1 \\ -5 \end{pmatrix}$$

“An eigenvector or characteristic vector of a linear transformation defines a direction that is invariant under the transformation.” – WP: Eigenvalues_and_eigenvectors



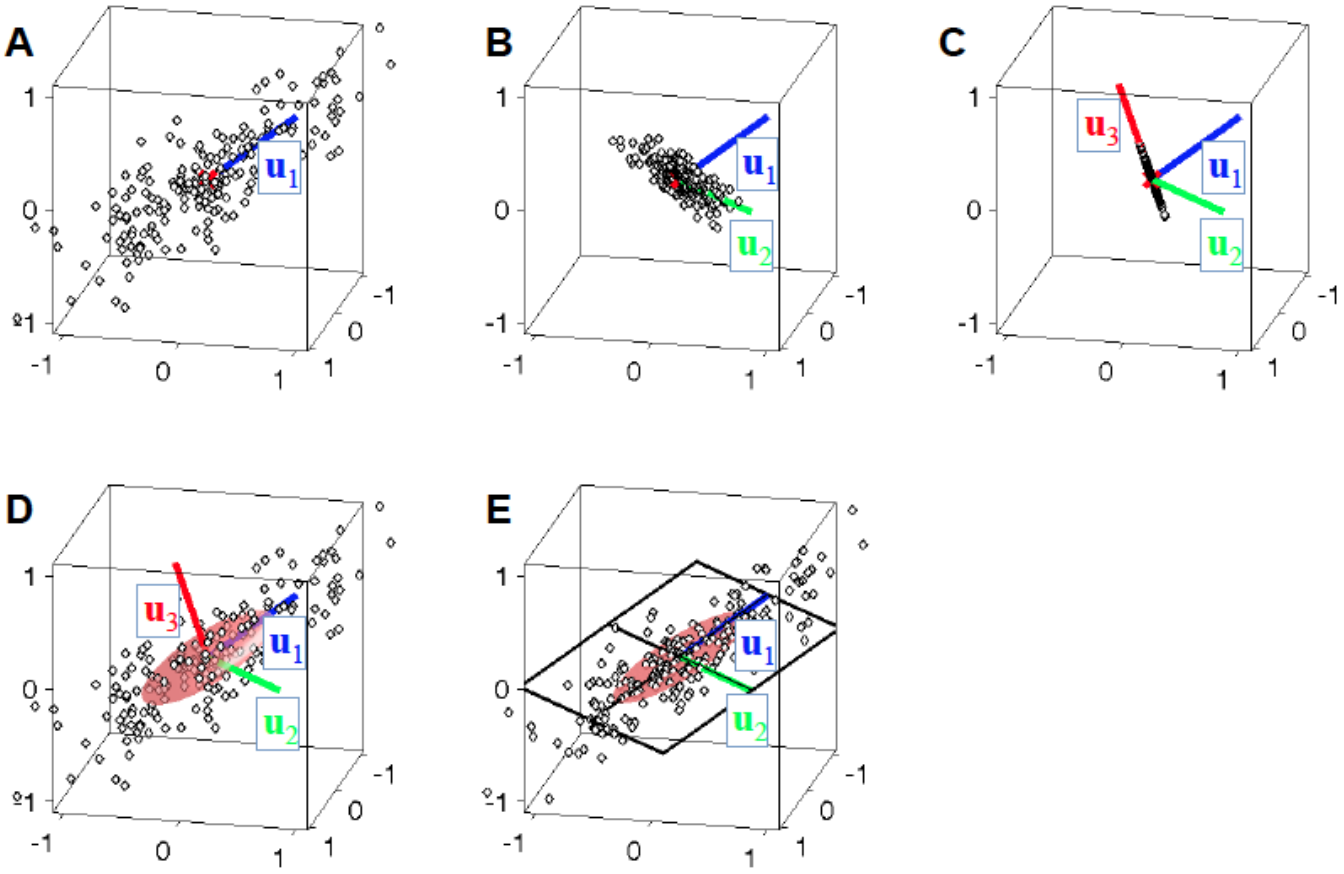


Wikipedia Commons: TreyGreer62

Principal Components Analysis

Principal components analysis (PCA) is a method to summarise, in a low-dimensional space, the variance in a multivariate scatter of points. In doing so, it provides an overview of linear relationships between your objects and variables.

Principal Components Analysis



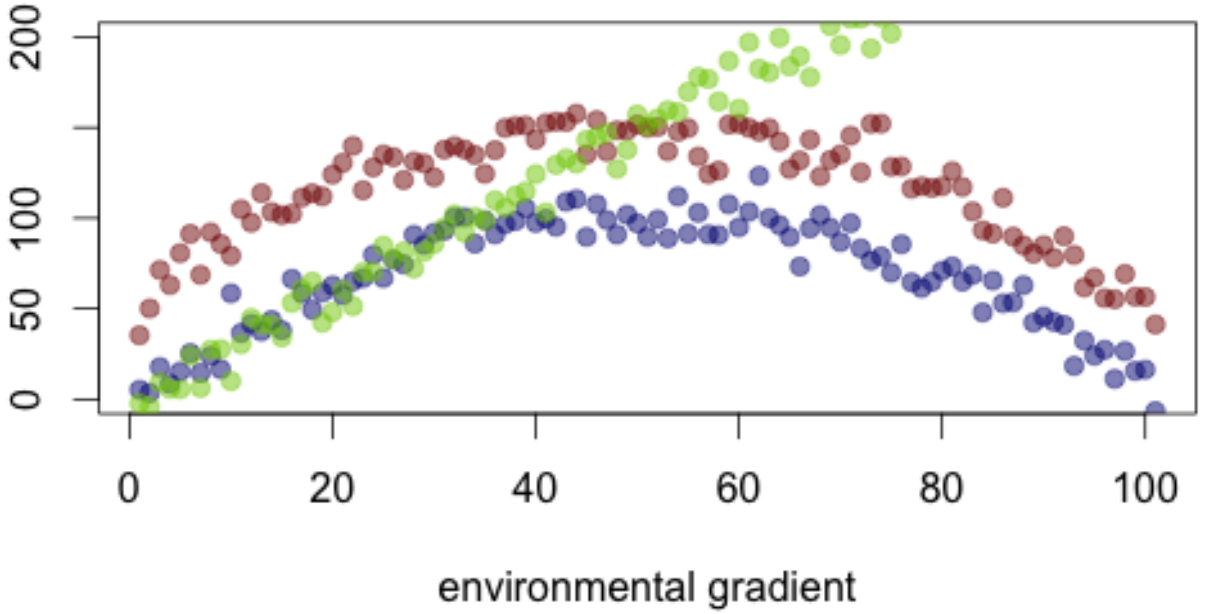
Principal Components Analysis

- PCA is useful when we expect species to be linearly (or even monotonically) related to each other.

- It is likely that species respond unimodally to an environmental gradient. This will be seen if you sample enough along the gradient. That is, species usually peak in abundance at some part of the environmental gradients.

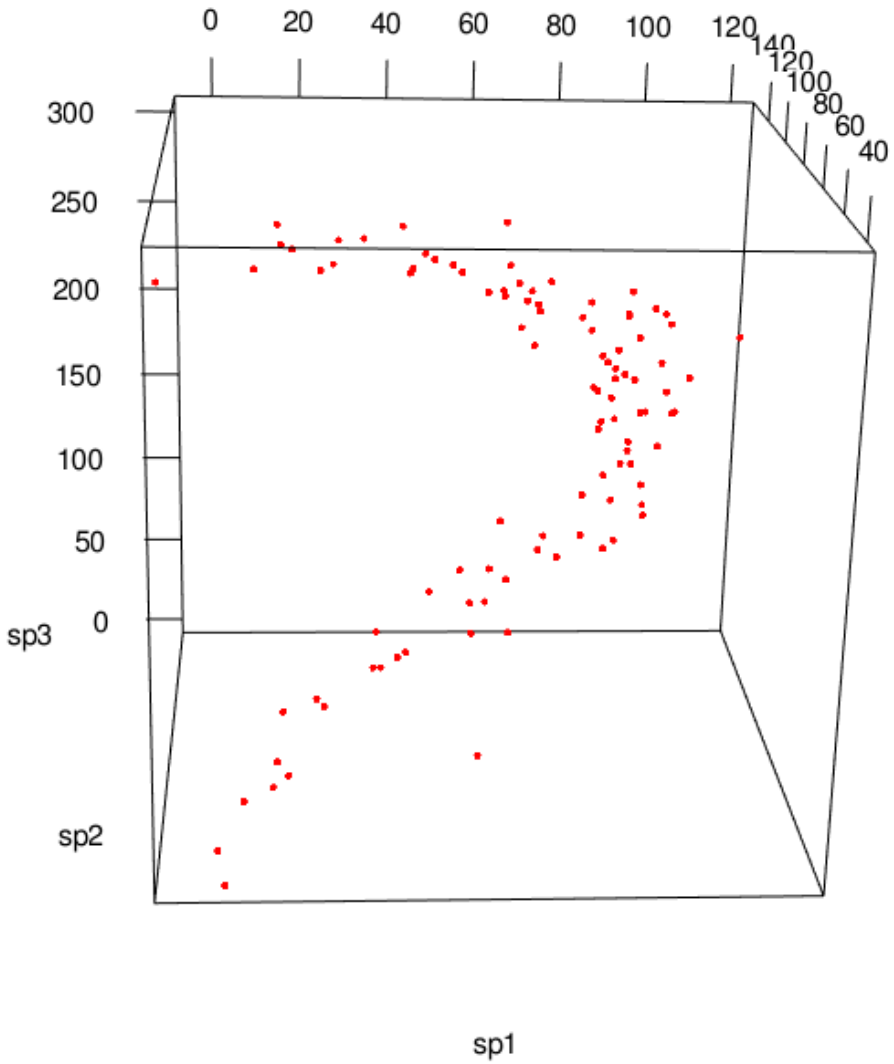
Principal Components Analysis

Species abundance response curve



Principal Components Analysis

Species abundance scatterplot



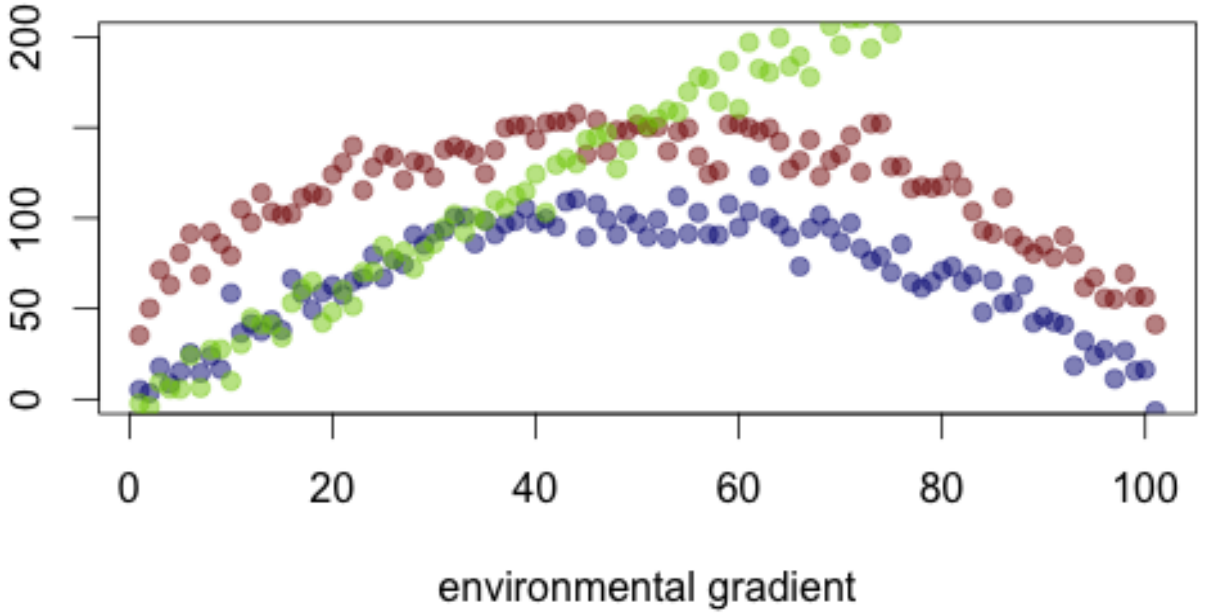
Principal Components Analysis

In PCA, **the horseshoe effect** is an artifact in which the second axis is curved and twisted relative to the first, and does not capture variation that is ecologically independent to the first PC.

This happens when the variables are not (approximately) linearly related to each other and show an unimodal response curve along a gradient

Principal Components Analysis

Species abundance response curve



Principal Components Analysis

Horseshoe ffect possible solutions:

1) Apply correspondence analysis (CCA)

2) Apply Hellinger transformation (what are the consequences for the interpretation?)

3) Subset your data keeping in mind the original question.

Principal Components Analysis

Input data:

1) Variables are linearly or at least monotonically related to one another

2) The variables should show a multivariate normal distribution (it practice this is rarely met, approximately normal is usually ok)

Note: It is better suited for environmental data, rather than species data, especially if the species are not linearly related to each other (given the horseshoe effect).

Principal Components Analysis

Standardized data: correlation matrix
Should be used when the data are in different units (e.g. environmental data).

Non-standardized data: covariance matrix
Should be used when the different variances in the data are informative (e.g. species data).

Biplot: variables and objects

There is no way to display objects (e.g. sites) and variables (e.g. species) in the same biplot diagram while visualizing all their inter relationships accurately.

Biplot: variables and objects

Scaling 1 - Choose this scaling if the main interest is to interpret relationships among objects.

1) Distances among objects in a biplot are approximations of their Euclidean distances (assuming there's no previous transformation).

2) Projecting an object at right angle on a variable approximates the value of the object along that variable.

3) The angles among variables are meaningless

Biplot: variables and objects

Scaling 2 - Choose this scaling if the main interest focuses on the relationships among variables (e.g. species).

1) Distances among objects in the biplot don't approximate their Euclidean distances.

2) Projecting an object at right angle on a variable approximates the value of the object along that variable.

3) The angles between variables in the biplot reflect their correlations.

Correspondance Analysis (CA)

- Similar to PCA, but applies to categorical rather than continuous data.
- Based on chi-squared distance rather than euclidean distance.
- Treats rows and columns equivalently.
- Traditionally applied to contingency tables: CA performs eigen decomposition on a transformation of the original data table. Each value in the transformed data table represents a contribution to the chi-squared statistic computed for the whole table.
- Tries to maximise the representation of 'correspondence' between the rows and columns, rather than the amount of variance.

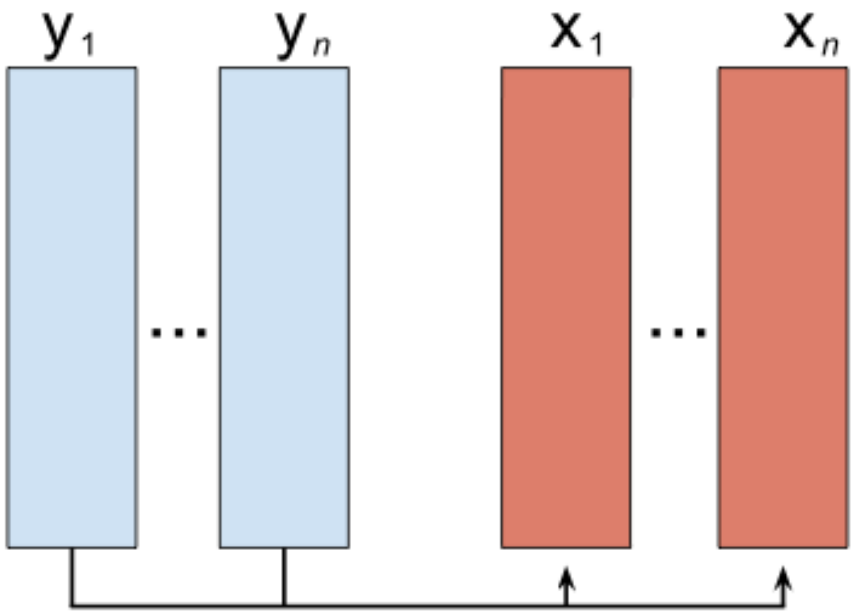
Canonical Correspondance Analysis (CCA)

It is a direct gradient analysis, wherein a matrix of explanatory variables intervenes in the calculation of the **CA** solution, only the **correspondence** that can be 'explained' by the matrix of explanatory variables is represented in the final results.

Redundancy Analysis (RDA)

It is a direct gradient analysis, wherein a matrix of explanatory variables intervenes in the calculation of the **PCA** solution, only the **variance** that can be 'explained' by the matrix of explanatory variables is represented in the final results.

CCA and RDA data



X: explanatory variables
Y: response variabels