Extension of Correspondence Analysis to multiway data-sets through HOSVD: a geometric framework

Martina Iannacito, Olivier Coulaud, Alain Franc

Extension of Correspondence Analysis to multiway data-sets through HOSVD: a geometric framework

Martina Iannacito*

Tensor Decompositions for Data Science - Part II of II

MDS22, San Diego 27-09-2022

Joint work with O. Coulaud* and A. Franc†
* Concace - joint team with Airbus and Cerfacs
† BioGeCo - Inrae, Pleiade - Inria

Table of contents

1

A case study

Malabar dataset

Figure: Aerial tour of the Arcachon basin, France

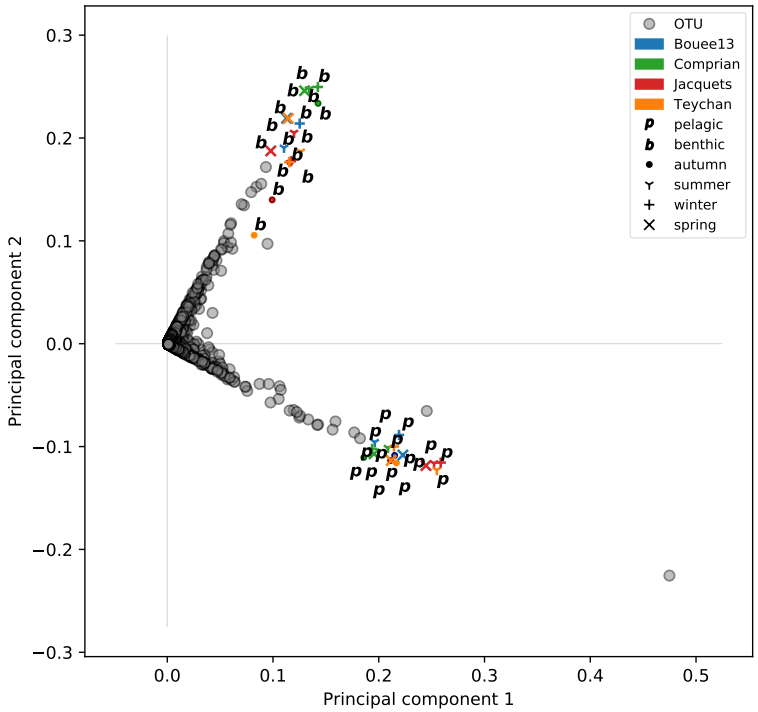
Malabar dataset

$d = 4$ mode dataset

- 1st mode of Operational Taxonomic Units (OTUs) with size $n_1 = 3539$
- 2nd mode of locations with size $n_2 = 4$, namely Bouee13, Comprian, Jacquets, Teychan
- 3rd mode of water column position with size $n_3 = 2$, that are pelagic and benthic
- 4th mode of seasons with size $n_4 = 4$

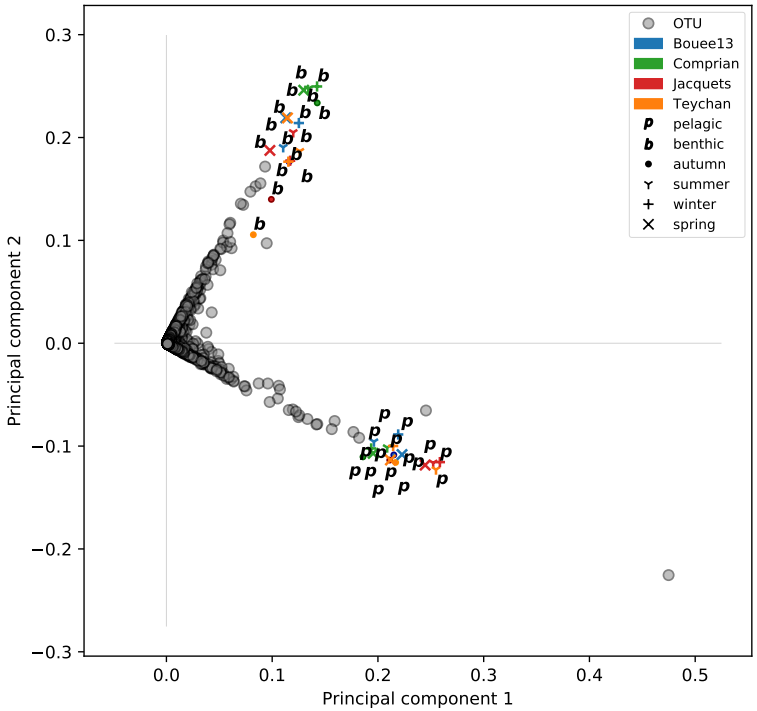
Figure: Aerial tour of the Arcachon basin, France

CA vs MWCA

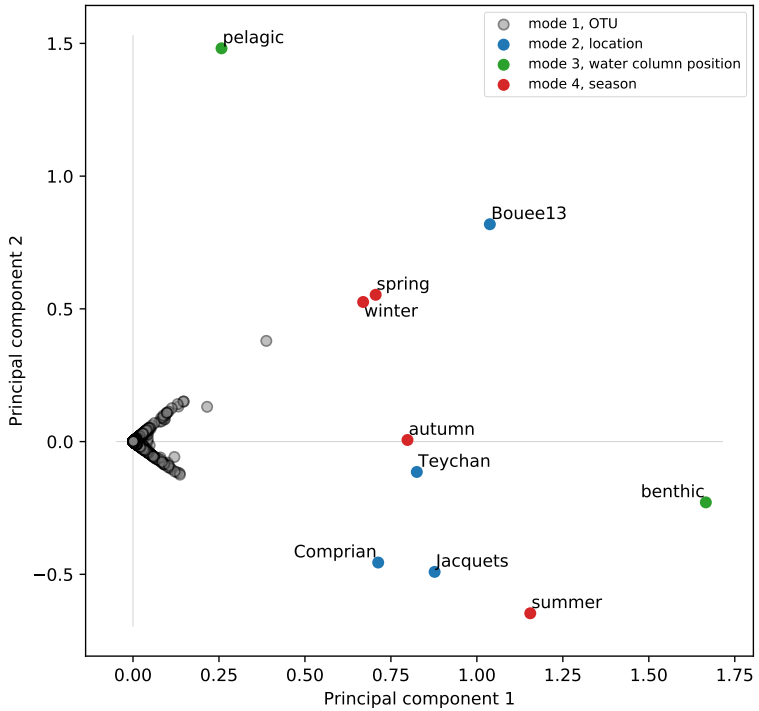


(A) Correspondence Analysis on mode 1

CA vs MWCA



(B) Correspondence Analysis on mode 1



(C) MultiWay Correspondence Analysis

Is it mathematically meaningful to disaply point clouds together?

2

Background on CA

Correspondence Analysis (CA) is a Principal Component Analysis (PCA) meant to investigate contingency tables through a **specific norm**

	Men				
Age group	**Very good**	**Good**	**Regular**	**Bad**	**Very bad**
16-24	145	402	84	5	3
25-34	112	414	74	13	2
35-44	80	331	82	24	4
45-54	54	231	102	22	6
55-64	30	219	119	53	12
65-74	18	125	110	35	4
+75	9	67	65	25	8

Table: Data from the Spanish National Health Survey of 1997 [Greenacre, 1984]

CA view point

Let A be a $m \times n$ real matrix

- columns (or rows) of A are realizations of n (m) random variables
- statistical: searching the r directions that maximize the variance

CA view point

Let A be a $m \times n$ real matrix

- columns (or rows) of A are realizations of n (m) random variables
- A has unknown rank

- statistical: searching the r directions that maximize the variance
- algebraic: searching the best rank r data approximation

CA view point

Let A be a $m \times n$ real matrix

- columns (or rows) of A are realizations of n (m) random variables
- A has unknown rank

- columns of A are coordinates of each column variable category in the space $\mathbb{R}^m$, rows of A coordinates of each row variable category in $\mathbb{R}^n$

- statistical: searching the r directions that maximize the variance
- algebraic: searching the best rank r data approximation
- **geometrical**: searching the subspace of dimension r that minimizes the columns or rows projection error

CA inner product

Let A be a contingency table with relative frequencies, the row and a_R and column marginals a_C are

$$a_R(i) = \sum_{j=1}^{n} A(i,j) \qquad \text{and} \qquad a_C(j) = \sum_{i=1}^{m} A(i,j).$$

- $\mathcal{S}'$ denotes the Euclidean space $\mathbb{R}^{m \times n}$ with the inner product induced by D_R^{-1} and D_C^{-1} such that

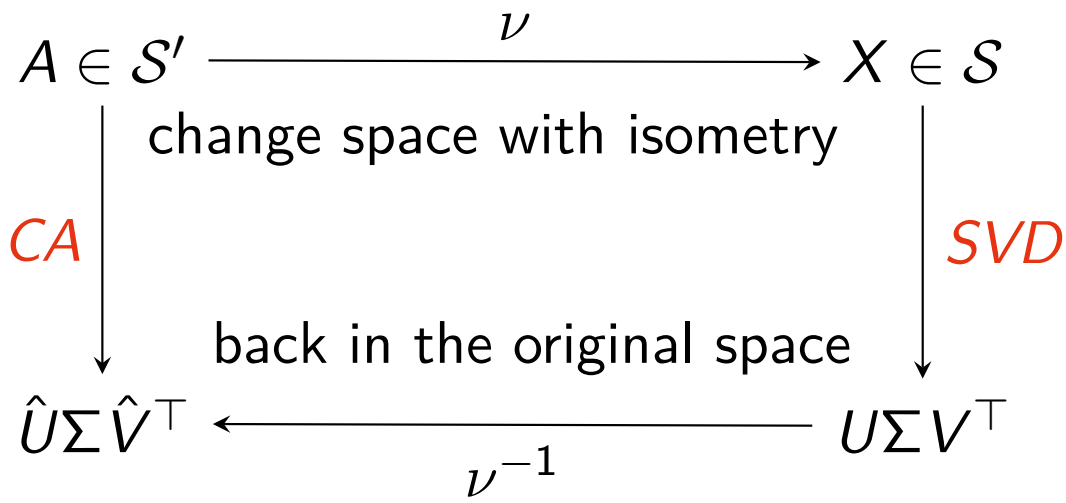
$$D_R = \text{diag}(\sqrt{a_R}) \qquad \text{and} \qquad D_C = \text{diag}(\sqrt{a_C}),$$

- $\mathcal{S}$ denotes the Euclidean space $\mathbb{R}^{m \times n}$ with the standard inner product

The two Euclidean spaces are isometric through ν defined as

$$\nu : \mathcal{S}' \to \mathcal{S} \qquad \text{such that} \qquad \nu(A) = D_R^{-1} A D_C^{-1}$$

CA scheme



Point cloud coordinates in $\mathcal{S}'$

$$W_R = \hat{U}\Sigma = D_R U\Sigma$$
$$W_C = \hat{V}\Sigma = D_C V\Sigma$$

Point cloud coordinates in $\mathcal{S}$

$$Y_R = U\Sigma$$
$$Y_C = V\Sigma$$

Barycentric relation [Lebart, 1982]

$$Z_R = D_R^{-2} A Z_C \Sigma^{-1} \quad \text{and} \quad Z_C = D_C^{-2} A^\top Z_R \Sigma^{-1}$$

with $Z_i = D_i^{-2} W_i$

Geometrical meaning of barycentric relation

$$Z_R(i,h) = \left(D_R^{-2} A Z_C \Sigma^{-1}\right)(i,h) = \frac{1}{\sigma_h} \sum_{j=1}^{n} \frac{A(i,j)}{a_R(i)} Z_C(j,h)$$

$$= \frac{1}{\sigma_h} \sum_{j=1}^{n} \rho_i(j) Z_C(j,h)$$

with σ_h the h-th singular value and $\rho_i \in \mathbb{R}^m$ such that

$$\sum_{j=1}^{n} \rho_i(j) = \sum_{j=1}^{n} \frac{A(i,j)}{a_R(i)} = \frac{a_R(i)}{a_R(i)} = 1$$

Geometrical meaning

The h-th Principal Coordinate of the i-th category of the row variable is the barycentre of the h-th Principal Coordinate of all the column variable categories

3

MultiWay Correspondence Analysis

How to study multiway contingency tables?

	Men					Women				
Age group	**Very good**	**Good**	**Regular**	**Bad**	**Very bad**	**Very good**	**Good**	**Regular**	**Bad**	**Very bad**
16-24	145	402	84	5	3	98	387	83	13	3
25-34	112	414	74	13	2	108	395	90	22	4
35-44	80	331	82	24	4	67	327	99	17	4
45-54	54	231	102	22	6	36	238	134	28	10
55-64	30	219	119	53	12	23	195	187	53	18
65-74	18	125	110	35	4	26	142	174	63	16
+75	9	67	65	25	8	11	69	92	41	9

Table: Data from the Spanish National Health Survey of 1997

How to study multiway contingency tables?

	Men					Women				
Age group	**Very good**	**Good**	**Regular**	**Bad**	**Very bad**	**Very good**	**Good**	**Regular**	**Bad**	**Very bad**
16-24	145	402	84	5	3	98	387	83	13	3
25-34	112	414	74	13	2	108	395	90	22	4
35-44	80	331	82	24	4	67	327	99	17	4
45-54	54	231	102	22	6	36	238	134	28	10
55-64	30	219	119	53	12	23	195	187	53	18
65-74	18	125	110	35	4	26	142	174	63	16
+75	9	67	65	25	8	11	69	92	41	9

Table: Data from the Spanish National Health Survey of 1997

Possible solutions

Multiple Correspondence Analysis

Inria

How to study multiway contingency tables?

	Men					Women				
Age group	**Very good**	**Good**	**Regular**	**Bad**	**Very bad**	**Very good**	**Good**	**Regular**	**Bad**	**Very bad**
16-24	145	402	84	5	3	98	387	83	13	3
25-34	112	414	74	13	2	108	395	90	22	4
35-44	80	331	82	24	4	67	327	99	17	4
45-54	54	231	102	22	6	36	238	134	28	10
55-64	30	219	119	53	12	23	195	187	53	18
65-74	18	125	110	35	4	26	142	174	63	16
+75	9	67	65	25	8	11	69	92	41	9

Table: Data from the Spanish National Health Survey of 1997

Possible solutions

Multiple Correspondence Analysis

MultiWay Correspondence Analysis (MWCA)

Inria

MWCA is a Principal Components Analysis of a multiway table with a **specific norm**

- statistical meaning [Kroonenberg, 2008]
- algebraic meaning [Kroonenberg, 1983]
- **geometric meaning** [Inria RR-9429, 2021]

Let $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be an order d tensor, then

Tucker format [Tucker 1963]

$$\mathbf{A}(i_1, \ldots, i_d) = \sum_{i_1, \ldots, i_d = 1}^{r_1, \ldots, r_d} \mathbf{C}(i_1, \ldots, i_d) u_{i_1}^{(1)} \otimes \cdots \otimes u_{i_d}^{(d)}$$

where $\mathbf{C}$ is the **core tensor**, $U_k = [u_1^{(k)}, \ldots, u_{r_k}^{(k)}]$ is an orthogonal $(n_k \times r_k)$ matrix with $r_k = \text{rank}(A^{(k)})$. More compactly

$$\mathbf{A} = (U_1, \ldots, U_d)\mathbf{C}$$

Tucker model

Let $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be an order d tensor, then

Tucker format [Tucker 1963]

$$\mathbf{A}(i_1, \ldots, i_d) = \sum_{i_1, \ldots, i_d = 1}^{r_1, \ldots, r_d} \mathbf{C}(i_1, \ldots, i_d) u_{i_1}^{(1)} \otimes \cdots \otimes u_{i_d}^{(d)}$$

where $\mathbf{C}$ is the **core tensor**, $U_k = [u_1^{(k)}, \ldots, u_{r_k}^{(k)}]$ is an orthogonal $(n_k \times r_k)$ matrix with $r_k = \text{rank}(A^{(k)})$. More compactly

$$\mathbf{A} = (U_1, \ldots, U_d)\mathbf{C}$$

Tucker decomposition is computed through the HOSVD, that performs a SVD for each matricization

$$\mathbf{A}^{(k)} = U_k \Sigma_k V_k^\top$$

and the core tensor is obtained projecting the tensor $\mathbf{A}$ in the new basis, i.e., $\mathbf{C} = (U_1^\top, \ldots, U_d^\top)\mathbf{A}$

MWCA inner product

Let $\mathbf{A}$ be a relative frequencies multiway contingency table, the k-th mode marginal is a_k such that

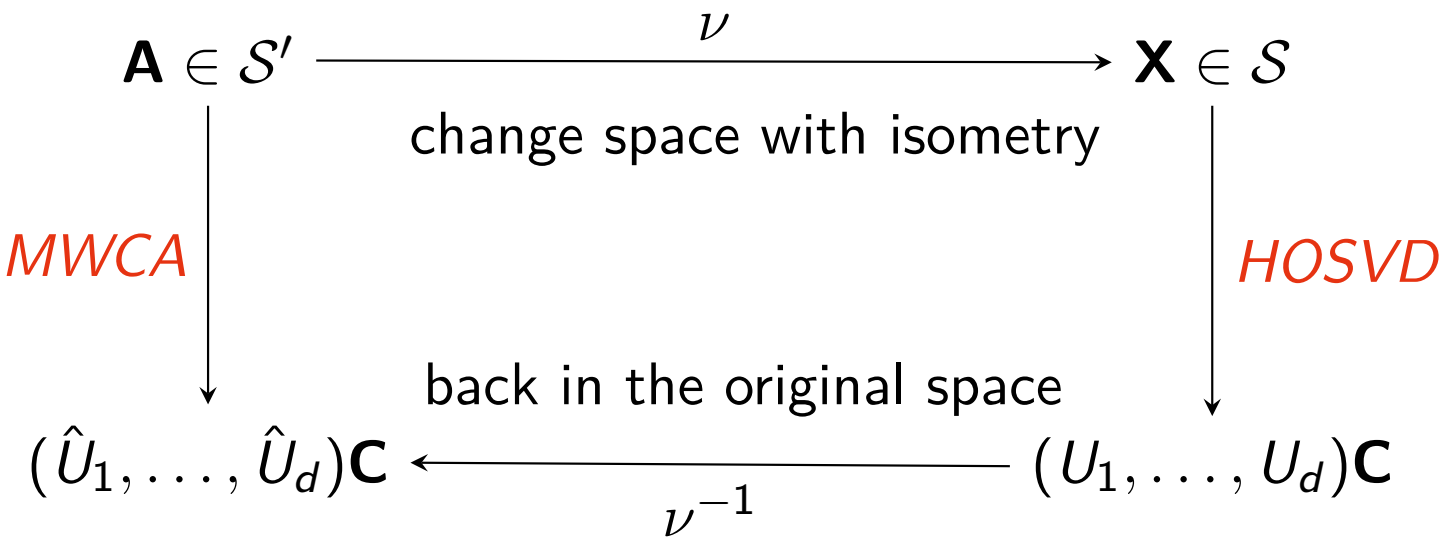
$$a_k(i_k) = \sum_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_d=1}^{n_1,\ldots,n_{k-1},n_{k+1},\ldots,n_d} \mathbf{A}(i_1,\ldots,i_d) \qquad \forall\, k \in \{1,\ldots,d\}$$

- $\mathcal{S}'$ denotes the Euclidean space $\mathbb{R}^{n_1\times\cdots\times n_d}$ with the inner product induced by $(D_1^{-1},\ldots,D_d^{-1})$, with $D_k = \mathrm{diag}(\sqrt{a_k})$
- $\mathcal{S}$ denotes the Euclidean space $\mathbb{R}^{n_1\times\cdots\times n_d}$ with the standard inner product.

The two Euclidean spaces are isometric through ν defined as

$$\nu : \mathcal{S}' \to \mathcal{S} \qquad \text{such that} \qquad \nu(\mathbf{A}) = (D_1^{-1},\ldots,D_d^{-1})\mathbf{A}$$

MWCA scheme



Point cloud coordinates in $\mathcal{S}'$

$$W_k = \hat{U}_k \Sigma_k = D_k U_k \Sigma_k$$

Point cloud coordinates in $\mathcal{S}$

$$Y_k = U_k \Sigma_k$$

where Σ_k are the singular values of $\mathbf{X}^{(k)}$

If $Z_k = D_k^{-2} W_k$, then

Barycentric relation

$$Z_k = D_k^{-2} \mathbf{A}^{(k)} (Z_d \otimes_{\mathrm{K}} \cdots \otimes_{\mathrm{K}} Z_{k+1} \otimes_{\mathrm{K}} Z_{k-1} \otimes_{\mathrm{K}} \cdots \otimes_{\mathrm{K}} Z_1)(\Sigma_k^{-1} \mathbf{B}_k^{(k)})^\top$$

where

$$\mathbf{B}_k = (\Sigma_1^{-1}, \ldots, \Sigma_{k-1}^{-1}, \mathbb{I}_k, \Sigma_{k+1}^{-1}, \ldots, \Sigma_d^{-1})\mathbf{C}$$

with $\mathbf{C}$ the core tensor of $\mathbf{X} = \nu(\mathbf{A})$

from the more general result of [Proposition 3.7; Kolda, 2006] if $\mathbf{W} = (M_1, \ldots, M_d)\mathbf{Y}$, then

$$\mathbf{W}^{(k)} = M_k \mathbf{Y}^{(k)} (M_d \otimes_{\mathrm{K}} \cdots \otimes_{\mathrm{K}} M_{k+1} \otimes_{\mathrm{K}} M_{k-1} \otimes_{\mathrm{K}} \cdots M_1)$$

MWCA barycentric relation II - meaning

$$Z_k(i_k, h_k) = \frac{1}{\sigma_{k,h_k}} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{d} \sum_{i_\ell, h_\ell = 1}^{n_\ell, r_\ell} \boldsymbol{\rho}_{i_k}(i_1, \ldots, i_{k-1}, i_k, \ldots, i_d)) Z_1(i_1, h_1) \cdots Z_{k-1}(i_{k-1}, h_{k-1})$$

$$Z_{k+1}(i_{k+1}, h_{k+1}) \cdots Z_d(i_d, h_d) \mathbf{B}_k(h_1, \ldots h_d)$$

with σ_{k,h_k} the h_k-th singular value of mode k and ρ_{i_k} such that

$$\sum_{\substack{\ell=1 \\ \ell \neq k}}^{d} \sum_{i_\ell=1}^{n_\ell} \boldsymbol{\rho}_{i_k}(i_1, \ldots, i_{k-1}, i_k, \ldots, i_d) = \sum_{\substack{\ell=1 \\ \ell \neq k}}^{d} \sum_{i_\ell=1}^{n_\ell} \frac{\mathbf{A}(i_1, \ldots, i_d)}{a_k(i_k)} = 1$$

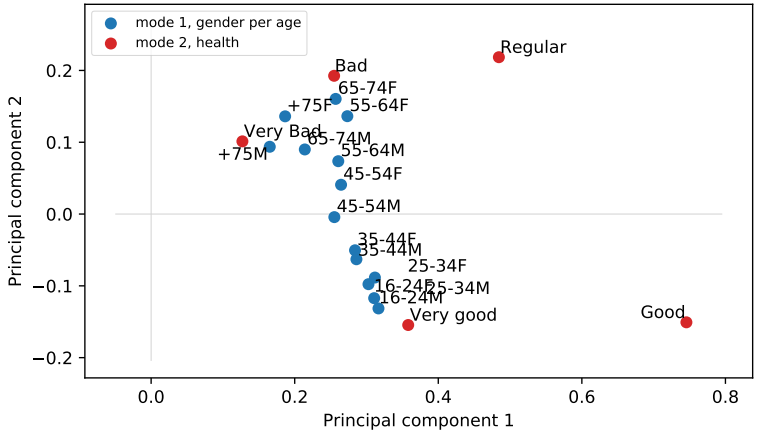
MWCA barycentric relation II - meaning

$$Z_k(i_k, h_k) = \frac{1}{\sigma_{k,h_k}} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{d} \sum_{i_\ell, h_\ell = 1}^{n_\ell, r_\ell} \rho_{i_k}(i_1, \ldots, i_{k-1}, i_k, \ldots, i_d)) Z_1(i_1, h_1) \cdots Z_{k-1}(i_{k-1}, h_{k-1})$$
$$Z_{k+1}(i_{k+1}, h_{k+1}) \cdots Z_d(i_d, h_d) \mathbf{B}_k(h_1, \ldots h_d)$$

Geometrical meaning

The h_k-th Principal Coordinate of the i_k-th category of the k-th variable is the barycentre of a linear combination of the h-th Principal Coordinate of all the other $(d-1)$ variable categories and coefficients expressed by $\mathbf{B}_k$ entries

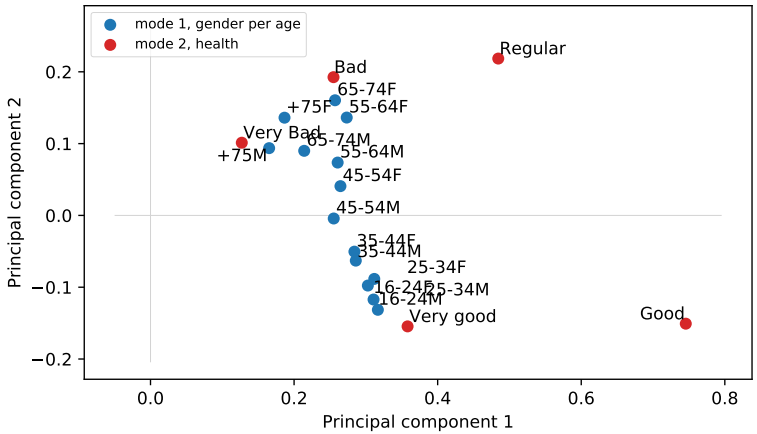
CA vs MWCA: example



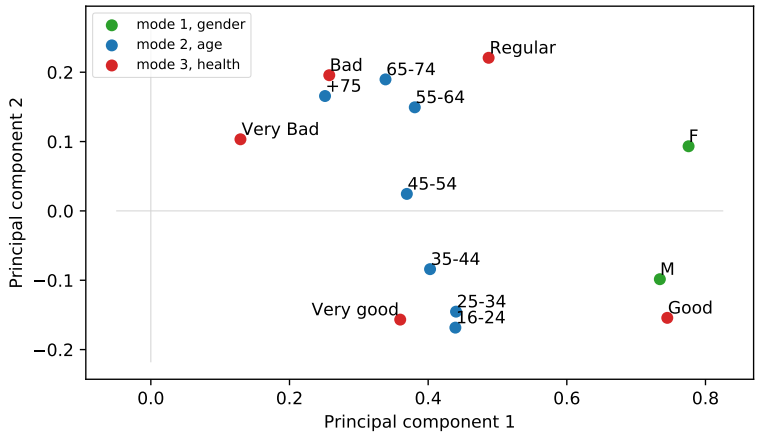
(D) CA on mode 1

- Age gradient from bottom to top
- Negative second component for good health, positive for bad health evaluation

CA vs MWCA: example



(E) CA on mode 1



(F) MWCA

- Age gradient from bottom to top
- Negative second component for good health, positive for bad health evaluation

- Age gradient from bottom to top
- Negative second component for good health, positive for bad health evaluation
- sex division

4

Malabar dataset analysis

Inria

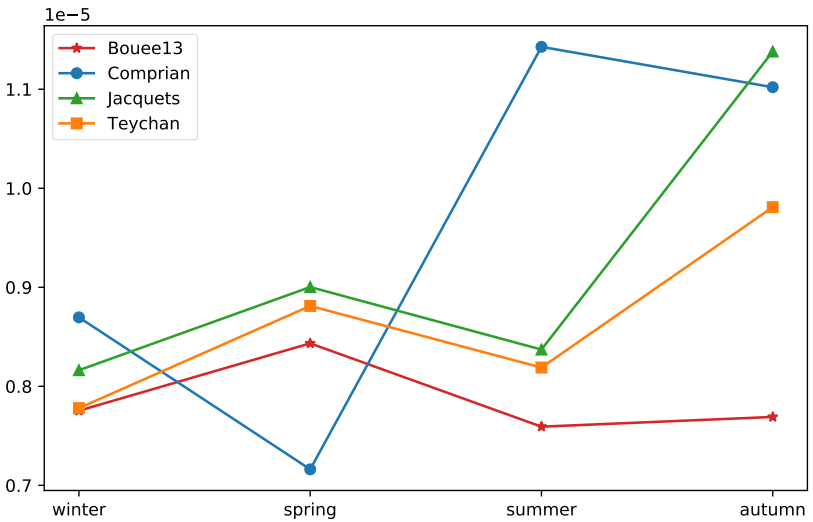
Malabar dataset

$d = 4$ mode dataset

- 1st mode of Operational Taxonomic Units (OTUs) with size $n_1 = 3539$
- 2nd mode of locations with size $n_2 = 4$, namely Bouee13, Comprian, Jacquets, Teychan
- 3rd mode of water column position with size $n_3 = 2$, that are pelagic and benthic
- 4th mode of seasons with size $n_4 = 4$

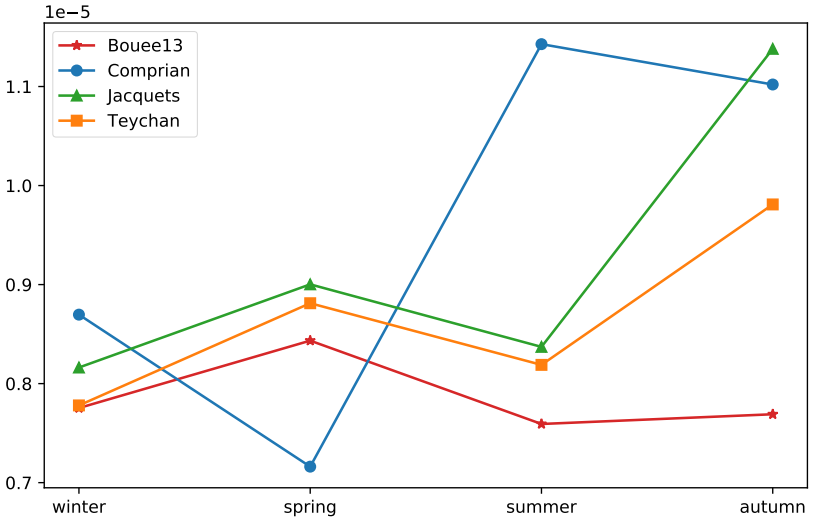
Figure: Aerial tour of the Arcachon basin, France

Preliminary dataset analysis

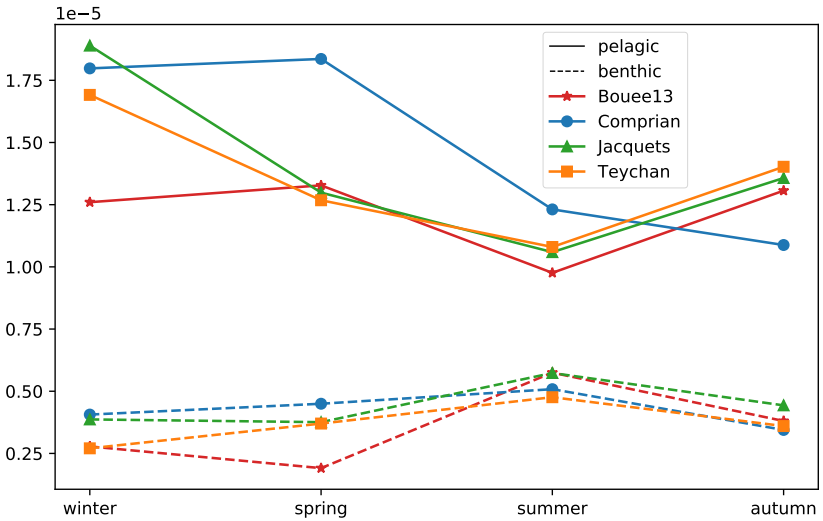


average number of OTU per location in function of the season over the entire water column

Preliminary dataset analysis

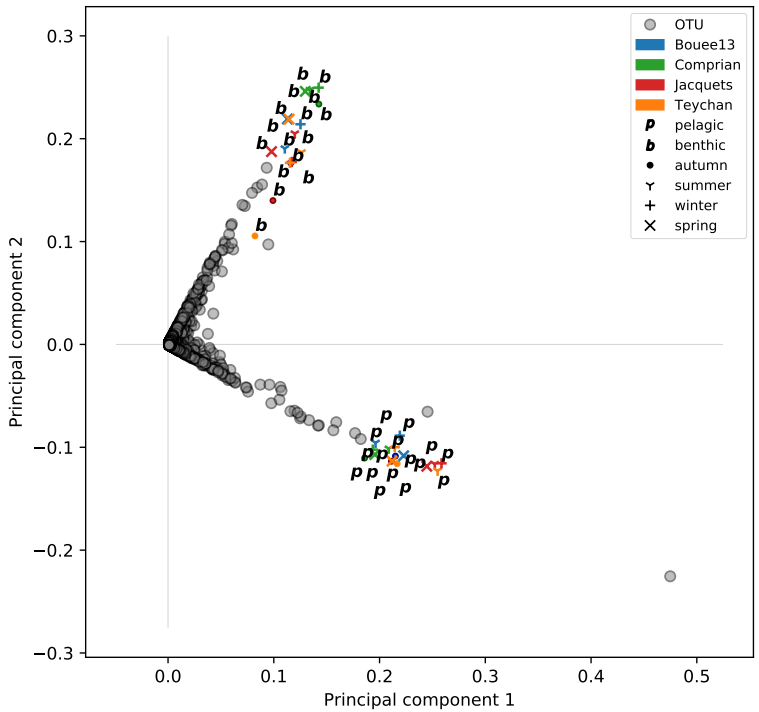


average number of OTU per location in function of the season over the entire water column



average number of OTU per location and position in the water column in function of the season

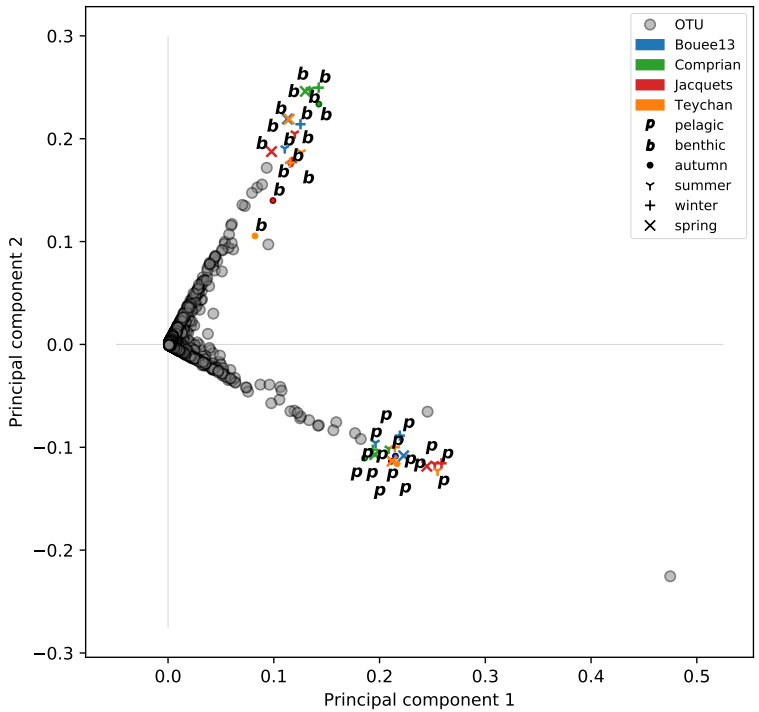
MWCA on Malabar dataset



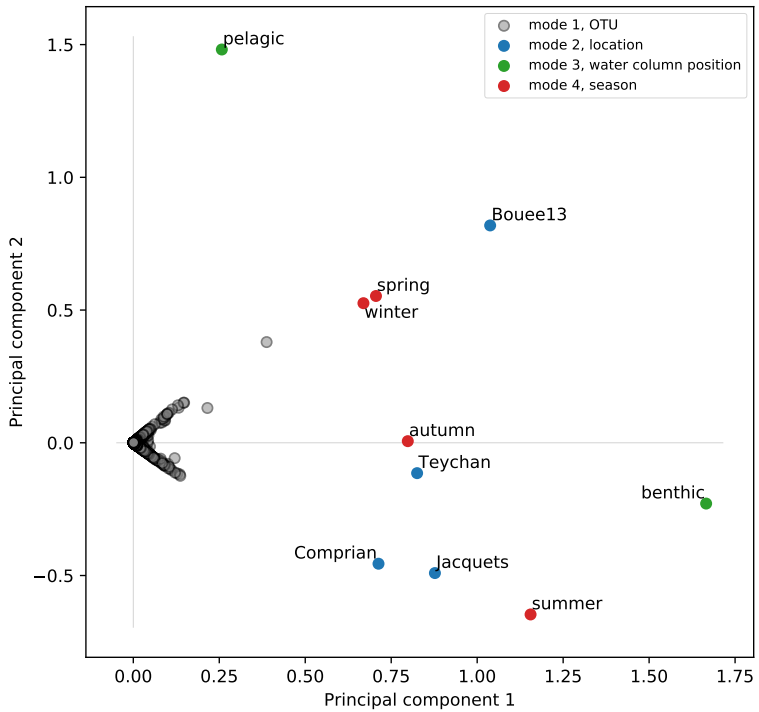
(D) CA on mode 1

- Orthogonality among water column positions
- Distribution of OTU along their directions

MWCA on Malabar dataset



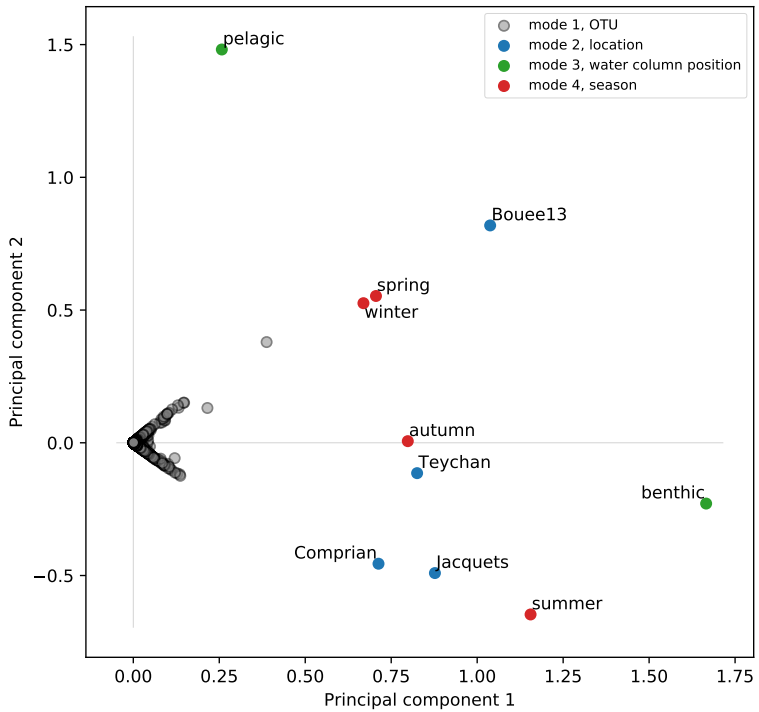
(E) CA on mode 1



(F) MWCA

MWCA on Malabar dataset

- water column positions still orthogonal, less determinant
- OTU distribution affected by seasons and locations too
- locations and seasons are clustered
- correlation between seasons and locations



(G) MWCA

Take home message

- MultiWay Correspondence Analysis has a triple nature as PCA and CA:
  - statistical meaning [Kroonenberg, 2008]
  - algebraic meaning [Kroonenberg, 1983]
  - **geometric meaning** [Inria RR-9429, 2021]
- the barycentric relation characterizing CA holds also for MWCA
- MWCA highlights inter-variable interactions
- CA appears more suitable for studying combinations of categories from different variables

Thanks for the attention.
[Inria RR-9429, 2021]

Questions?