# An introduction to $K$-tables Analyses

AB Dufour, ade4 team

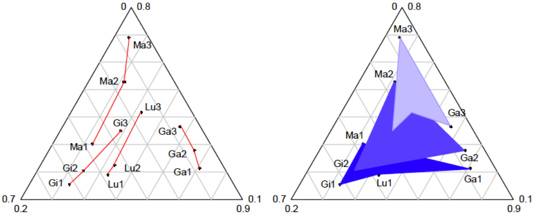
UCB Lyon 1

January 2015

# Example 1:

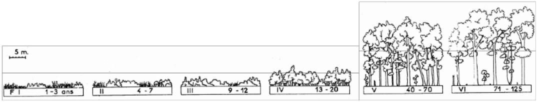
Proportion of people employed in the primary, secondary and tertiary sectors for 4 town councils in the South of France: Gignac (Gi), Ganges (Ga), Matellas (Ma) and Lunel (Lu) during 3 census (1968, 1975, 1982).
What is the objective of the study ?

- Studying the evolution of each town council ?
- Studying the evolution of economic typologies ?



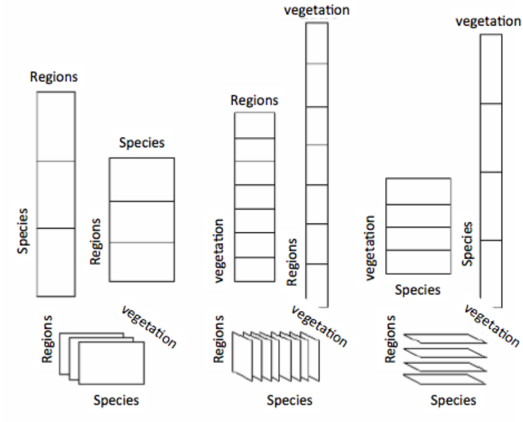
# Example 2: Influence of vegetation successions on bird communities composition

The dataset contains the number of birds of 79 species observed in four regions (Burgundy, Provence, Corsica and Poland) along a gradient of six stages of vegetation succession:




```
 data(bf88)
 names(bf88)
[1] "S1" "S2" "S3" "S4" "S5" "S6"
 dim(bf88$S1)
[1] 79  4
 head(bf88$S1)
     Pol Bur Pro Cor
MOAL   7   1   0   0
LULU   5   0   0   0
LOFL   0   0   0   0
LACO   0   3   0   1
HIIC   0   0   0   0
EMCT  21  30   0   0
```

Data can be organized in one table. But which one ? and for what study ?

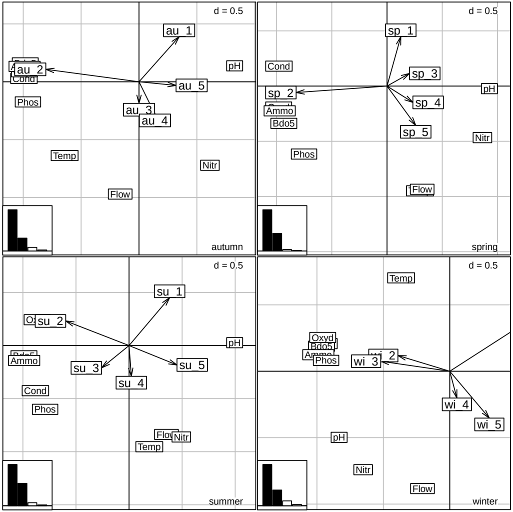


# Example 3: The Meaudret, environment and seasons

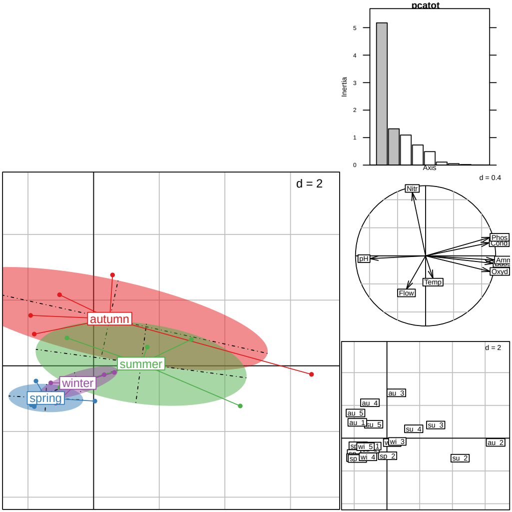
4 seasons

4 PCA

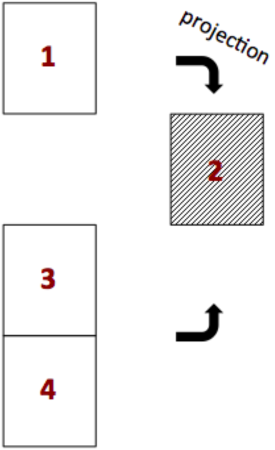
How to compare the factorial maps ?

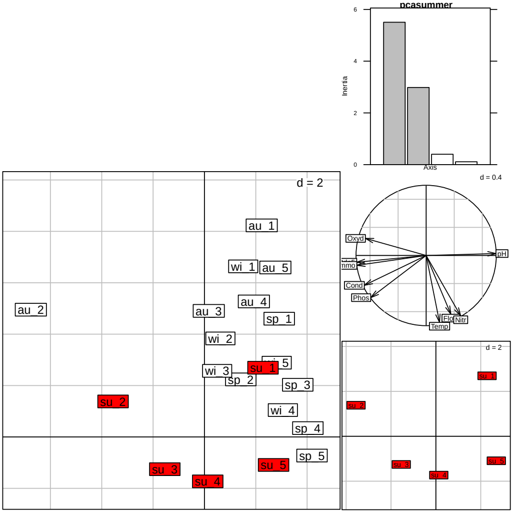


# First Analysis: concataining all datasets

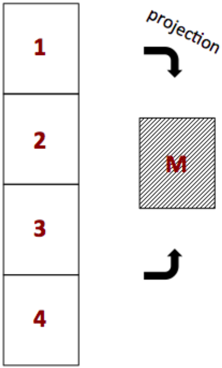



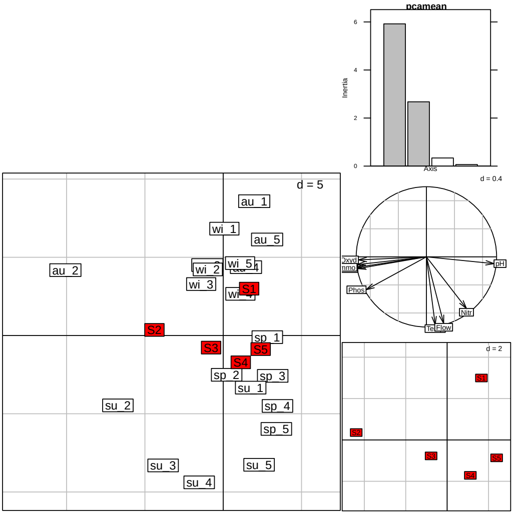
# Second Analysis: one season as reference



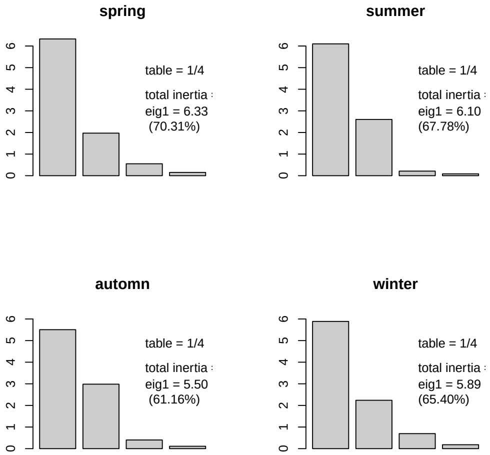



# Third Analysis: season average table as reference



# Is this mean table a good reference?



# Conclusion

Objective: link the environmental structure and seasons

- Analysing four tables separatly: no common space
- Creating a common space: a reference table
  - one season as reference
  - the average table as reference

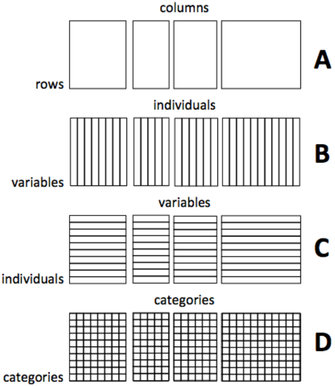
We link the individuals in a same variable space.
BUT
the question was about the relationships between season tables.

# The $K$-table methods

| Function name | Analysis name |
|---|---|
| sepan | K-table separate analyses |
| pta | Partial triadic analysis |
| foucart | Foucart analysis |
| statis | STATIS analysis |
| mfa | Multiple factor analysis |
| mcoa | Multiple coinertia analysis |
| statico | 2 K-table analysis |
| co-statis | 2 K-table analysis |

# Preparing a $K$-table analysis

$K$ tables are stored in objects of type `ktab`. A `ktab` is a list of dataframes that share the same row names.



# Preparing a $K$-table analysis

The tables must share the same row names and row weights. If the common dimension of the tables is the columns, they must be transposed to have their common dimension as rows.

| Element of `ktab` | Definition |
|---|---|
| `lw` | row weights, common to all the tables |
| `cw` | column weights |
| `blo` | number of columns of each table |
| `TL` | index for rows (table and row names) |
| `TC` | index for columns (table and columns names) |
| `T4` | index for 4 elements of an array |
| `call` | function call |

# Creating a $K$-table object

Series of tables are stored in object of class `ktab` which can be created using:

- `ktab.list.df`: a list of data frames
- `ktab.list.dudi`: a list of `dudi` objects
- `ktab.data.frame`: one dataframe (and number of columns of each table)
- `ktab.within`: an object created by a `withinpca` analysis

# Examples

A way for creating a `ktab` object:

```
data(meaudret)
pcadat <- withinpca(meaudret$env,meaudret$design$season,scaling="partial",scann=F)
ktadat <- ktab.within(pcadat, colnames = rep(c("S1", "S2", "S3", "S4", "S5"),4))
```

Another way for creating a `ktab` object:

```
prep <- split(meaudret$env, meaudret$design$season)
prep <- lapply(prep, function(x) data.frame(t(scalewt(x))))
ktaseason <- ktab.list.df(prep, colnames = rep(c("S1", "S2", "S3", "S4", "S5"),4))
```

# The $K$-table objects

```
names(ktaseason)
```

```
 [1] "autumn" "spring" "summer" "winter" "blo"    "lw"     "cw"     "TL"     "TC"
[10] "T4"     "call"
```

```
ktaseason$spring
```

```
             S1         S2          S3         S4         S5
Temp -1.3728129 -0.3922323 -0.39223227  0.5883484  1.5689291
Flow -1.6151562 -0.4251482 -0.01830782  0.8157149  1.2428973
pH    0.2041241 -1.8371173  0.20412415  1.2247449  0.2041241
Cond  0.0000000  1.9069252 -0.47673129 -0.4767313 -0.9534626
Bdo5 -0.9621078  1.9351487 -0.41545565 -0.3061252 -0.2514600
Oxyd -0.6364382  1.9923283 -0.49808208 -0.4980821 -0.3597259
Ammo -0.7376580  1.9837018 -0.45854414 -0.4087024 -0.3787973
Nitr -0.5281643 -1.6155614  0.40389035  0.4038903  1.3359450
Phos -1.0473231  1.6189391 -1.06650488  0.1419449  0.3529441
```

```
ktaseason$blo
```

```
autumn spring summer winter
     5      5      5      5
```

```
ktaseason$lw
```

```
[1] 1 1 1 1 1 1 1 1 1
```

```
ktaseason$cw[1:15]
```

```
 [1] 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2
```

# The $K$-table objects

```
ktaseason$TL[1:18,]

        T    L
1  autumn Temp
2  autumn Flow
3  autumn   pH
4  autumn Cond
5  autumn Bdo5
6  autumn Oxyd
7  autumn Ammo
8  autumn Nitr
9  autumn Phos
10 spring Temp
11 spring Flow
12 spring   pH
13 spring Cond
14 spring Bdo5
15 spring Oxyd
16 spring Ammo
17 spring Nitr
18 spring Phos

       ⋮
```

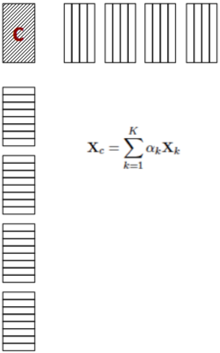
```
ktaseason$TC

        T  C
1  autumn S1
2  autumn S2
3  autumn S3
4  autumn S4
5  autumn S5
6  spring S1
7  spring S2
8  spring S3
9  spring S4
10 spring S5
11 summer S1
12 summer S2
13 summer S3
14 summer S4
15 summer S5
16 winter S1
17 winter S2
18 winter S3
19 winter S4
20 winter S5
```

```
ktaseason$T4

        T 4
1  autumn 1
2  autumn 2
3  autumn 3
4  autumn 4
5  spring 1
6  spring 2
7  spring 3
8  spring 4
9  summer 1
10 summer 2
11 summer 3
12 summer 4
13 winter 1
14 winter 2
15 winter 3
16 winter 4
```

# The $K$ tables




Let $\mathbf{X}_1, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_K$ be $K$ tables of quantitative variables with the same $n$ rows (samples) and the same $p$ columns (variables).

Let
$(\mathbf{X}_1, \mathbf{Q}, \mathbf{D}), \ldots, (\mathbf{X}_k, \mathbf{Q}, \mathbf{D}), \ldots, (\mathbf{X}_K, \mathbf{Q}, \mathbf{D})$
be the $K$ associated statistical triplets.

The partial triadic analysis is decomposed in three steps.

# The RV-coefficient

Let $\mathbf{X}_k$ and $\mathbf{X}_\ell$ be two tables of quantitative variables with the same $n$ rows (samples) and the same $p$ columns (variables).
Let ($\mathbf{X}_k$, $\mathbf{Q}$, $\mathbf{D}$) and ($\mathbf{X}_\ell$, $\mathbf{Q}$, $\mathbf{D}$) the two statistical triplets. The inner-product between the tables are defined by:

$$Covv(\mathbf{X}_k, \mathbf{X}_\ell) = Trace({\mathbf{X}_k}^\top \mathbf{D} \mathbf{X}_\ell \mathbf{Q}) = Trace({\mathbf{X}_\ell}^\top \mathbf{D} \mathbf{X}_k \mathbf{Q})$$

Let note that $Covv(\mathbf{X}_k, \mathbf{X}_k) = Trace({\mathbf{X}_k}^\top \mathbf{D} \mathbf{X}_k \mathbf{Q}) = Trace({\mathbf{X}_k}^\top \mathbf{D} \mathbf{X}_k \mathbf{Q})$ is called the vectorial variance $Vav(\mathbf{X}_k)$.

The Vectorial correlation coefficient called RV-coefficient is :

$$RV(\mathbf{X}_k, \mathbf{X}_\ell) = \frac{Covv(\mathbf{X}_k, \mathbf{X}_\ell)}{\sqrt{Vav(\mathbf{X}_k)}\sqrt{Vav(\mathbf{X}_\ell)}}$$

# Step 1: the interstructure

For each couple of triplets ($\mathbf{X}_k$, $\mathbf{Q}$, $\mathbf{D}$) and ($\mathbf{X}_\ell$, $\mathbf{Q}$, $\mathbf{D}$), we can compute the RV coefficient and put all them in the $\mathbf{RV}$ matrix :

$$\begin{pmatrix} RV(\mathbf{X}_1,\mathbf{X}_1) & \dots & RV(\mathbf{X}_1,\mathbf{X}_k) & \dots & RV(\mathbf{X}_1,\mathbf{X}_K) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ RV(\mathbf{X}_k,\mathbf{X}_1) & \dots & RV(\mathbf{X}_k,\mathbf{X}_k) & \dots & RV(\mathbf{X}_k,\mathbf{X}_K) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ RV(\mathbf{X}_K,\mathbf{X}_1) & \dots & RV(\mathbf{X}_K,\mathbf{X}_k) & \dots & RV(\mathbf{X}_K1,\mathbf{X}_K) \end{pmatrix}$$

We look for the eigenvalues $\Lambda_{is}$ and eigenvectors $\mathbf{U}_{is}$ of $\mathbf{RV}$ giving the scores of tables $\mathbf{S} = \mathbf{U}_{is}\Lambda_{is}^{1/2}$ which can be viewed through a correlation circle.

# Step 2: the compromise

Let $\mathbf{u}_{is}$ the first eigenvector of the interstructure analysis:
$\mathbf{u}_{is}{}^{\top} = (\alpha_1 \ \dots \ \alpha_k \ \dots \ \alpha_K)$.
For all $k = 1, K$, $\alpha_k$ represents the weighting of the $\mathbf{X}_k$ table. And we can therefore built the compromise table :

$$\mathbf{X}_c = \sum_{k=1}^{K} \alpha_k \mathbf{X}_k$$

The analysis of the compromise is the analysis of the triplet $(\mathbf{X}_c, \mathbf{Q}, \mathbf{D})$ in the PCA sense under the following constraint $\sum_{k=1}^{K} \alpha_k^2 = 1$

# Step 3: the intrastructure

Let $\Lambda$ and $\mathbf{U}$, the eigenvalues and the eigenvectors of $(\mathbf{X}_c, \mathbf{Q}, \mathbf{D})$.

Projection
of the rows
of each table $\mathbf{X}_k$
onto the principal axes:
$\mathbf{R}_k = \mathbf{X}_k\mathbf{Q}\mathbf{U}$

Projection
of the columns
of each table $\mathbf{X}_k$
onto the principal components:
$\mathbf{C}_k = \mathbf{X}^\top\mathbf{D}\mathbf{X}_k\mathbf{Q}\mathbf{U}\Lambda^{-1/2}$