M2 Écologie,Évolution,Biométrie 
UE Description Statistique des Structures Biologiques

Co-inertie, co-structures et compromis

D. Chessel & A.-B. Dufour

Notes de cours cssb8

Table des matières

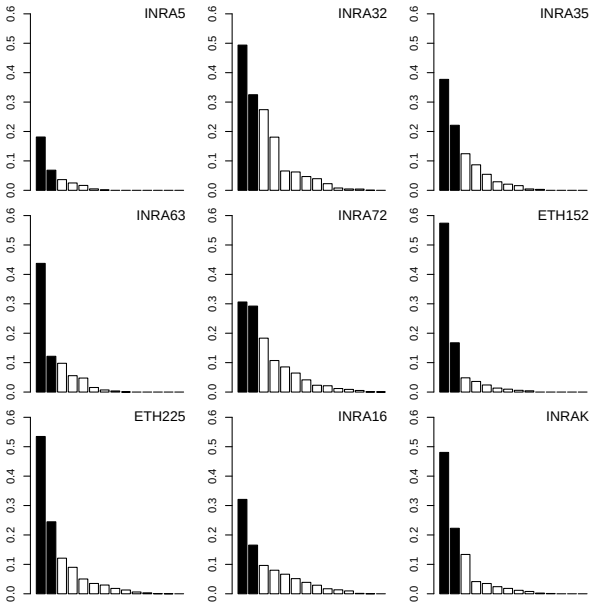
1 Introduction

Nous avons abordé la question des K-tableaux par les analyses élémentaires des cubes. La demande la plus abondante provient des tableaux multiples associés par les lignes ou les colonnes. Un cube donne plusieurs objets de ce type. Nous abordons ici les problèmes de co-structures à deux ou plusieurs tableaux. On suppose que chaque tableau donne, à lui seul, une typologie de lignes et de colonnes : on veut discuter de relations entre tableaux au niveau des relations entre structures. De même qu'en analyse multivariée, on examine avec soin le comportement des variables, de même en analyse multi-tableaux on examine avec soin le comportement de chacune des analyses séparées. La fonction `sepan` le fait simplement. L'ensemble de données `microsatt` est décrit dans :

`http://pbil.univ-lyon1.fr/R/pps/pps055.pdf`

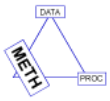
Un tableau important, proposé par D. Laloë, regroupe les fréquences alléliques pour 18 races bovines (taurines ou zébu), d'origine française ou africaine, typées sur 9 microsatellites.


```
library(ade4)
data(microsatt)
fac <- factor(rep(microsatt$loci.names, microsatt$loci.eff))
w <- dudi.coa(data.frame(t(microsatt$tab)), scann = FALSE)
wit <- within(w, fac, scann = FALSE)
microsatt.ktab <- ktab.within(wit)
plot(sepan(microsatt.ktab))
```



La figure indique clairement le champ dans lequel on se trouve. Un marqueur microsatellite fait une typologie des 18 races, avec plus ou moins d'intensité (inertie) et de simplicité. Comment mesure-t-on et comment compare-t-on des analyses de tableaux ? C'est l'objet de ce cours, qu'on essaiera le moins mathématisé possible.

2 Eléments de base



On commence par deux tableaux, situation déjà rencontrée dans l'ordination directe et très répandue. Il convient maintenant de bien distinguer deux choses.

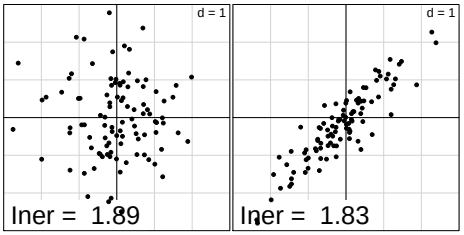
2.1 Inertie et Variance vectorielle

L'inertie est un concept simple qui généralise la notion de variance. Soit un tableau $\mathbf{X}$ avec n lignes et p colonnes. La ligne de rang i de $\mathbf{X}$ notée $\mathbf{X}_i$ est un point de $\mathbb{R}^p$. La somme des produits du carré de la distance à l'origine par le poids du point est l'inertie. Ceci n'a de sens que si les poids sont définis et si on sait calculer un carré de distance. Il n'y a pas d'analyse de données sans triplet statistique $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. Pour simplifier l'accès à des notions essentielles, on s'en tiendra à une pondération uniforme des individus et une pondération unitaire des variables, soit $\mathbf{Q} = \mathbf{I}_p$ et $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$. C'est le cas de l'ACP. Mesurer la variabilité se fait par :

$$iner(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} d^2(\mathbf{X}_i) = \sum_{j=1}^{p} \mathrm{v}\left(\mathbf{X}^j\right) = \sum_{k=1}^{r} \lambda_k$$

L'inertie mesure la grosseur du nuage, pas sa forme. L'intensité d'une structure intègre l'inertie qui en fait partie mais y ajoute autre chose qui exprime les liens entre les éléments. Noter aussi, dans la formule qui précède la vue dans $\mathbb{R}^p$ (les carrés des distances à l'origine), dans $\mathbb{R}^n$ (les variances des variables) et dans les deux (les valeurs propres de l'analyse). Deux tableaux peuvent avoir la même inertie sans avoir la même forme :

```
set.seed(20092006)
library(MASS)
fun1 <- function(x) {
    res <- paste("Iner = ", round(sum(diag(t(x) %*% x))/100, dig = 2))
    s.label(x, clab = 0, xlim = c(-3, 3), ylim = c(-3, 3), sub = res,
        csub = 2)
}
x1 <- mvrnorm(100, c(0, 0), matrix(c(1, 0, 0, 1), 2))
x2 <- mvrnorm(100, c(0, 0), matrix(c(1, 0.9, 0.9, 1), 2))
par(mfrow = c(1, 2))
par(mar = rep(0, 4))
fun1(x1)
fun1(x2)
```



Deux nuages montrent des variabilités voisines. Celui de droite possède en plus de la covariance. On ne peut cependant pas ajouter directement une covariance à une variance, car la somme d'une covariance positive et d'une covariance négative ne fait pas une structure nulle. On peut ajouter des carrés de covariance

qui seront ajoutés à des carrés de variance pour mesurer la structure :

$$vav(\mathbf{X}) = \sum_{j=1}^{p}\sum_{k=1}^{p} \mathrm{cov}^2\left(\mathbf{X}_j, \mathbf{X}_k\right) = trace(\frac{1}{n}\mathbf{X}^T\mathbf{X}\frac{1}{n}\mathbf{X}^T\mathbf{X}) = trace(\mathbf{V}\mathbf{V})$$

De l'autre côté (dans $\mathbb{R}^p$), l'inertie fait intervenir les carrés des distances à l'origine (au lieu de la variance). Qu'est ce qui généralise un carré de distance pour deux points ? Le produit scalaire évidemment et on a la même quantité sour la forme :

$$vav(\mathbf{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{m=1}^{n} \left\langle \mathbf{X}^i | \mathbf{X}^m \right\rangle^2 = trace(\frac{1}{n}\mathbf{X}\mathbf{X}^T\frac{1}{n}\mathbf{X}\mathbf{X}^T) = trace(\frac{1}{n}\mathbf{W}\frac{1}{n}\mathbf{W})$$

Dans un cas, la matrice $\mathbf{V}$ est la matrice des covariances. Dans l'autre, $\mathbf{W}$ est la matrice des produits scalaires. La première est $p \times p$, la seconde est $n \times n$, les deux sont carrées et symétriques, les deux sont appelées opérateurs d'Escoufier. Quand on ne peut plus comparer deux tableaux sur les mêmes variables, on pourra comparer les $\mathbf{V}$; quand on ne pourra plus comparer deux tableaux sur les mêmes individus on pourra comparer les $\mathbf{W}$. Dans tous les cas :

$$vav(\mathbf{X}) = \sum_{k=1}^{r} \lambda_k^2$$

La variance vectorielle d'un tableau est donc la somme des carrés de covariance étendue à tous les couples de deux variables. Cette variance vectorielle se décompose en somme de carrés de valeurs propres. On pourrait tracer le graphe des carrés de valeurs propres pour décider du nombre d'axes. Il peut être beaucoup plus clair que celui des valeurs propres. Noter aussi que remplacer les données par les coordonnées, c'est conserver la variance vectorielle en ne l'exprimant que comme une somme de carrés de variance (toutes les covariances sont nulles), c'est concentrer l'organisation dans de la variance.

2.2 Variance et covariance vectorielles

On pourrait se demander à quoi ça sert. Tout simplement à passer de l'analyse d'un tableau à celle de plusieurs. Comment caractériser la co-inertie, c'est-à-dire la covariance vectorielle ? Introduisons un second tableau $\mathbf{Y}$ avec n lignes et q colonnes. Ses variables sont toujours dans $\mathbb{R}^n$ mais ses lignes sont dans un autre espace $\mathbb{R}^q$. Pour étendre la variance vectorielle à deux tableaux (la covariance d'un tableau avec lui-même doit être la variance vectorielle) il suffit de :

$$cov(\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^{p}\sum_{k=1}^{q} \mathrm{cov}^2\left(\mathbf{X}_j, \mathbf{Y}_k\right) = trace(\frac{1}{n}\mathbf{X}^T\mathbf{Y}\frac{1}{n}\mathbf{Y}^T\mathbf{X}) = trace(\mathbf{C}_{\mathbf{XY}}\mathbf{C}_{\mathbf{YX}})$$

On se retrouve dans une situation bien connue : celle de l'analyse d'un nouveau tableau, la matrice des covariances entre les deux paquets de variables. L'analyse de co-inertie va concentrer la co-structure en quelques couples de coordonnées comme l'analyse d'inertie concentre la structure dans la variance de quelques coordonnées. La covariance vectorielle se comprend aussi dans la géométrie des deux nuages de n points par l'extension :

$$vav(\mathbf{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{m=1}^{n} \left\langle \mathbf{X}^i | \mathbf{X}^m \right\rangle \left\langle \mathbf{Y}^i | \mathbf{Y}^m \right\rangle = trace(\frac{1}{n}\mathbf{X}\mathbf{X}^T\frac{1}{n}\mathbf{Y}\mathbf{Y}^T) = trace(\frac{1}{n}\mathbf{W}_{\mathbf{X}}\frac{1}{n}\mathbf{W}_{\mathbf{Y}})$$

2.3 Corrélation vectorielle

La cohérence de ces définitions vient du fait qu'on est passé sans le savoir de la géométrie des vecteurs à celle des matrices. La quantité :

$$0 \leq RV(\mathbf{X},\mathbf{Y}) = \frac{cov(\mathbf{X},\mathbf{Y})}{\sqrt{vav(\mathbf{X})}\sqrt{vav(\mathbf{Y})}} \leq 1$$

a le statut d'un R^2 et rapporte la co-structure à chacune des structures. Un des arguments les plus convaincants de la pertinence de cette notion est rapportée dans Holmes[11]. Pour un seul tableau $\mathbf{X}$, le tableau $\mathbf{Y}$ des q premières coordonnées est le tableau de rang q qui maximise le RV et le maximum atteint est :

$$RV_{max}(\mathbf{X},\mathbf{Y}) = \sqrt{\frac{\sum_{i=1}^{q}\lambda_i^2}{\sum_{i=1}^{q}\lambda_i^2}}$$

Pour résumer, la structure d'un tableau se mesure par la somme des carrés des variances des variables augmentée des carrés des covariances entre tous les couples. Quand on passe aux coordonnées, cette structure reste inchangée mais est concentrée dans les carrés des variances. La co-structure de deux tableaux se mesure par la somme des carrés des covariances de chaque couple formé d'une variable de l'un et d'une variable de l'autre. Cette quantité est décomposée par l'analyse de co-inertie. La mesure de co-structure divisée par la racine du produit des mesures de structures est un cosinus carré : le RV est à un couple de tableaux ce que le r^2 est à un couple de variables. Le RV est introduit en statistique dans [7] par Y. Escoufier. Le test non paramétrique de signification est décrit dans [10]. Ces notions s'étendent à tous les types d'analyse à schémas de dualité.

3 L'analyse de co-inertie

3.1 Pêches et nectarines

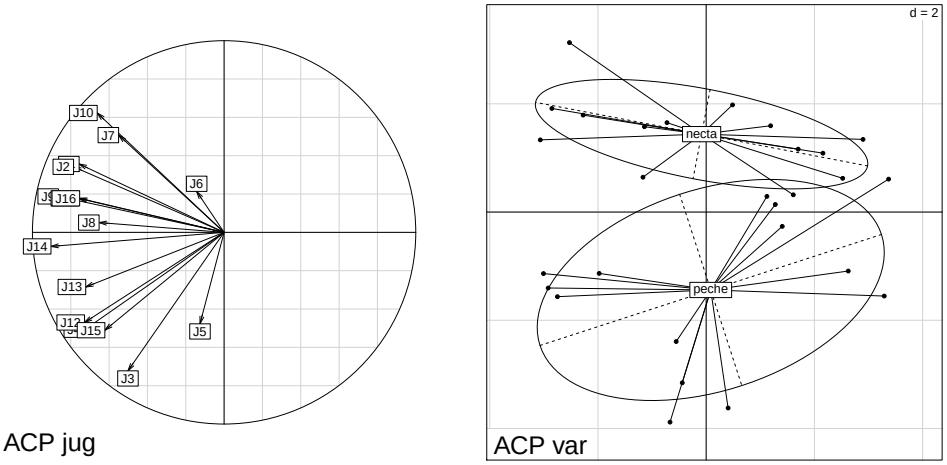
Considérons le jeu de données [13] conservé dans `fruits` :

```
data(fruits)
names(fruits)
names(fruits$var) <- c("taches", "stries", "abmucr", "irform", "allong",
    "suroug", "homlot", "homfru", "pubesc", "verrou", "foncee",
    "comucr", "impres", "coldom", "calibr")
```

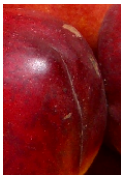
Le tableau `fruits$jug` a 28 lignes (produits) et 16 colonnes (juges). Chacun des juges a exprimé son opinion sur 28 lots de pêches en attribuant à chacun des lots un rang qui va de 1 (le plus apprécié) à 28 (le moins apprécié). Toutes les moyennes par colonne sont égales. Toutes les variances sont identiques. La corrélation mesure la ressemblance entre deux juges :

```
pcajug <- dudi.pca(fruits$jug, scan = F, nf = 4)
s.corcircle(pcajug$co, sub = "ACP jug", csub = 2)
```

Interpréter. Le vecteur `fruits$type` est un facteur à deux modalités `peche` et `necta` qui sépare les lots de fruits en pêches et nectarines. Pêches et nectarines sont de la même espèce et diffèrent par un seul gène qui affecte la peau. La pêche se reconnaît à sa peau duveteuse et veloutée. Celle des nectarines est lisse.




```
pcavar <- dudi.pca(fruits$var, scan = F, nf = 4)
s.class(pcavar$li, fruits$typ, sub = "ACP var", csub = 2)
```



Le tableau `fruits$var` contient un descriptif objectif du lot de fruits à l'aide de 15 variables :

taches la quantité de taches liègeuses (0=absente - 5) [1]

stries la quantité de stries (1/aucune - 4)

abmucr l'abondance du mucron (forme pointue à la base du fruit : 1/absent - 4/important) [2]

irform l'irrégularité de la forme (0/nulle - 3)

allong l'allongement du fruit (1/fruit rond - 4)

suroug le pourcentage de surface rouge (minimum 40

homlot l'homogénéité de coloration intra lot (1/forte - 4)

homfru l'homogénéité de coloration intra fruit (1/forte - 4)

pubesc la pubescence (0=nulle - 4)

verrou l'intensité du vert en zone rouge (1/nulle - 4)

foncee l'intensité des zones foncées (0/rose - 4)

comucr l'intensité de couleur du mucron (1=non contrasté - 4/foncé)

impres le type d'impression (1/lavé - 4/pointillé)

coldom l'intensité de la couleur dominante (0/claire - 4) [3]

calibr le calibre (1/<90g - 5/>200g)

[1] `http://www.srpv-centre.com/ulf/SRPV_biblio/ijhtml/IJ807_3.JPG`
[2] `\http://www.eecs.harvard.edu/~ilan/nectarines.jpg`
[3] `http://www.pomeroyfarm.com/nectarines.jpg`

On peut voir l'ACP comme un changement de base qui permet de remplacer les variables covariantes par des combinaisons linéaires non covariantes et de variances décroissantes.

```
foo1 <- function(x, y) round(t(as.matrix(x)) %*% as.matrix(y)/28,
    3)
foo1(pcajug$li, pcajug$li)
```

```
      Axis1 Axis2 Axis3 Axis4
Axis1 7.319 0.000 0.000   0.0
Axis2 0.000 2.612 0.000   0.0
Axis3 0.000 0.000 1.797   0.0
Axis4 0.000 0.000 0.000   1.3
```

```
round(pcajug$eig[1:4], 3)
```

```
[1] 7.319 2.612 1.797 1.300
```

```
foo1(pcavar$li, pcavar$li)
```

```
      Axis1 Axis2 Axis3 Axis4
Axis1 4.392 0.000 0.000  0.00
Axis2 0.000 3.229 0.000  0.00
Axis3 0.000 0.000 2.433  0.00
Axis4 0.000 0.000 0.000  1.51
```

```
round(pcavar$eig[1:4], 3)
```

```
[1] 4.392 3.229 2.433 1.510
```

La co-inertie de deux ACP étend cette propriété partiellement. En effet, la diagonalisation du schéma $(\mathbf{Y}^T\mathbf{X}, \mathbf{I}_p, \mathbf{I}_q)$ garantit que, si r est le rang de la matrice $\mathbf{Y}^T\mathbf{X}$, on dispose d'un système orthonormé $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ dans $\mathbb{R}^p$, d'un système orthonormé $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ dans $\mathbb{R}^q$ et d'un ensemble de valeurs propres non nulles $\{\omega_1, ..., \omega_r\}$ ayant les propriétés générales de tous les schémas de dualité. Si $\mathbf{U}$ et $\mathbf{V}$ sont les matrices $p - r$ et $q - r$ qui contiennent les systèmes de vecteurs propres, on a la décomposition de la co-inertie sous la forme :

$$S = trace\left(w\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\right) = \sum_{k=1}^{r} \omega_k$$

avec $w = \frac{1}{n^2}$. Comme $\mathbf{Y}^T\mathbf{X}\mathbf{U}\Omega^{-\frac{1}{2}} = \mathbf{V}$ et comme les systèmes de vecteurs sont orthonormés dans les deux espaces :

$$k \neq k' \Rightarrow \mathbf{u}_k^T\mathbf{X}^T\mathbf{Y}\mathbf{v}_{k'} = 0$$

et

$$S = \sum_{k=1}^{p}\sum_{j=1}^{q}\left(\frac{1}{n}\mathbf{X}^{k\,T}\mathbf{Y}^j\right)^2 = \sum_{k=1}^{r}\left(\frac{1}{n}\mathbf{u}_k{}^T\mathbf{X}^T\mathbf{Y}\mathbf{v}_k\right)^2 = \sum_{k=1}^{r}\omega_k$$

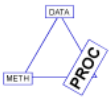
On est passé d'une co-inertie se décomposant en pq carrés de covariances à une décomposition en r valeurs rangées par ordre décroissant. On est passé de $p+q$ variables à $r+r$ variables par combinaisons linéaires dans chaque paquet. Mais on ne peut pas tout faire à la fois. La propriété fondamentale des coordonnées en ACP est d'être non corrélées, en co-inertie c'est d'être non corrélées avec les coordonnées de l'autre paquet à l'exception de celles de même rang.

```
copca <- coinertia(pcajug, pcavar, scan = F, nf = 4)
foo1(copca$lX, copca$lY)
```

```
      AxcY1 AxcY2 AxcY3 AxcY4
AxcX1  3.89 0.000 0.000 0.000
AxcX2  0.00 2.388 0.000 0.000
AxcX3  0.00 0.000 1.652 0.000
AxcX4  0.00 0.000 0.000 0.926
```

```
round(sqrt(copca$eig[1:4]), 3)
```

```
[1] 3.890 2.388 1.652 0.926
```



3.2 La classe coi

Une analyse de co-inertie est un objet des classes `dudi` et `coi`. Il est disponible pour deux normes diagonales. Les composantes de la liste ont une signification générale mais pourra prendre dans chaque type de couplage une signification particulière. Commençons par deux ACP normées.

```
copca
```

```
Coinertia analysis
call: coinertia(dudiX = pcajug, dudiY = pcavar, scannf = F, nf = 4)
class: coinertia dudi
$rank (rank)     : 15
$nf (axis saved) : 4
$RV (RV coeff)   : 0.4927474

eigen values: 15.13 5.704 2.728 0.8568 0.5648 ...

  vector length mode    content
1 $eig   15     numeric eigen values
2 $lw    15     numeric row weigths (crossed array)
3 $cw    16     numeric col weigths (crossed array)

   data.frame nrow ncol content
1  $tab       15   16   crossed array (CA)
2  $li        15   4    Y col = CA row: coordinates
3  $l1        15   4    Y col = CA row: normed scores
4  $co        16   4    X col = CA column: coordinates
5  $c1        16   4    X col = CA column: normed scores
6  $lX        28   4    row coordinates (X)
7  $mX        28   4    normed row scores (X)
8  $lY        28   4    row coordinates (Y)
9  $mY        28   4    normed row scores (Y)
10 $aX        4    4    axis onto co-inertia axis (X)
11 $aY        4    4    axis onto co-inertia axis (Y)
```

tab est la matrice $\mathbf{Y}^T\mathbf{DX}$ des produits scalaires entre colonnes de $\mathbf{X}$ et colonnes de $\mathbf{DX}$. Les éléments peuvent être des moyennes, des covariances, des corrélations, des cosinus suivant les tableaux d'origine. Ici, ce sont des corrélations entre les variables des deux groupes.

```
round(copca$tab[, 1:5], 2)
```

```
          J1    J2    J3    J4    J5
taches  0.00 -0.18  0.17 -0.37 -0.14
stries  0.42  0.36 -0.21 -0.03 -0.19
abmucr  0.24  0.13 -0.20  0.00  0.11
irform  0.23  0.12 -0.05  0.08  0.16
allong  0.05 -0.03 -0.24 -0.27  0.29
suroug -0.54 -0.45 -0.13 -0.21  0.29
homlot  0.27  0.15  0.43  0.38  0.16
homfru  0.57  0.49  0.15  0.23 -0.30
pubesc  0.42  0.50 -0.40 -0.04 -0.22
verrou  0.40  0.38  0.31  0.22 -0.14
foncee -0.06  0.02 -0.22 -0.26 -0.22
comucr  0.50  0.39 -0.16  0.09 -0.01
impres  0.09 -0.10  0.02 -0.11 -0.19
coldom -0.52 -0.41  0.10 -0.16  0.09
calibr  0.06  0.01 -0.77 -0.68 -0.16
```

cw est la métrique diagonale de $\mathbb{R}^p$ (poids des colonnes de $\mathbf{X}$, ici tous égaux à 1).

lw est la métrique diagonale de $\mathbb{R}^q$ (poids des colonnes de $\mathbf{Y}$, ici tous égaux à 1).

eig contient les valeurs propres de l'analyse, carrés des produits scalaires (en général des covariances) entre les coordonnées de co-inertie de même rang.

c1 donne les axes de co-inertie dans $\mathbb{R}^p$, vecteurs normés en colonnes. Ici, on peut les identifier à des *loadings*. **l1** donne les axes de co-inertie dans $\mathbb{R}^q$, vecteurs normés en colonnes. Même remarque.

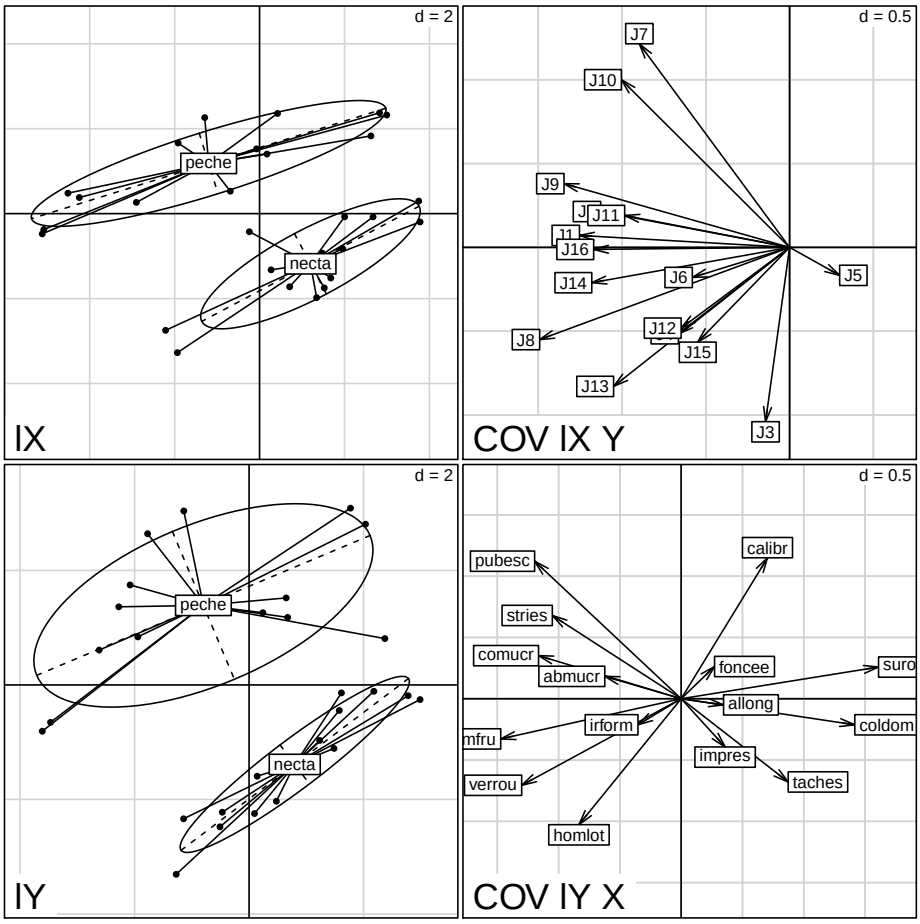


FIGURE 1 – A gauche plans de co-inertie par projection de nuages de points. A droite illustration du plan par les covariances avec les variables de l'autre groupe. Le compromis des juges se fait au détriment des pêches. En effet, une **note élevée** correspond à un **avis défavorable**. Une autre partie du compromis se fait à l'intérieur de chaque type de fruits. Les variables qui séparent les types et les autres participent donc à la formation du jugement majoritaire des juges.

co contient les produits scalaires entre colonnes de $\mathbf{X}$ et coordonnées de co-inertie dans $\mathbb{R}^q$. Ici, ce sont des covariances, dans la logique covariances des variables d'un bloc avec les scores de l'autre.

li contient les produits scalaires entre colonnes de $\mathbf{Y}$ et coordonnées de co-inertie dans $\mathbb{R}^p$. Même remarque.

lX est le tableau des coordonnées de co-inertie dans $\mathbb{R}^p$ donnant les projections des lignes de $\mathbf{X}$ sur les axes de co-inertie dans $\mathbb{R}^p$. Ce sont de vraies coordonnées d'ACP et peuvent être utilisées de toutes sortes de manières comme des scores.

lYest le tableau des coordonnées de co-inertie dans $\mathbb{R}^q$ donnant les projections des lignes de $\mathbf{Y}$ sur les axes de co-inertie dans $\mathbb{R}^q$. Même remarque.

Nous faisons deux ACP coordonnées avec cette pratique. Les scores sont optimaux par leurs liens avec les variables de l'autre. On voit la variabilité de chaque nuage dans ce qu'elle s'organise par rapport à l'autre (figure 1).

```
par(mfrow = c(2, 2))
s.class(copca$lX, fruits$typ, sub = "lX", csub = 2)
s.arrow(copca$co, sub = "COV lX Y", csub = 2)
s.class(copca$lY, fruits$typ, sub = "lY", csub = 2)
s.arrow(copca$li, sub = "COV lY X", csub = 2)
```

RV conserve la valeur du coefficent RV. Il est compris entre 0 et 1. Ici, pour les métriques canoniques de $\mathbb{R}^p$ et $\mathbb{R}^q$, la pondération uniforme et les tableaux normalisés des deux ACP normées de départ, on retrouve directement ce résultat par :

```
copca$RV
```

```
[1] 0.4927474
```

```
sum(cor(pcajug$tab, pcavar$tab)^2)/sqrt(sum(cor(pcajug$tab, pcajug$tab)^2) *
    sum(cor(pcavar$tab, pcavar$tab)^2))
```

```
[1] 0.4927474
```

3.3 Aides à l'interprétation

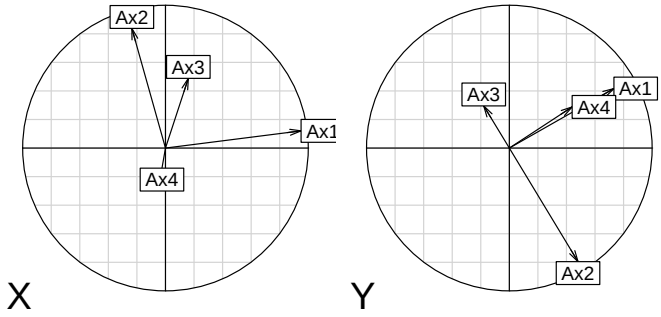
Trois compléments doivent être compris pour tirer tout le possible de cette pratique.

Le **premier** est canonique (fait règle). Dans l'espace des deux nuages de points, nous avons deux systèmes d'axes, ceux des analyses simples et ceux de l'analyse de co-inertie. Il est toujours intéressant de savoir si on s'est éloigné sensiblement des structures internes en imposant le couplage.

aX contient les coordonnées de la projection des axes d'inertie dans $\mathbb{R}^p$ (analyse initiale de $\mathbf{X}$ sur les axes de co-inertie dans $\mathbb{R}^p$.

aY contient les coordonnées de la projection des axes d'inertie dans $\mathbb{R}^q$ (analyse initiale de $\mathbf{Y}$ sur les axes de co-inertie dans $\mathbb{R}^q$.

```
par(mfrow = c(1, 2))
s.corcircle(copca$aX, sub = "X", csub = 2)
s.corcircle(copca$aY, sub = "Y", csub = 2)
```



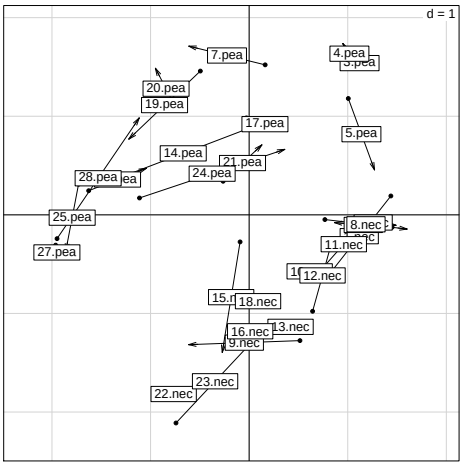
On peut voir ainsi comment il fallait tourner ou retourner certains plans d'une analyse pour retrouver certains plans de l'autre.

Le **second** complément n'est pas canonique mais est bien pratique.

mX contient les scores normés ($\mathbb{R}^n$) obtenus ici en ramenant à l'unité la variance des coordonnées de **lX**.

mY contient les scores normés ($\mathbb{R}^n$) obtenus ici en ramenant à l'unité la variance des coordonnées de **lY**. On peut donc superposer les deux plans, opération qui montre la partie corrélation dans la partie covariance qui est optimisée.

```
s.match(copca$mX, copca$mY)
```



On renforce ainsi la perception commune des deux plans.

Enfin, le **troisième** complément est numérique :

```
summary(copca)
```

```
Eigenvalues decomposition:
        eig    covar       sdX      sdY      corr
1 15.1338350 3.890223 2.6075806 1.864335 0.8002263
2  5.7037338 2.388249 1.5506657 1.776134 0.8671329
3  2.7282313 1.651736 1.4713560 1.433355 0.7831937
4  0.8568131 0.925642 0.9718156 1.312019 0.7259706
Inertia & coinertia X:
       inertia       max     ratio
1     6.799477  7.318882 0.9290322
12    9.204041  9.930650 0.9268317
```

```
123  11.368929 11.727775 0.9694021
1234 12.313355 13.027723 0.9451655

Inertia & coinertia Y:
      inertia      max     ratio
1     3.475745  4.391663 0.7914416
12    6.630397  7.620306 0.8700960
123   8.684903 10.053072 0.8639054
1234 10.406297 11.563570 0.8999208

RV:
 0.4927474
```

Les racines de valeurs propres sont décomposées en écart-type de score de $\mathbf{X}$, écart-type de score de $\mathbf{Y}$ et corrélation entre les deux. Les inerties cumulées qui se projettent sur les axes de co-inertie sont comparées aux valeurs optimales définies par les analyses simples. Ces éléments permettent de comprendre les particularités d'une double analyse d'inertie coordonnée. Cette méthode s'étend à tout couple de deux analyses ayant même pondération des lignes. On pourra trouver des références bibliographiques et des compléments théoriques et des exemples sur ce site.

- ⋆ La méthode niche est une co-inertie adaptée à l'étude de la marginalité des espèces :

  `http://pbil.univ-lyon1.fr/R/querep/qrb.pdf`

- ⋆ Pour l'utilisation d'un très grand nombre de variables dans les deux tableaux :

  `http://pbil.univ-lyon1.fr/R/querep/qr2.pdf`

- ⋆ La place de la co-inertie dans l'ensemble des méthodes de couplage :

  `pbil.univ-lyon1.fr/R/stage/stage5.pdf`

- ⋆ Le lien avec les rotations procustéennes :

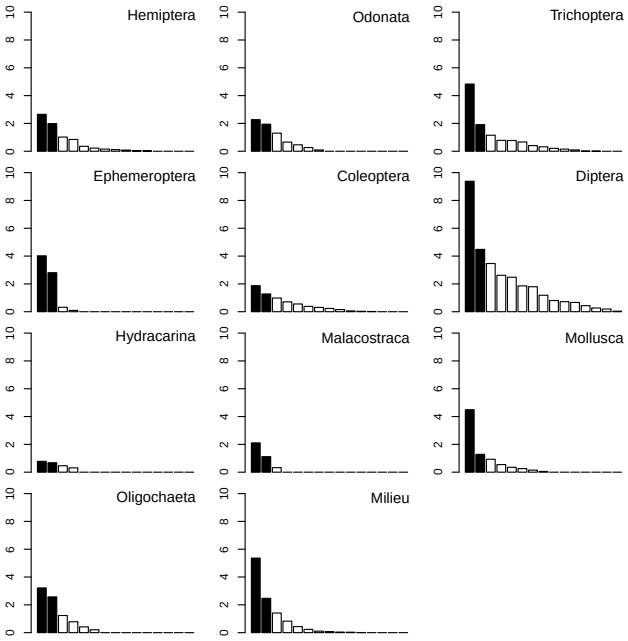
  `pbil.univ-lyon1.fr/R/fichestd/tdr64.pdf`

L'introduction est ici utile pour comprendre les stratégies K-tableaux. Dès qu'on introduit un tableau supplémentaire, la confrontation va passer d'une comparaison binaire à une comparaison de chaque élément à un élément commun appelé compromis. Il y a deux grande stratégies. La première définit le compromis puis le décompose, la seconde compose le compromis élément par élément.

4 Compromis : calcul et analyse par *STATIS*

Nous avons vu que les tableaux sont appariés par les lignes mais qu'on utilise cette dimension pour parler aussi bien des individus (multiplicité des ensembles de descripteurs) que des variables (multiplicité des ensembles de points). Dans un cas comme dans l'autre, on a affaire à K schémas de dualité.

```
data(friday87)
w1 <- cbind.data.frame(scalewt(friday87$fau, scale = F), scalewt(friday87$mil))
kta1 <- ktab.data.frame(w1, c(friday87$fau.blo, 11), tabnames = c(friday87$tab.names,
    "Milieu"))
sep1 <- sepan(kta1)
plot(sep1)
```



4.1 Inter-structure

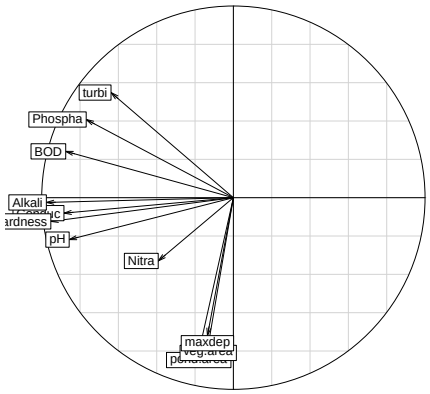
Chaque tableau définit une structure. Selon les critères habituels, la mesure de la structure d'un tableau est son inertie totale. Avant cela le nombre de variables est essentiel. *STATIS* [15] [16] propose de mesurer la valeur d'une analyse non par la somme des valeurs propres d'une analyse élémentaire (qui vaut l'inertie totale) mais par la somme des carrés de ces valeurs propres. Il s'agit d'abord dans cette modification d'une conséquence de la logique algébrique sous-jacente. En fait, la signification écologique de cette innovation ne saurait échapper à un ... écologue. En effet, on sait que plus les valeurs propres d'une analyse sont différentes (c'est-à-dire, plus les premières valeurs propres sont grandes et les suivantes petites), meilleure est l'analyse (au sens que l'expression issue de la projection ne peut être un artefact). En passant de la somme (ou la moyenne) à la somme des carrés, on passe de la moyenne à la variance, de l'abondance à la diversité. Plus l'inertie exprimée sur les premiers axes est grande, pour une valeur d'inertie donnée, plus la norme de l'opérateur est grande. Le groupe des éphéméroptères l'emporte sans conteste et c'est un autre groupe pauvre qui prend la seconde valeur.

```
nvar <- kta1$blo
rank <- sep1$rank
fac <- factor(rep(1:11, sep1$rank))
inertia <- tapply(sep1$Eig, fac, sum)
vav <- tapply(sep1$Eig, fac, function(x) sum(x * x))
round(cbind(nvar, rank, inertia, vav), dig = 2)
```

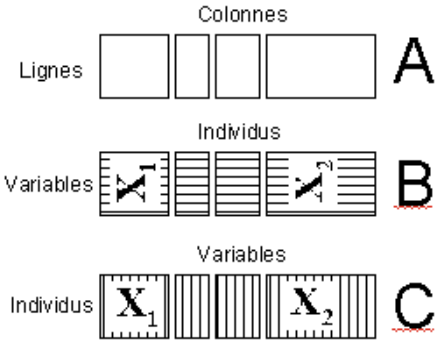
```
             nvar rank inertia    vav
Hemiptera      11   11    7.52  12.97
Odonata         7    7    7.00  11.36
Trichoptera    13   13   11.32  30.33
Ephemeroptera   4    4    7.23  24.10
Coleoptera     13   12    6.58   7.22
Diptera        22   15   30.39 143.10
Hydracarina     4    4    2.23   1.38
Malacostraca    3    3    3.55   5.79
Mollusca        8    8    8.05  23.17
Oligochaeta     6    6    8.43  19.30
Milieu         11   11   11.00  37.76
```

On reconnaît aussi l'inertie de l'ACP normée qui est égale au nombre de variables et qui donne une structure simple à deux gradients (taille sur le facteur 2, charge minérale et organique sur le facteur 1) :

```
s.corcircle(dudi.pca(friday87$mil, scan = F)$co)
```



On s'attend à ce qu'un groupe faunistique comportant de nombreuses espèces fournisse une inertie totale (somme des variances des abondances des taxons) plus grande. Il n'en est rien et la corrélation inertie-richesse est nulle. On peut dire qu'utilisés seuls, les groupes 1, 2, 5, 7, 8 et 10 n'auraient pas grand chose à dire sur une éventuelle typologie de relevés. On garderait pour les groupes 3, 6 et 9 le premier axe et pour le seul groupe 4 les deux premiers. Curieusement il n'y a que 4 espèces d'éphéméroptères et c'est le tableau qui semble le plus structuré. A la notion de mesure de structure, on ajoute la mesure de co-structure par le produit scalaire de la covariance vectorielle qui est la co-inertie totale associée au couple de tableaux qui se décompose dans la somme des valeurs propres de l'analyse de co-inertie. Il s'en suit qu'on peut mesurer la corrélation entre deux triplets par le coefficient de corrélation vectoriel ou RV. Nous ne ferons pas la distinction habituellement pratiquée entre les opérateurs **VQ** et les opérateurs **WD** de la théorie des opérateurs d'Escoufier. Cette distinction s'exprime par :



La structure commune (**A**) recouvre le cas **B** (mêmes variables, *STATIS* sur les **VQ**) et le cas **C** (mêmes individus *STATIS* sur les **WD**). La fonction `statis` d'`ade4` recouvre les deux cas de manière identique. Chaque groupe de variables fait ici une typologie d'individus et *STATIS* mesure la corrélation entre ces typologies :

```
statis1 <- statis(kta1, scannf = F)
names(statis1)
```

```
 [1] "RV"        "RV.eig"    "RV.tabw"   "RV.coo"    "C.eig"     "C.nf"
 [7] "C.rank"    "C.li"      "C.Co"      "C.T4"      "cos2"      "tab.names"
[13] "TL"        "TC"        "T4"
```

Éditer la matrices des *RV* :

```
w <- round(statis1$RV, 2)
dimnames(w)[[2]] <- abbreviate(dimnames(w)[[2]], 2)
w
```

```
             Hm   Od   Tr   Ep   Cl   Dp   Hy  Mlc  Mll   Ol  Mil
Hemiptera  1.00 0.44 0.53 0.44 0.50 0.43 0.50 0.35 0.41 0.39 0.34
Odonata    0.44 1.00 0.51 0.57 0.41 0.64 0.31 0.43 0.49 0.41 0.43
Trichoptera 0.53 0.51 1.00 0.54 0.46 0.61 0.43 0.51 0.42 0.44 0.49
Ephemeroptera 0.44 0.57 0.54 1.00 0.31 0.62 0.32 0.60 0.61 0.41 0.42
Coleoptera 0.50 0.41 0.46 0.31 1.00 0.45 0.47 0.31 0.50 0.25 0.29
Diptera    0.43 0.64 0.61 0.62 0.45 1.00 0.34 0.49 0.53 0.51 0.66
Hydracarina 0.50 0.31 0.43 0.32 0.47 0.34 1.00 0.42 0.59 0.28 0.34
Malacostraca 0.35 0.43 0.51 0.60 0.31 0.49 0.42 1.00 0.64 0.34 0.33
Mollusca   0.41 0.49 0.42 0.61 0.50 0.53 0.59 0.64 1.00 0.24 0.37
Oligochaeta 0.39 0.41 0.44 0.41 0.25 0.51 0.28 0.34 0.24 1.00 0.71
Milieu     0.34 0.43 0.49 0.42 0.29 0.66 0.34 0.33 0.37 0.71 1.00
```

Ici est le cœur de la méthode. Tous les coefficients sont positifs car ce sont des carrés de corrélation (le signe d'un *RV* n'a pas de sens). Mais ils sont assez élevés et homogènes. Retrouver l'un d'entre eux par une co-inertie.

```
coinertia(dudi.pca(kta1[[1]], , , F, F, F), dudi.pca(kta1[[2]],
    , , F, F, F), F)$RV
```

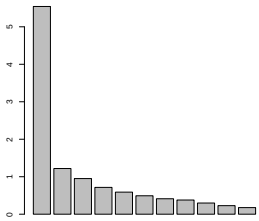
```
[1] 0.4417286
```

```
statis1$RV[1, 2]
```

```
[1] 0.4417286
```

Il est hors de question de dépouiller 55 analyses de co-inertie ! La matrice des *RV* est diagonalisée. On obtient une image euclidienne des tableaux pour ce produit scalaire. Mais contrairement au cas général d'une ACP normée, les *RV* sont toujours positifs ou nuls et le premier vecteur propre définit toujours une pondération des tableaux (théorème de Perron-Frobenius).

```
barplot(statis1$RV.eig)
```



Cette forme est naturelle et dérive directement du caractère positif des RV. Si la première valeur propre n'était pas nettement dominante, on aurait l'indication d'une absence de compromis et donc d'une analyse sans objet. Les opérateurs d'inertie ont été normés avant d'en faire une combinaison linéaire. C'est un choix simplifié qui élimine le rôle en général écrasant des opérateurs "volumineux". On a donc débarrassé chaque tableau de son inertie et de son importance typologique en prenant comme compromis la moyenne des matrices de produits scalaires entre sites avec une variance vectorielle ramenée à l'unité. On écrit en toute généralité :

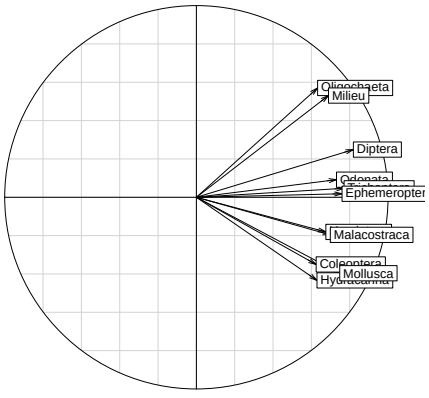
$$\mathbf{WD} = \sum_{k=1}^{K} \alpha_k \frac{\mathbf{W}_k\mathbf{D}}{\|\mathbf{W}_k\mathbf{D}\|}$$

Mais il est surtout important d'en comprendre le sens. Pour un tableau, on a un nuage de points-sites qui est ordonné. La géométrie de ce nuage est dans la matrice des produits scalaires. Celle-ci est mise à l'échelle commune pour éviter que dans le mélange un tableau emporte le tout. La moyenne de ces matrices est calculée pour des poids optimaux (on fait un mélange le plus interprétable possible). Ces poids sont ici :

```
statis1$RV.tabw
[1] 0.2869362 0.3105288 0.3273297 0.3228044 0.2654266 0.3479269 0.2672668 0.2968233
[9] 0.3177330 0.2686765 0.2927640
```

Chaque tableau, quelle que soit sa structure initiale, participe à égalité à la constitution du compromis dont la diagonalisation fait une typologie moyenne. Remarquer que le tableau de milieu n'occupe pas une position centrale.

```
s.corcircle(statis1$RV.coo)
```



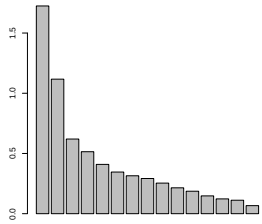
A retenir : on peut faire un cercle de corrélation dont les éléments sont des tableaux. Vue la liste des marqueurs complexes que la biologie invente, on peut se demander pourquoi ceci est pratiquement inconnu. Voir une discussion dans :

http://pbil.univ-lyon1.fr/R/articles/chessel_atm.pdf

4.2 Compromis

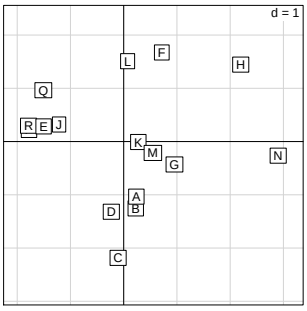
La cohérence des tableaux n'est pas très grande mais la constitution d'un compromis est légitime. La suite change alors de point de vue. Observer si les tableaux se ressemblent, donner une image euclidienne de l'ensemble des tableaux, c'est examiner l'inter-structure. Moyenner les opérateurs pour faire un compromis, c'est en quelque sorte trouver la moyenne des structures. Analyser cette moyenne, c'est faire une nouvelle analyse dite analyse du compromis. Ici, le compromis est franchement de dimension 2.

```
barplot(statis1$C.eig)
```



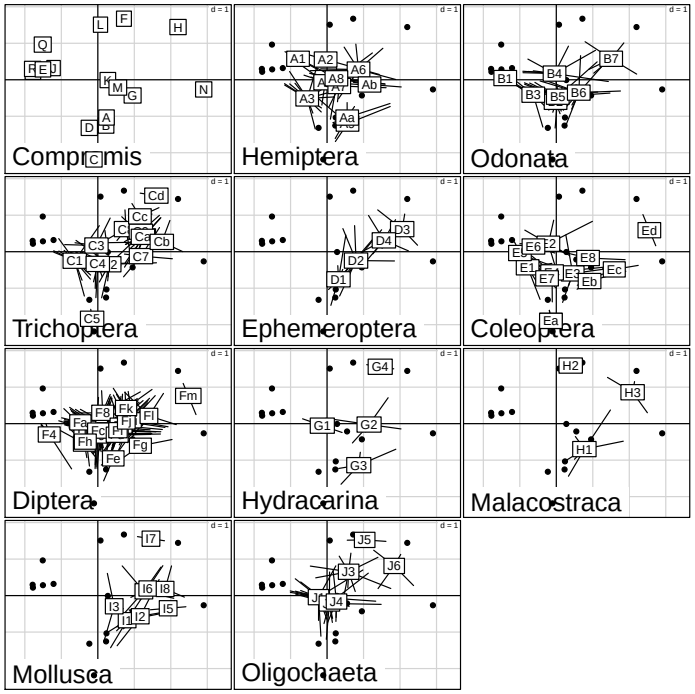
On génère donc une typologie de référence :

```
s.label(statis1$C.li)
```



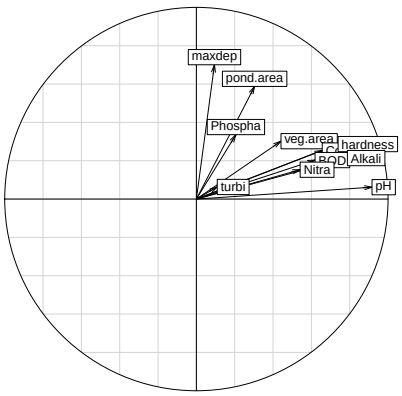
L'interprétation dépend de chaque cas particulier. Ici, on pourra chercher à regarder comment chaque groupe d'espèces réagit à cette typologie :

```
kta2 <- ktab.data.frame(friday87$fau, friday87$fau.blo)
par(mfrow = c(4, 3))
s.label(statis1$C.li, clab = 1.5, sub = "Compromis", csub = 3)
for (j in 1:10) {
    s.distri(statis1$C.li, kta2[[j]], cell = 0, axesell = F, cstar = 0.5,
        clab = 1.5, cpoi = 2, sub = names(kta2)[[j]], csub = 3)
}
```



Les trichoptères, les éphéméroptères et les quatre derniers groupes sont largement absents des stations de gauche, alors que les hémiptères, les coléoptères, les diptères (dans une mesure moindre) ne présentent pas cet effet. En fait on a là l'effet conjugué de deux éléments : le nombre d'espèces augmente avec la taille du lac, d'une part, l'acidité des eaux étant le facteur limitant d'autre part :

```
s.corcircle(statis1$C.Co[statis1$TC[, 1] == 11, ])
```



On retrouvera tous ces éléments dans les graphes génériques de `statis` :

```
kplot(statis1)
plot(statis1)
```

Un facteur écologique limitant (pH) est omniprésent dans cette observation, son rôle suivant les groupes d'espèces varie fortement. Foster [8] apporte des arguments très cohérents avec ces résultats. Nous n'avons pas introduit l'étude de l'intra-structure qui est la représentation des trajectoires (chaque individu dans chaque tableau) car l'autre stratégie fait sensiblement mieux en la matière. L'intra-structure a un sens précis, par exemple dans [9]. Il existe enfin des cas où *STATIS* sur les tableaux et *STATIS* sur les opérateurs sont en concurrence. On peut très bien avoir des tableaux totalement appariés et utiliser une méthode portant sur les opérateurs. On multiplie malheureusement les choix possibles (figure 2). On trouvera la question des traits biologiques abordée de ce point de vue dans :

`http://pbil.univ-lyon1.fr/R/querep/qr9.pdf`

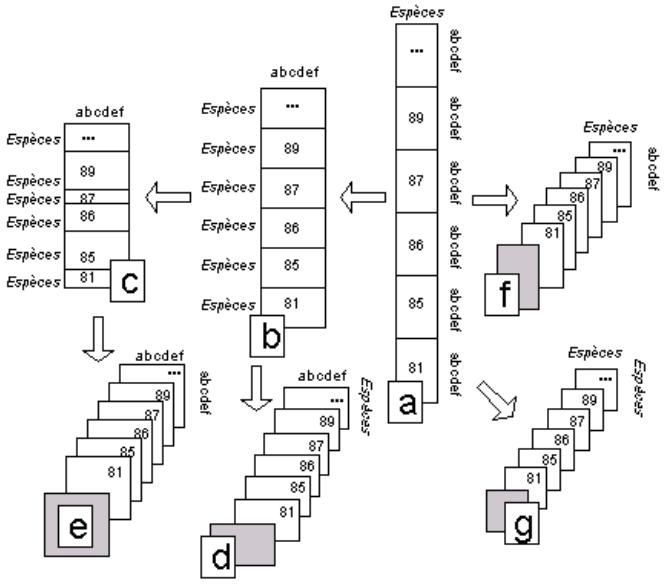


FIGURE 2 – Multiples manières de relier un cube de données espèces-sites-dates à une analyse K-tableaux. C'est sans doute une des difficultés principales : le format d'enregistrement ne guide pas l'analyse statistique.

5 Compromis d'analyse d'inertie

5.1 Microsatellites et races bovines

Reprendre la liste `microsatt` . On y trouve les fréquences alléliques chez 18 races bovines :

sur 9 sites microsatellites (13792 typages à 1 euro 50 pièce). De quoi se poser la question de la pertinence des descripteurs [17].

```
data(microsatt)
sum(microsatt$tab)
```

```
[1] 13792
```

```
colo = 1:18
colo[c(1:2, 7, 11, 14)] = "red"
colo[c(3:6, 8:10, 12:13, 15)] = "blue"
colo[16:18] = "green"
```

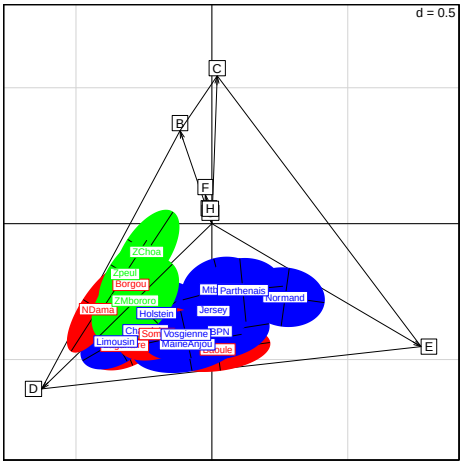
Consulter la fiche des données pour donner un sens à la coloration mise en place. Identifer le type de données en isolant la premier locus :

```
w <- microsatt$tab[, 1:8]
names(w) = LETTERS[1:8]
w
```

```
           A  B  C  D  E F G H
Baoule     0  0  5 37 32 0 0 0
Borgou     0 17 18 56  9 0 0 0
BPN        0  0 10 36 32 0 0 0
Charolais  0  0 16 64 20 0 0 0
Holstein   0  0 21 57 20 0 0 0
Jersey     0  0 20 44 36 0 0 0
Lagunaire  0  1 10 70 16 0 1 0
Limousin   0  0 12 66 12 0 0 0
MaineAnjou 0  0  6 35 21 0 0 0
Mtbeliard  0  0 27 38 35 0 0 0
NDama      0  0 15 43  2 0 0 0
Normand    0  0 22 24 54 0 0 0
Parthenais 0  0 26 33 41 0 0 0
Somba      0  1 12 60 23 2 0 0
Vosgienne  0  0 13 55 32 0 0 0
ZChoa      1  8 17 29  4 5 1 1
ZMbororo   2  5  8 30  6 1 0 0
Zpeul      0 10 26 56  5 2 0 1
```

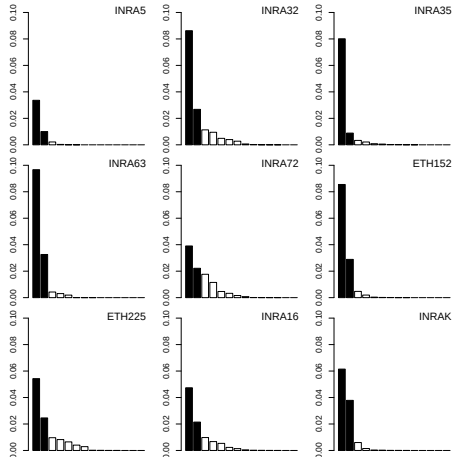
C'est un problème d'AFC (dénombrements alléliques) ou d'ACP (fréquences alléliques). Le premier point de vue est plus justifié, plus rarement utilisé en génétique, plus difficile en théorie à étendre à plusieurs loci. Le second est plus habituel, il est cohérent avec la notion de FST. Si on veut résumer ce tableau, on fera :

```
wpc <- sweep(w, 1, rowSums(w), "/")
wsco <- dudi.pca(wpc, scal = F, scan = F)$c1
s.multinom(wsco, w, coul = colo)
```



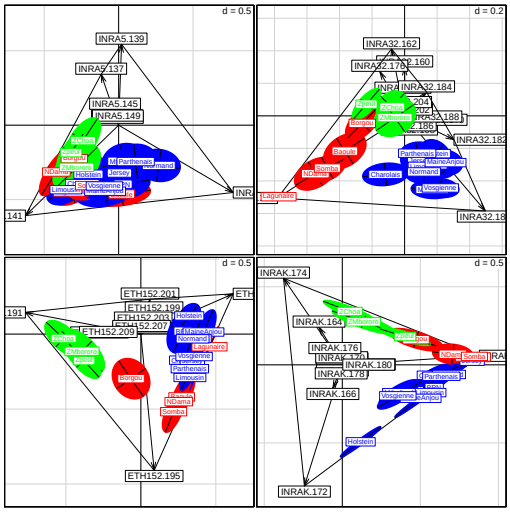
Cette figure est très simple à comprendre. Las allèles sont placés par deux scores de moyennes nulles et de covariance nulle. La somme des carrés des scores vaut l'unité. Cette disposition est la meilleure qui soit pour séparer les populations placées au score moyen des allèles qu'on y trouve. On peut utiliser cette pratique pour la description des contenus stomacaux [2]. On y rajoute ici la représentation de la variabilité de cette position sous l'hypothèse d'un tirage multinomial (intervalle de confiance à 95%) de la position de la moyenne des allèles avec les fréquences observées. Ici, le marqueur a de piètres capacités descriptives ! Ce graphe contient toute l'information si on ne garde que deux facteurs. C'est généralement le cas.

```
fuz = prep.fuzzy.var(microsatt$tab, microsatt$loci.eff)
pcafuz = dudi.fpca(fuz, scan = F)
miktab = ktab.data.frame(pcafuz$tab, microsatt$loci.eff, rownames = row.names(fuz),
    tabnames = microsatt$loci.names)
misep <- sepan(miktab)
plot(misep)
```



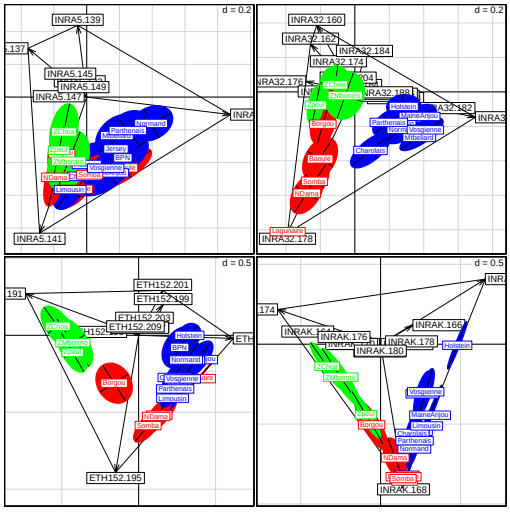
Mais la question est clairement posée dans la figure :

```
par(mfrow = c(2, 2))
for (k in c(1, 2, 6, 9)) {
    X <- misep$Ci[misep$TC[, 1] == k, ]
    index <- c(1, cumsum(microsatt$loci.eff) + 1)
    Y <- microsatt$tab[, index[k]:(index[k + 1] - 1)]
    row.names(X) <- names(Y)
    s.multinom(X, Y, coul = colo)
}
```



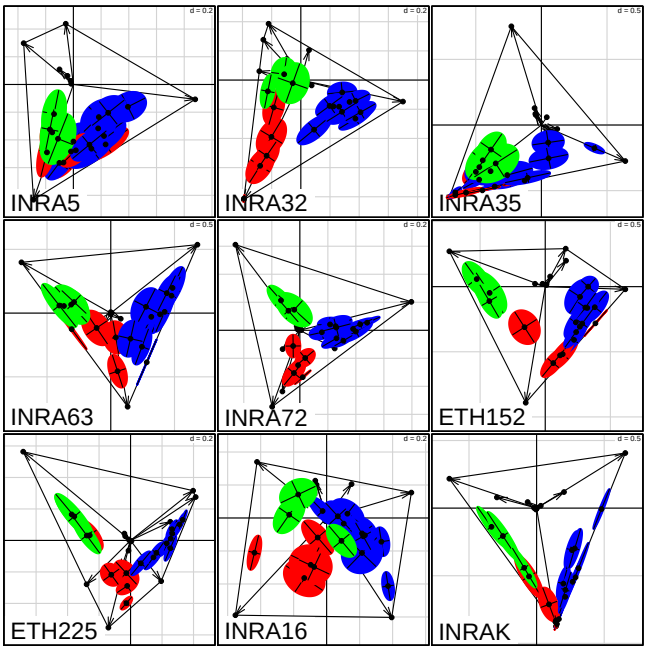
La question est bien posée : on veut la même chose mais de manière cohérente ! Rien de plus simple [12], en apparence :

```
mcoa1 <- mcoa(miktab, scannf = F)
par(mfrow = c(2, 2))
index <- c(1, cumsum(microsatt$loci.eff) + 1)
for (k in c(1, 2, 6, 9)) {
    X <- mcoa1$axis[mcoa1$TC$T == k, ]
    Y <- microsatt$tab[, index[k]:(index[k + 1] - 1)]
    row.names(X) <- names(Y)
    s.multinom(X, Y, coul = colo)
}
```



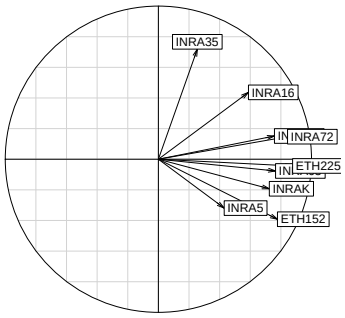
Finalement, le tableau entier se résume par :

```
par(mfrow = c(3, 3))
for (k in 1:9) {
    X <- mcoa1$axis[mcoa1$TC$T == k, ]
    Y <- microsatt$tab[, index[k]:(index[k + 1] - 1)]
    row.names(X) <- names(Y)
    s.multinom(X, Y, coul = colo, clabelcat = 0, clabelrowprof = 0,
        sub = tab.names(miktab)[k], csub = 3)
}
```



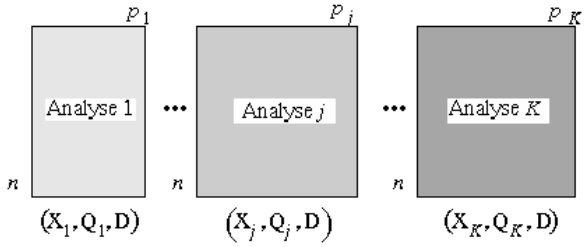
L'axe 1, c'est Afrique contre Europe ; l'axe 2, c'est taurins contre zébus en Afrique, mais quelle diversité de réalisation de cette structure ! Les bovins africains sont parfois aussi bovins que les européens, parfois aussi africains que les zébus et parfois originaux. Les marqueurs sont loin d'être totalement redondants. La synthèse que fait alors `statis` de leur fonction est saisissante. Le premier marqueur est le plus confus :

```
s.corcircle(statis(miktab, scan = F)$RV.coo)
```



5.2 Schéma de principe

Dans l'exemple qui précède, on ne sait plus si ce sont les données qui aident à comprendre la méthode ou si c'est la méthode qui aide à comprendre les données. Chacun fera selon son intérêt du moment. L'analyse de co-inertie multiple ($ACOM$) prend sa source dans l'analyse factorielle multiple (AFM). L'AFM d'origine n'a envisagé que le cas des tableaux appariés par les individus alors que la théorie ne semble pas imposer ce point de vue unique. C'est cependant le cas de l'exemple en cours. On dispose de K tableaux ayant en commun les lignes-individus, chacun d'entre eux correspondant à un groupe de variables-colonnes.



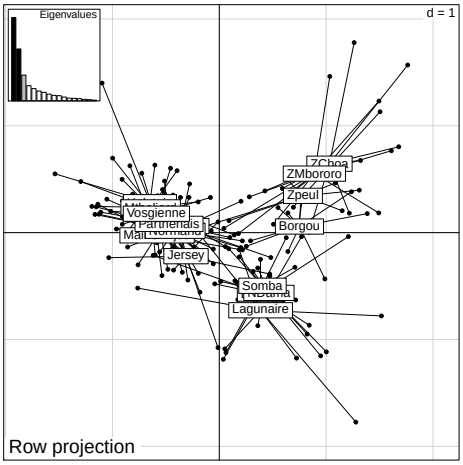
On dispose donc de K triplets statistiques. L'objectif est d'ordonner simultanément K tableaux, de réaliser K analyses portant sur un même ensemble d'individus, ou encore de comparer K groupes de variables définis sur le même ensemble d'individus comme indiqué dans le premier texte présentant la méthode [4] [3]. L'AFM est basée sur l'analyse du tableau accolé $[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K]$ avec les pondérations lignes et colonnes commune (pour les lignes) et concaténées (pour les colonnes). Mais une opération préliminaire essentielle est introduite pour uniformiser le rôle des tableaux dans l'analyse simultanée. On multiplie chaque tableau par un poids qui diminue l'importance des grands tableaux et augmente celle des petits.

```
args(mfa)
```

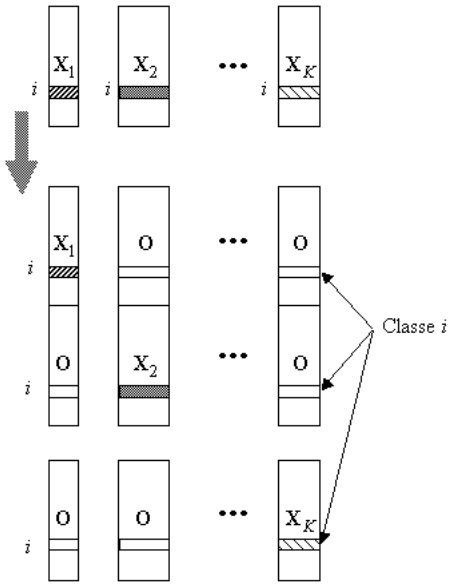
```
function (X, option = c("lambda1", "inertia", "uniform", "internal"),
    scannf = TRUE, nf = 3)
NULL
```

Ce poids est par défaut l'inverse de la première valeur propre (`lambda1`) ou si on préfère l'inverse de l'inertie du tableau (`inertia`), ou encore un poids uniforme (`uniform`) si on pense que cette opération ne doit pas avoir lieu ou enfin un poids prédéfini dans le K-tableaux par une composante `tabw` (`internal`). Dans le premier cas, les tableaux juxtaposés et pondérés ont une même première valeur propre égale à 1, dans le second cas ils ont tous même inertie totale égale à 1. En première approche l'AFM est donc l'analyse du tableau juxtaposé avec une correction de poids. Comme on fait une analyse avec n individus et $p_1 + p_2 + ... + p_k$ variables, on a directement une carte des variables et un graphe de valeurs propres comme d'habitude. Sur les composantes principales, on peut projeter les composantes principales de chaque tableau et on observe (en bas à gauche) que dans $\mathbb{R}^n$, $STATIS$ a construit un compromis pour l'analyser alors que l'AFM a trouvé des composantes principales pour constituer un compromis mais que les deux ont trouvé des plans très voisins.

```
mimfa <- mfa(miktab, scan = F)
plot.mfa(mimfa, op = 1)
```



Une différence essentielle introduit la représentation simultanée d'un point et de chacune de ses réalisation par tableau. Nous savons qu'en tant qu'analyse d'un triplet particulier l'*AFM* renvoie à une analyse duale des individus. Un individu est une ligne du tableau juxtaposant les K tableaux. Les lignes du tableau k sont dans l'espace $\mathbb{R}^{p_k}$ et seule l'*ACOM* s'intéresse directement à ces K nuages de points. Les lignes du tableau global sont dans $\mathbb{R}^p$ où $p = p_1 + p_2 + ... + p_K$. Le nuage des n points de $\mathbb{R}^p$ donne la carte ordinaire. Mais une analyse peut en cacher une autre, ce qui s'exprime dans le graphe :

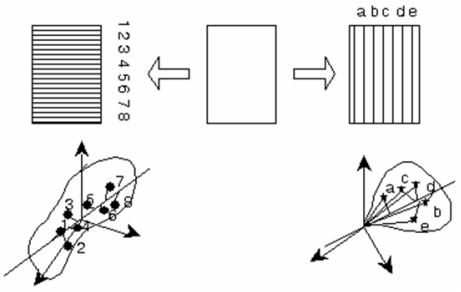


Une ligne y correspond à une ligne de départ et un tableau. Le reste est rempli de 0. Il s'agit évidemment d'une idée mathématique qui justifie que l'*AFM* est une inter-classe cachée qui permet de représenter globalement un point et chacune de ses expressions [5]. (Escofier et Pagès 1984). Ceci donne la représentation simultanée. On peut utiliser ces coordonnées par tableau soit pour les assembler (c'est la variance des positions moyennes qui est optimisée) et exprimer la cohérence des positions d'un même point, soit les dispatcher pour les

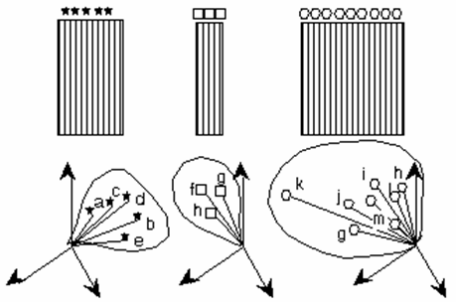
représenter avec le nuage des variables correspondant pour observer ensemble les K analyses, mais l'*ACOM* fera cela plus clairement avec une contrainte supplémentaire. La procédure `mfa` généralise la méthode à tous types de variables et renvoie à une indication [6] qui sous-tend la discussion :

> First applications of M.F.A. were highly encouraging as regards to the practical value of the method. ... The last example clearly shows that to take into account variable groups is not only a technical problem, susceptible of being solved by an appropriate method, but it also can be considered as a methodological problem which enriches the field of data Analysis.

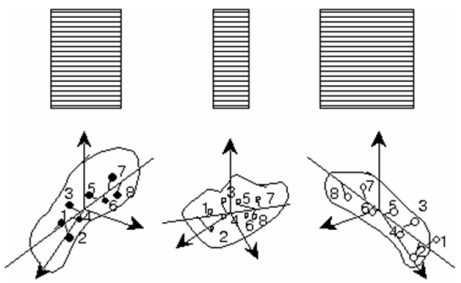
Très proche de la précédente, l'analyse de co-inertie multiple traite de la même situation. Un tableau individus-variables engendre deux nuages de points :



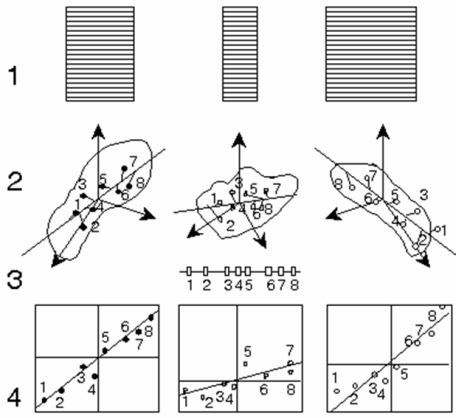
K tableaux de données donnent donc K nuages de variables dans un même espace et ce fait est exploité par l'*AFM* :



Ils donnent K opérateurs d'inertie dans un même espace et ce fait est exploité par *STATIS*. Ils donnent aussi K nuages de points appariés dans K espaces différents et ceci est pris en compte par l'*ACOM* :



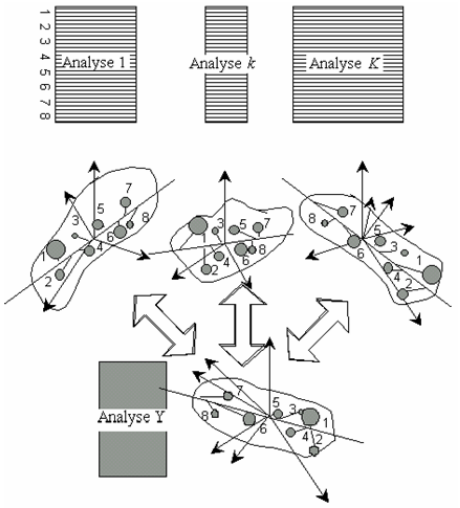
On voit qu'en partant des objets créés par les mêmes données plusieurs approches sont concurrentes. L'analyse linéaire d'un tableau a ceci de caractéristique : elle aborde les deux nuages de points dualement. C'est la base de son succès : tout résultat obtenu sur un des deux objets se retrouve exprimé directement sur l'autre et réciproquement. Mais dès qu'il y a plus d'un tableau, la symétrie lignes-colonnes est détruite et plusieurs voies sont ouvertes suivant qu'on aborde la question par l'une ou l'autre.



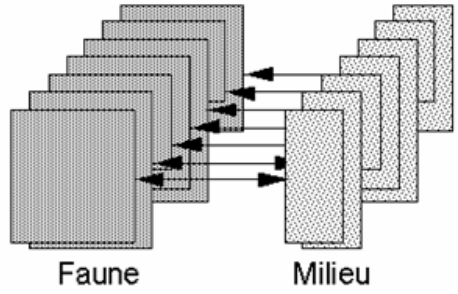
Le schéma de principe de l'*ACOM* se résume alors par :

1. K tableaux ont les mêmes lignes.
2. Ils définissent K nuages de points dans K espaces euclidiens. Les points sont pondérés de la même manière dans chaque nuage. Dans chacun des espaces, on cherche un vecteur normé (axe) sur lequel on projette le nuage.
3. On définit un code numérique de référence de variance unitaire.
4. Les axes et le code de référence optimisent la somme pondérée des carrés des covariances entre le code de référence et les coordonnées de chaque projection.
5. On recommence sous la contrainte d'orthogonalité sur les axes *et* sur les codes.

On fait donc, axe par axe, l'analyse d'inertie de chacun des tableaux en coordonnant les systèmes de coordonnées par des variables de synthèse [1]. Le premier système est directement donné par la première composante de l'*AFM*. Ensuite on est plus précis sur la géométrie séparée des nuages au prix de la perte de la représentation optimale des variables. Si vous êtes intéressé par ces notions, vous pouvez continuer par la notion de concordances (librairie `conco` de R. Lafosse), par exemple dans la question des $K + 1$ tableaux [14] :



ou celle des $2K$ tableaux [19] [18] [20] :



Références

[1] D. Chessel and M. Hanafi. Analyses de la co-inertie de k nuages de points. *Revue de Statistique Appliquée*, 44 :35–60, 1996.

[2] V. de Crespin de Billy, S. Doledec, and D. Chessel. Biplot presentation of diet composition data : an alternative for fish stomach contents analysis. *Journal of Fish Biology*, 56(4) :961–973, 2000.

[3] B. Escofier and J. Pagès. Comparaison de groupes de variables. 2ème partie : un exemple d'applications. Technical report, INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le chesnay cedex, France, 1982. Rapport de recherche n°165.

[4] B. Escofier and J. Pagès. Comparaison de groupes de variables définies sur le même ensemble d'individus. Technical report, INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le chesnay cedex, France, 1982. Rapport de recherche n°149, ISSN 0249-6399.

[5] B. Escofier and J. Pagès. L'analyse factorielle multiple : une méthode de comparaison de groupes de variables. In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone, editors, *Data Analysis and Informatics, III*, pages 41–55. Elsevier, North-Holland, 1984.

[6] B. Escofier and J. Pagès. Multiple factor analysis : results of a three-year utilization. In R. Coppi and S. Bolasco, editors, *Multiway data analysis*, pages 277–285. Elsevier Science Publishers B.V., North-Holland, 1989.

[7] Y. Escoufier. Le traitement des variables vectorielles. *Biometrics*, 29 :750–760, 1973.

[8] G.N. Foster. Evidence for ph insensitivity in some insects inhabiting peat pools in the loch fleet catchment. *Chemistry and Ecology*, 9 :207–215, 1995.

[9] J.C. Gaertner, D. Chessel, and J. Bertrand. Stability of spatial structures of demersal assemblages : a multitable approach. *Aquatic Living Resources*, 11 :75–85, 1998.

[10] M. Heo and K.R. Gabriel. A permutation test of association between configurations by means of the rv coefficient. *Communications in Statistics - Simulation and Computation*, 27 :843–856, 1998.

[11] S. Holmes. Multivariate analysis : The french way. In *Festschrift for David Freedman*. IMS Lecture Notes - Monograph Series (in press), 2006.

[12] T. Jombart, K. Moazami-Goudarzi, A.-B. Dufour, and D. Laloë. Fréquences alléliques et cohérence entre marqueurs moléculaires : des outils descriptifs. *Les Actes du BRG*, 6 :25–39, 2006.

[13] J. Kervella. Analyse de l'attrait d'un produit : exemple d'une comparaison de lots de pêches. In *2émes journées européennes Agro-Industrie et Méthodes Statistiques*, pages 103–106. Association pour la Statistique et ses Utilisations, Paris, Nantes 13-14 juin 1991, 1991.

[14] R. Lafosse and M. Hanafi. Concordance d'un tableau avec k tableaux : définition de k+1 uples synthétiques. *Revue de Statistique Appliquée*, 45 :111–126, 1997.

[15] Ch. Lavit. *Analyse conjointe de tableaux quantitatifs*. Masson, Paris, 1988.

[16] Ch. Lavit, Y. Escoufier, R. Sabatier, and P. Traissac. The act (statis method). *Computational Statistics and Data Analysis*, 18 :97–119, 1994.

[17] K. Moazami-Goudarzi and D. Laloë. Is a multivariate consensus representation of genetic relationships among populations always meaningful ? *Genetics*, 162(1) :473–484, 2002.

[18] M. Simier, L. Blanc, F. Pellegrin, and D. Nandris. Approche simultanée de k couples de tableaux : Application à l'étude des relations pathologie végétale - environnement. *Revue de Statistique Appliquée*, 47 :31–46, 1998.

[19] M. Simier, M. Hanafi, and D. Chessel. Approche simultanée de k couples de tableaux. In *Recueil des résumés des communications des XXVIIIèmes Journées de Statistique, Université laval, Québec*, pages 673–676. Québec(Canada), 1996.

[20] J. Thioulouse, M. Simier, and D. Chessel. Simultaneous analysis of a sequence of pairs of ecological tables with the statico method. *Ecology*, 85 :272–283, 2003.