Fiche TD avec le logiciel R : tdr64

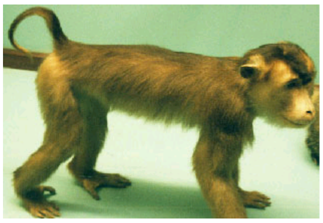
La co-structure de deux nuages de points

D. Chessel A.B. Dufour & S. Dray

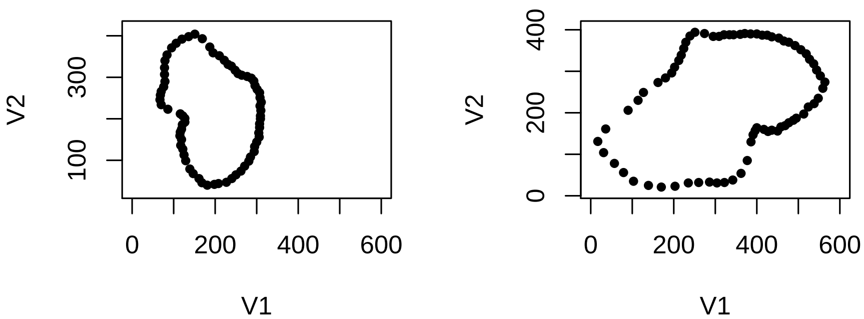
La fiche introduit à l'usage de l'analyse de co-inertie. On définit cette notion à partir des rotations procustéennes. La co-inertie désigne une classe d'analyse de couples de tableaux. Pour deux ACP on a l'analyse inter-batterie des psychométriciens (Tucker 1958). Pour deux ACM on retrouve l'analyse canonique sur variables qualitatives de Cazes (1980). L'analyse des correspondances d'un tableau de profils écologiques (Romane 1972) en fait partie. La CCA est une méthode voisine mais en diffère par les contraintes sous-jacentes.

Table des matières

1 Rotations procustéennes


[1]

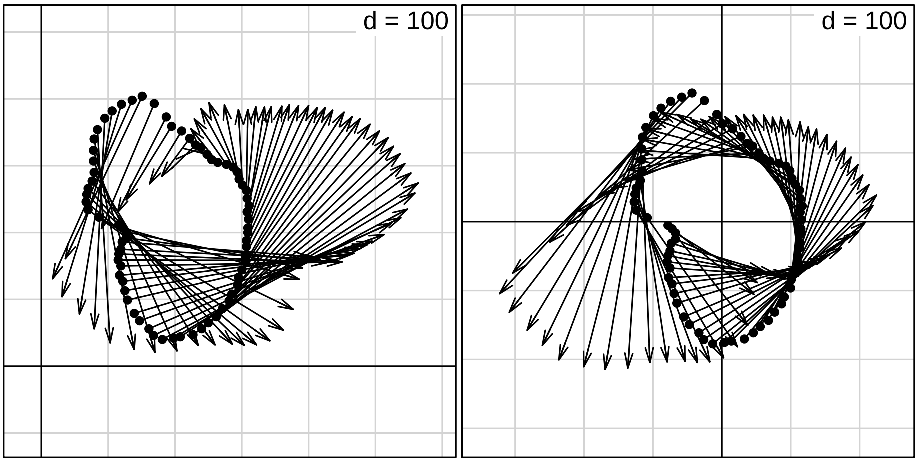
On s'intéresse (Olshan et al. 1982[20]) à la croissance céphalofaciale d'un mâle *Macaca nemestrina* étudié aux âges respectifs de 0.90 et 5.77 années. 72 points de repères fixes sont enregistrés.



Utiliser la liste `macaca` pour refaire cette figure (voir `asp` et `pch`). Le tableau xy1 (à gauche) et le tableau xy2 (à droite) contiennent 72 points définis par deux coordonnées dans un plan. C'est le même individu à deux âges différents. Le dessin de la tête a été redéfini en coordonnées polaires par rapport à un point fixe (il y a 72 points associés à 360° divisés en unités de 5°). Pour cet exercice, on a ignoré le point de référence, tourné le petit d'un angle de 90° et digitalisé la série des 72 points. L'objectif est de replacer une figure sur l'autre. Ce problème, fondamental en morphométrie, est l'objet d'une intense réflexion méthodologique (nombreuses citations et discussions dans l'article cité). On ne s'en sert ici que comme illustration, sans prétendre participer au débat. Il est idiot de superposer brutalement ces deux figures sans repère commun (ci-dessous à gauche).

```
par(mfrow = c(1, 2))
s.match(macaca$xy1, macaca$xy2, clab = 0)
s.match(scalewt(macaca$xy1, scale = F), scalewt(macaca$xy2, scale = F),
    clab = 0)
```

[1]Source : `http://members.tripod.com/uakari/macaca_nemestrina.html`



Il faut évidemment faire tourner le petit pour le recaler sur le grand (ou le grand pour le recaler sur le petit). C'est l'objectif de la rotation procustéenne (du nom d'une terreur de l'antiquité qui périt de la torture qu'il infligeait à ses victimes). Si $\mathbf{X}$ contient n points en lignes sur p coordonnées en colonnes et si $\mathbf{Y}$ contient n points en lignes sur p coordonnées en colonnes, l'accord entre les deux nuages se mesure par :

$$d^2\left(\mathbf{X},\mathbf{Y}\right) = \left\|\mathbf{X}-\mathbf{Y}\right\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij}-y_{ij}\right)^2$$

On améliore immédiatement cette mesure en donnant aux deux nuages le même centre de gravité. Centrer les deux tableaux et recommencer la superposition (ci dessus à droite).

Les deux nuages de points sont dans le même espace. $\mathbf{X}$ et $\mathbf{Y}$ sont maintenant centrés. Supposons que $\mathbf{X}$ soit la cible (le nuage en place). On veut appliquer à $\mathbf{Y}$ une rotation $\mathbf{R}$ pour optimiser l'ajustement, donc minimiser :

$$d^2\left(\mathbf{X}, {}_{\frown}\!\rightarrow \mathbf{Y}\right) = \left\|\mathbf{X}-\mathbf{Y}\mathbf{R}^T\right\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij}- {}_{\frown}\!\rightarrow y_{ij}\right)^2$$

$\mathbf{R}$ est une matrice $p\times p$ qui conserve les angles et les distances, donc les produits scalaires et vérifie $\mathbf{R}^T\mathbf{R} = \mathbf{I}_p = \mathbf{R}\mathbf{R}^T$. La mesure se réécrit (noter que $Trace\left(\mathbf{X}^T\mathbf{X}\right) = Trace\left(\mathbf{X}\mathbf{X}^T\right)$) :

$$d^2\left(\mathbf{X},\mathbf{Y}\right) = Trace\left(\left(\mathbf{X}-\mathbf{Y}\right)^T\left(\mathbf{X}-\mathbf{Y}\right)\right) = Trace\left(\mathbf{X}^T\mathbf{X}\right) + Trace\left(\mathbf{Y}^T\mathbf{Y}\right) - 2Trace\mathbf{Y}^T\mathbf{X}$$

d'où :

$$d^2\left(\mathbf{X}, {}_{\frown}\!\rightarrow \mathbf{Y}\right) = \left\|\mathbf{X}-\mathbf{Y}\mathbf{R}^T\right\|^2 = Trace\left(\mathbf{X}^T\mathbf{X}\right) + Trace\left(\mathbf{Y}^T\mathbf{Y}\right) - 2Trace\mathbf{R}\mathbf{Y}^T\mathbf{X}$$

On cherche donc $\mathbf{R}$ qui minimise :

$$\left\|\mathbf{Y}-\mathbf{R}\mathbf{X}\right\|^2 = Cte - 2Trace\left(\mathbf{R}\mathbf{Y}^T\mathbf{X}\right)$$

La décomposition en valeurs singulières de $\mathbf{Y}^T\mathbf{X}$ s'écrit :

$$\mathbf{Y}^T\mathbf{X} = \mathbf{V}\Theta\mathbf{U}^T$$

On cherche donc $\mathbf{R}$ qui maximise :

$$Trace\left(\mathbf{RV}\Theta\mathbf{U}^T\right) = \sum_{k=1}^{n} \theta_k \left(\mathbf{u}_k | \mathbf{R}\mathbf{v}_k\right) \leqslant \sum_{k=1}^{n} \theta_k$$

La borne est atteinte pour $\mathbf{R} = \mathbf{UV}^T$. Cette isométrie envoie le nuage des lignes de $\mathbf{Y}$ au plus près du nuage des lignes de $\mathbf{X}$. Pour assurer la rotation procustéenne, il suffit donc de faire la décomposition en valeurs singulières de $\mathbf{Y}^T\mathbf{X}$.

```
pro1 <- procuste(macaca$xy1, macaca$xy2, scal = F)
```

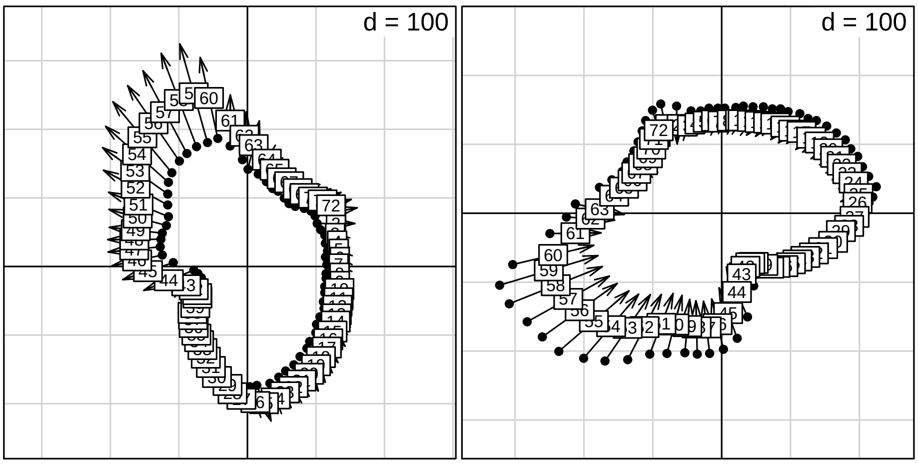
`scal=F` indique que les deux tableaux sont centrés sans changement d'échelle.

```
names(pro1)
```

```
 [1] "d"     "rank"  "nfact" "rot1"  "rot2"  "tab1"  "tab2"  "load1" "load2" "scor1"
[11] "scor2" "call"
```

Le résultat est une liste dont seul les quatre premiers éléments nous intéressent ici.

```
par(mfrow = c(1, 2))
s.match(pro1$tab1, pro1$rot2, clab = 0.7)
s.match(pro1$tab2, pro1$rot1, clab = 0.7)
```



Le petit a tourné pour se mettre dans le grand et/ou le grand a tourné pour se mettre sur le petit. Les deux figures, à une isométrie près, sont identiques.

Cette méthode a été introduite par J.C. Gower[10]. Une solution analytique avait été introduite par Sneath (1967)[28]. Ces sources sont citées dans Rohlf et Slice (1990)[21] qui décrivent cette procédure (p.42) sous le terme *Orthogonal Procustes Analysis* avec le *scaling* (voir ci-après). Si on quitte la morphométrie pour l'écologie, deux problèmes se posent immédiatement.

1.1 Problème d'échelle

Le premier touche les échelles, le second les dimensions. Pour les échelles, la situation est assez simple. La rotation procuste ne déforme pas le nuage qui est ajusté sur la cible. Les deux dessins sont strictement équivalents car aucune déformation n'est en jeu. Ajuster par rotation X à Y ou ajuster par rotation Y à X ne change rien aux objets mais ajuste la représentation de l'un à l'autre. Si les deux objets ne sont pas à la même échelle expérimentale, on sent la nécessité de renforcer l'adéquation en multipliant un tableau par une constante, ce qui ne change pas sa forme ou à multiplier certaines composantes par une constante, ce qui introduit des dilatations et un changement de formes. Il y a alors deux voies.

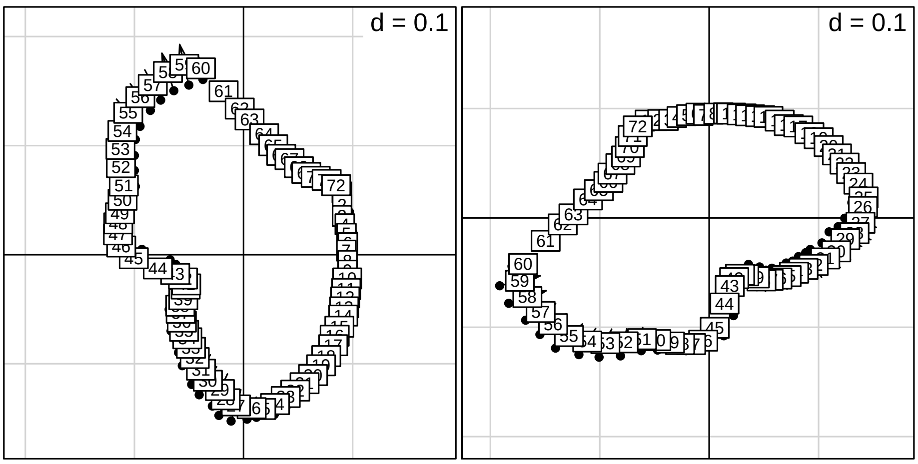
L'une est foncièrement dissymétrique en disant qu'un tableau est la cible et qu'il définit les modifications nécessaires pour ajuster l'autre. Ces propositions, issues des stratégies d'analyse factorielle, sont initiées par Schönemann et Carrol (1970)[26] qui cite déjà une *standard orthogonal Procustes subroutine* (Green 1952[12], Cliff 1966[4], Schönemann 1966[24]).Le terme *Procustes* est attribué par Schönemann [25] à Hurley and Cattel (1962)[13].

L'autre conserve une certaine symétrie au problème en mettant à une échelle commune, avant toute autre pratique, les deux tableaux. C'est le *scaling* de J.C. Gower[10] repris par Rohlf et Slice (1990)[21]. La variabilité totale des mesures est ($\mathbf{X}$ est le tableau centré) :

$$T(\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2 = n \times Iner(\mathbf{X}) = Trace\left(\mathbf{X}\mathbf{X}^T\right) = Trace\left(\mathbf{X}^T\mathbf{X}\right)$$

Les tableaux mis à une échelle commune sont alors $\frac{\mathbf{X}}{\sqrt{T(\mathbf{X})}}$ et $\frac{\mathbf{Y}}{\sqrt{T(\mathbf{Y})}}$. Pour éviter de multiplier les notations, on appellera dorénavant $\mathbf{X}$ et $\mathbf{Y}$ les tableaux centrés et mis à l'échelle, donc vérifiant $T(\mathbf{X}) = T(\mathbf{Y}) = 1$. C'est le cas d'utilisation par défaut de la fonction procuste.

```
par(mfrow = c(1, 2))
pro2 <- procuste(macaca$xy1, macaca$xy2)
s.match(pro2$tab1, pro2$rot2, clab = 0.7)
s.match(pro2$tab2, pro2$rot1, clab = 0.7)
```



La mise à l'échelle modifie sensiblement le résultat et illustre les changements de forme. En morphométrie, le *rescaling* ne s'impose pas car les changements de taille ont un sens. Mais il peut être utilisé. Par exemple dans Klingenberg et McIntyre (1998)[15] les auteurs étudient l'asymétrie des ailes de mouches tsé-tsé en superposant par rotation procustéenne une aile sur le symétrique de l'autre après *scaling to unit centroid size* pour se concentrer sur les variations de forme. En écologie, ajuster deux modèles sans mise à l'échelle n'aurait pas de sens.

1.2 Problème de dimensions

Le second problème est de loin le plus important. Ajuster deux objets à deux dimensions n'a pas besoin de plus de commentaires. L'objet de référence et celui qui lui est ajusté se voient intégralement dans le plan. Par contre, si le premier est dans $\mathbb{R}^p$ et le second est dans $\mathbb{R}^q$ avec p et q grands et différents, il faut une méthode pour voir le résultat, d'une part, la question des rotations est plus difficile, d'autre part.

A priori les deux nuages de points ne sont pas dans le même espace. On peut, d'un point de vue formel, compléter le plus petit des tableaux par des colonnes de valeurs nulles pour lui donner la dimension du plus grand. La décomposition en valeurs singulières définit deux bases orthonormées du même espace de dimension $\max(p, q)$:

$$\mathbf{Y}^T\mathbf{X} = \mathbf{V}\Theta\mathbf{U}^T$$

$\mathbf{X}_{rot} = \mathbf{X}\mathbf{U}\mathbf{V}^T$ contient le nuage issu de $\mathbf{X}$ qui s'ajuste à $\mathbf{Y}$ et $\mathbf{Y}_{rot} = \mathbf{Y}\mathbf{V}\mathbf{U}^T$ contient le nuage issu de $\mathbf{Y}$ qui s'ajuste à $\mathbf{X}$. La qualité de l'ajustement s'écrit :

$$\|\mathbf{Y} - \mathbf{X}\mathbf{R}\|^2 = 2 - 2Trace\left(\mathbf{R}\mathbf{Y}^T\mathbf{X}\right) = 2 - 2\sum_{k=1}^{r}\theta_k$$

où r est le rang de $\mathbf{Y}^T\mathbf{X}$. Ajouter des zéros dans un tableau ne change rien puisqu'on ajoute des valeurs singulières nulles. Pour voir l'ajustement, il faut donc projeter sur un plan les lignes de $\mathbf{X}$ et de $\mathbf{Y}_{rot}$ ou celles de $\mathbf{Y}$ et de $\mathbf{X}_{rot}$. On obtient des approximations procustéennes. Il y a de nombreuses possibilités, en particulier les deux ACP des tableaux accolés et les deux ACP des tableaux moyens (par exemple Mouttet 1981[19]) :

$$\begin{bmatrix} \mathbf{X}_{rot} \\ \mathbf{Y} \end{bmatrix} \quad \begin{bmatrix} \mathbf{X} \\ \mathbf{Y}_{rot} \end{bmatrix} \quad \frac{1}{2}\left(\mathbf{X}_{rot} + \mathbf{Y}\right) \quad \frac{1}{2}\left(\mathbf{X} + \mathbf{Y}_{rot}\right)$$

La question est clairement explicitée dans l'article fondamental de Sibson (1978)[27] :

> It might be thought that (by analogy with the centroid matching used to fit X and it might be thought that (by analogy with the centroid matching used to fit X and optimally under translation) rotation/reflexion fit would have something to do with matching principal axes. This is simply not the case ; all that can be said is that if X and Y can be well matched, and if the principal variances are well distinguished, then the principal axes will themselves correspond reasonably closely after fitting under rotation/reflexion. (p. 237)

Toutes ces analyses donnent des représentations voisines mais différentes. Or les rotations ne déformant pas les nuages, les configurations de $2n$ points (les n

points d'un tableau et les n points de l'autre après rotation) sont strictement les mêmes. La représentation en dimensions réduites devrait être unique. Pour résoudre ce problème, l'option utilise le résultat de l'analyse de co-inertie de deux tableaux totalement appariés[33].

Considérons deux tableaux $\mathbf{A}$ et $\mathbf{B}$ portant sur les mêmes n lignes et les mêmes s colonnes. Ils donnent deux nuages de n points dans $\mathbb{R}^s$. Si on cherche un axe d'ACP $\mathbf{z}$ *commun* à ces deux tableaux, il doit avoir une propriété d'axe principal pour $\mathbf{A}$ ($Var(\mathbf{Az})$ grand), une propriété d'axe principal pour $\mathbf{B}$ ($Var(\mathbf{Bz})$ grand), et une propriété de cohérence entre les deux systèmes de coordonnées, donc une propriété d'analyse canonique ($corr^2(\mathbf{Az}, \mathbf{Bz})$ grand). On cherche à maximiser :

$$f(\mathbf{z}) = \langle \mathbf{Az} \mid \mathbf{Bz} \rangle = Cte \times \sqrt{Var(\mathbf{Az})} \times \sqrt{Var(\mathbf{Bz})} \times corr(\mathbf{Az}, \mathbf{Bz})$$

Le maximum est atteint pour le premier vecteur propre de la matrice :

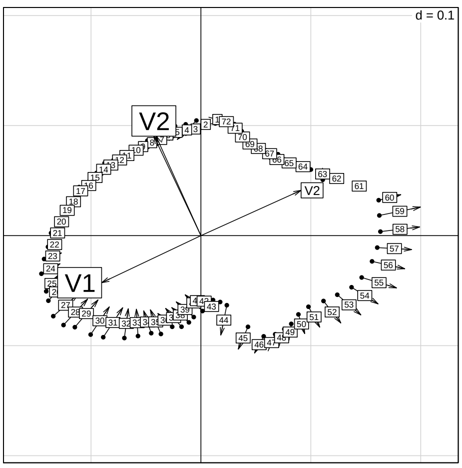
$$\mathbf{W} = \frac{1}{2}\left(\mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{A}\right)$$

Les axes successifs de co-inertie sont les vecteurs propres normés de $\mathbf{W}$. Pour le couple $\mathbf{X}_{rot} = \mathbf{XUV}^T$ et $\mathbf{Y}$ on a :

$$\begin{aligned}\mathbf{W} = \tfrac{1}{2}\left(\mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{A}\right) &= \tfrac{1}{2}\left(\mathbf{VU}^T\mathbf{X}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{XUV}^T\right) \\ &= \tfrac{1}{2}\left(\mathbf{VU}^T\mathbf{U}\Theta\mathbf{V}^T + \mathbf{V}\Theta\mathbf{U}^T\mathbf{UV}^T\right) = \mathbf{V}\Theta\mathbf{V}^T\end{aligned}$$

Il s'en suit immédiatement que la co-inertie entre $\mathbf{X}_{rot} = \mathbf{XUV}^T$ et $\mathbf{Y}$, d'une part, et entre $\mathbf{Y}_{rot} = \mathbf{YVU}^T$ et $\mathbf{X}$ d'autre part, conduit à la même représentation et à la même qualité de représentation. Les deux systèmes de coordonnées sont $\mathbf{XU}$ et $\mathbf{YV}$ et les critères associés sont les valeurs singulières de $\mathbf{Y}^T\mathbf{X}$ (ou $\mathbf{X}^T\mathbf{Y}$) c'est-à-dire les racines des valeurs propres de $\mathbf{Y}^T\mathbf{XX}^T\mathbf{Y}$ (ou $\mathbf{YY}^T\mathbf{XX}^T$ ou $\mathbf{X}^T\mathbf{YY}^T\mathbf{X}$ ou $\mathbf{XX}^T\mathbf{YY}^T$). Les composantes `scor1` et `scor2` contiennent les matrices $\mathbf{XU}$ et $\mathbf{YV}$. Les composantes `load1` et `load2` contiennent les matrices $\mathbf{U}$ et $\mathbf{V}$ vues comme projections des bases canoniques dans deux ACP simultanées.

```
s.match(pro2$scor1, pro2$scor2, clab = 0.7)
s.arrow(0.1 * pro2$load1, add.plo = T)
s.arrow(0.1 * pro2$load2, add.plo = T, clab = 2)
```



Le résultat est une représentation des deux figures ramenées à la même échelle avec l'indication des rotations utilisées par les deux bases canoniques (de longueur 0.1 pour tenir dans la figure). Évidemment l'une avait été tournée de 90° mais maintenant les deux ont tourné pour mettre la plus grande variabilité sur l'axe des x.

2 Inertie et co-inertie

Soit un tableau $\mathbf{X}$ avec n lignes et p colonnes. Les n lignes de $\mathbf{X}$, notées $\mathbf{X}_i$ sont des vecteurs de $\mathbb{R}^p$qui forment un nuage de n points. Si $\mathbb{R}^p$est muni d'un produit scalaire $\mathbf{Q}$ et si chaque point est muni d'un poids w_i ($\mathbf{D} = diag\left(w_1, ..., w_n\right)$), le nuage a une inertie autour de l'origine :

$$I_\mathbf{0} = \sum_{i=1}^{n} w_i \left\| \mathbf{X}_i \right\|_\mathbf{Q}^\mathbf{2} = trace\left(\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}\right)$$

Pour un vecteur **u, Q**-normé, quelconque de $\mathbb{R}^p$, le nuage des projections des points $\mathbf{X}_i$ sur $\mathbf{u}$ est l'inertie projetée sur **u :**

$$I\left(\mathbf{u}\right) = \mathbf{u}^T\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}$$

Dans la base des axes principaux de l'analyse du triplet $\left(\mathbf{X}, \mathbf{Q}, \mathbf{D}\right)$ on a la décomposition :

$$I_\mathbf{0} = \sum_{k=1}^{r} I\left(\mathbf{u}_k\right)$$

Ceci est la base de toutes les analyses linéaires élémentaires. On a d'ailleurs pour une base $\mathbf{Q}$-orthonormale quelconque de $\mathbb{R}^p$:

$$I_\mathbf{0} = \sum_{k=1}^{p} I\left(\mathbf{v}_k\right)$$

Lorsque le nuage est centré, ce qui est le cas le plus général l'inertie est une somme de variances. Pour une base orthonormale arbitraire dans $\mathbb{R}^p$ $\left\{\mathbf{v}_1, ..., \mathbf{v}_p\right\}$:

$$I_\mathbf{0} = \sum_{k=1}^{p} I\left(\mathbf{v}_k\right) = \sum_{k=1}^{p} \mathbf{v}_k^T\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{v}_k = \sum_{k=1}^{p} \left\| \mathbf{X}\mathbf{Q}\mathbf{v}_k \right\|_\mathbf{D}^\mathbf{2}$$

Soit alors un second tableau $\mathbf{Y}$ avec n lignes et q colonnes. Les n lignes sont des vecteurs de $\mathbb{R}^q$qui forment un nuage de n points. Si $\mathbb{R}^q$est muni d'un produit scalaire $\mathbf{R}$ et si chaque point est muni du **même poids** w_i ($\mathbf{D} = diag\left(w_1, ..., w_n\right)$), le nuage a une inertie autour de l'origine :

$$J_\mathbf{0} = trace\left(\mathbf{Y}\mathbf{R}\mathbf{Y}^T\mathbf{D}\right)$$

qui peut se décomposer comme précédemment. Il est plus difficile de parler de la géométrie conjointe des deux nuages. Les rotations procustéennes le font au prix de l'élimination de la question des poids et des métriques, tout se passant, soit explicitement, soit implicitement, dans un même espace muni de la métrique canonique. Si on veut étendre la notion d'inertie d'un nuage à celle de co-inertie de deux nuages, on pense naturellement à dire que, puisque la première est une somme de variances, la seconde pourrait être une somme de covariances. La remarque qui suit est très dissuasive. Supposons que $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$soit une base $\mathbf{Q}$-orthonormée de $\mathbb{R}^p$ et que $\{\mathbf{v}_1, ..., \mathbf{v}_q\}$ soit une base $\mathbf{R}$-orthonormée de $\mathbb{R}^q$. $\mathbf{XQu}_k$ contient les coordonnées du premier nuage sur le vecteur de rang k et $\mathbf{YRv}_j$ contient les coordonnées du second nuage sur le vecteur de rang j. Si on somme les covariances sur tous les couples formés d'un vecteur dans le premier espace et d'un vecteur dans le second espace :

$$S = \sum_{k=1}^{p} \sum_{j=1}^{q} \mathbf{u}_k^T \mathbf{QX}^T \mathbf{DYRv}_j = \sum_{k=1}^{p} \sum_{j=1}^{q} \mathbf{X}^{k^T} \mathbf{DY}^j$$

on trouve un invariant indépendant des bases, ce qui est souhaitable, et des métriques ce qui l'est moins. La covariance n'est pas l'extension de la variance à un couple, simplement parce que la covariance de $\mathbf{x}$ et de -$\mathbf{x}$ vaut l'opposé de la variance, alors que *d'un point de vue typologique* $\mathbf{x}$ et -$\mathbf{x}$ ont strictement la même fonction. D'ailleurs on ne s'étonne pas de voir un plan factoriel dans un sens ou dans l'autre parce qu'on sait que le tirage est aléatoire. On peut alors définir la co-inertie des deux nuages par la somme des carrés des covariances de tous les couples de coordonnées :

$$S = \sum_{k=1}^{p} \sum_{j=1}^{q} \left(\mathbf{u}_k^T \mathbf{QX}^T \mathbf{DYRv}_j \right)^2 = \sum_{k=1}^{p} \sum_{j=1}^{q} \left(\mathbf{X}^{k^T} \mathbf{DY}^j \right)^2$$

On a :

$$\begin{aligned} S &= trace \left(\mathbf{U}^T \mathbf{QX}^T \mathbf{DYRVV}^T \mathbf{RY}^T \mathbf{DXQU} \right) \\ &= trace \left(\mathbf{XQUU}^T \mathbf{QX}^T \mathbf{DYRVV}^T \mathbf{RY}^T \mathbf{D} \right) \\ &= trace \left(\mathbf{XQX}^T \mathbf{DYRY}^T \mathbf{D} \right) \qquad (1) \end{aligned}$$

La quantité est invariante par un double changement de base.

L'analyse de co-inertie repose alors sur le schéma :

$$\begin{tikzcd} \boxed{p} \arrow[r, "\mathbf{Q}"] & \boxed{p} \arrow[d, "\mathbf{X}"] \\ \boxed{n} \arrow[u, "\mathbf{X}^T"] & \boxed{n} \arrow[l, "\mathbf{D}"] \end{tikzcd} \qquad \begin{tikzcd} \boxed{q} \arrow[r, "\mathbf{R}"] & \boxed{q} \arrow[d, "\mathbf{Y}"] \\ \boxed{n} \arrow[u, "\mathbf{Y}^T"] & \boxed{n} \arrow[l, "\mathbf{D}"] \end{tikzcd} \quad \Rightarrow \quad \begin{tikzcd} \boxed{p} \arrow[r, "\mathbf{Q}"] & \boxed{p} \arrow[d, "\mathbf{Y}^T\mathbf{DX}"] \\ \boxed{q} \arrow[u, "\mathbf{X}^T\mathbf{DY}"] & \boxed{q} \arrow[l, "\mathbf{R}"] \end{tikzcd}$$

3 Décomposition de la co-inertie

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. La co-inertie de deux ACP étend cette propriété partiellement. En effet, la diagonalisation du schéma garantit que, si r est le rang de la matrice $\mathbf{Y}^T\mathbf{DX}$, on dispose d'un système $\mathbf{Q}$-orthonormé $\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ de

$\mathbb{R}^p$, d'un système $\{\mathbf{v}_1, ..., \mathbf{v}_r\}$ $\mathbf{R}$-orthonormé de $\mathbb{R}^q$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 :

$$S = trace\left(\mathbf{XQX}^T\mathbf{DYRY}^T\mathbf{D}\right) = \sum_{k=1}^{r} \omega_k$$

Comme $\mathbf{Y}^T\mathbf{DXQU}\Omega^{-1/2} = \mathbf{V}$ et comme les systèmes de vecteurs sont orthonormés dans les deux espaces :

$$k \neq k' \Rightarrow \mathbf{u}_k^T\mathbf{QX}^T\mathbf{DYRv}_{k'} = 0$$

$$S = \sum_{k=1}^{p}\sum_{j=1}^{q}\left(\mathbf{X}^{k^T}\mathbf{DY}^j\right)^2 = \sum_{k=1}^{r}\mathbf{u}_k^T\mathbf{QX}^T\mathbf{DYRv}_k = \sum_{k=1}^{r}\omega_k$$

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.

Illustrer cette propriété en utilisant deux paquets de deux variables de corrélations connues.

```
 sd <- matrix(0.7, 4, 4)
 diag(sd) <- 1
 sd
     [,1] [,2] [,3] [,4]
[1,]  1.0  0.7  0.7  0.7
[2,]  0.7  1.0  0.7  0.7
[3,]  0.7  0.7  1.0  0.7
[4,]  0.7  0.7  0.7  1.0
 library(MASS)
 w <- mvrnorm(20, mu = rep(0, 4), Sigma = sd)
 w1 <- data.frame(w[, 1:2])
 w2 <- data.frame(w[, 3:4])
 pca1 <- dudi.pca(w1, scann = F, scal = F)
 pca2 <- dudi.pca(w2, scann = F, scal = F)
 zapsmall(cor(pca1$li))
      Axis1 Axis2
Axis1     1     0
Axis2     0     1
 zapsmall(cor(pca2$li))
      Axis1 Axis2
Axis1     1     0
Axis2     0     1
 coi <- coinertia(pca1, pca2, scan = F)
 summary(coi)
Eigenvalues decomposition:
           eig       covar      sdX       sdY       corr
1 1.754815e+00 1.324694451 1.1787801 1.2486329 0.90001163
2 1.183033e-05 0.003439525 0.4686636 0.5972808 0.01228736
Inertia & coinertia X:
    inertia      max     ratio
1  1.389523 1.391662 0.9984629
12 1.609168 1.609168 1.0000000

Inertia & coinertia Y:
    inertia      max    ratio
1  1.559084 1.559106 0.999986
12 1.915829 1.915829 1.000000

RV:
 0.7789408
```

```
 cor(coi$lX)
            AxcX1       AxcX2
AxcX1  1.00000000 -0.09063422
AxcX2 -0.09063422  1.00000000
 cor(coi$lY)
             AxcY1        AxcY2
AxcY1  1.000000000 -0.006890195
AxcY2 -0.006890195  1.000000000
 zapsmall(cor(coi$lX, coi$lY))
          AxcY1     AxcY2
AxcX1 0.9000116 0.0000000
AxcX2 0.0000000 0.0122874
 x <- as.matrix(pca1$tab)
 y <- as.matrix(pca2$tab)
 t(x) %*% y/20
          X1        X2
X1 0.4107683 0.6782491
X2 0.5447958 0.9106439
 sum(t(x) %*% y/20^2)
[1] 0.1272229
 sum((t(x) %*% y/20)^2)
[1] 1.754827
 x <- as.matrix(coi$lX)
 y <- as.matrix(coi$lY)
 zapsmall(t(x) %*% y/20)
         AxcY1     AxcY2
AxcX1 1.324695 0.0000000
AxcX2 0.000000 0.0034395
 sum((t(x) %*% y/20)^2)
[1] 1.754827
 sum(coi$eig)
[1] 1.754827
```

Le schéma d'une analyse de co-inertie est d'abord celui d'une analyse d'inertie :

$$\begin{tikzcd} \boxed{p} \arrow[r, "\mathbf{Q}"] & \boxed{p} \arrow[d, "\mathbf{Y}^T\mathbf{DX}"] \\ \boxed{q} \arrow[u, "\mathbf{X}^T\mathbf{DY}"] & \boxed{q} \arrow[l, "\mathbf{R}"] \end{tikzcd}$$

Mais ses produits s'interprète de multiple manière en fonction de la nature du tableau croisé $\mathbf{X}^T\mathbf{DY}$. Un tableau croisé peut, en effet suivant les cas avoir le statut de tableau de fréquences, de moyenne, de covariances ou de corrélations.

```
 coi
Coinertia analysis
call: coinertia(dudiX = pca1, dudiY = pca2, scannf = F)
class: coinertia dudi
$rank (rank)     : 2
$nf (axis saved) : 2
$RV (RV coeff)   : 0.7789408

eigen values: 1.755 1.183e-05

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

   data.frame nrow ncol content
1  $tab       2    2    crossed array (CA)
```

```
2  $li        2    2    Y col = CA row: coordinates
3  $l1        2    2    Y col = CA row: normed scores
4  $co        2    2    X col = CA column: coordinates
5  $c1        2    2    X col = CA column: normed scores
6  $lX        20   2    row coordinates (X)
7  $mX        20   2    normed row scores (X)
8  $lY        20   2    row coordinates (Y)
9  $mY        20   2    normed row scores (Y)
10 $aX        2    2    axis onto co-inertia axis (X)
11 $aY        2    2    axis onto co-inertia axis (Y)
```

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. Les composantes dont il est utile de connaître le contenu sont :

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

cw la métrique diagonale de $\mathbb{R}^p$(poids des colonnes de$\mathbf{X}$)

lw la métrique diagonale de $\mathbb{R}^q$(poids des colonnes de $\mathbf{Y}$)

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

c1 les axes de co-inertie dans $\mathbb{R}^p$, vecteurs normés en colonnes

l1 les xes de co-inertie dans $\mathbb{R}^q$, vecteurs normés en colonnes

co les produits scalaires entre colonnes de $\mathbf{X}$et coordonnées de co-inertie dans $\mathbb{R}^q$

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

lX les coordonnées de co-inertie dans $\mathbb{R}^p$donnant les projections des lignes de $\mathbf{X}$sur les axes de co-inertie dans $\mathbb{R}^p$

lY les coordonnées de co-inertie dans $\mathbb{R}^q$donnant les projections des lignes de $\mathbf{Y}$sur les axes de co-inertie dans $\mathbb{R}^q$

mX les scores normés obtenus en normant dans $\mathbb{R}^n$les coordonnées de lX

mY les scores normés obtenus en normant dans $\mathbb{R}^n$les coordonnées de lY

aX 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 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 $\mathbb{R}^q$

4 Couple d'Analyses en Composantes Principales

C'est la cas le plus simple. *Sans son interprétation géométrique*, les calculs ont été décrits dans l'analyse inter-batteries de Tucker ....(1958). Utiliser la liste `fruits` décrite à :

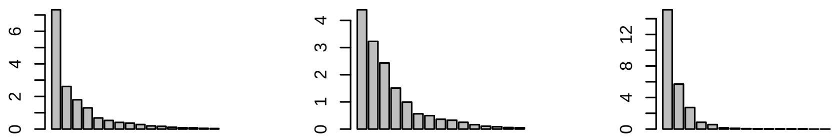
`http://pbil.univ-lyon1.fr/R/pps/pps081.pdf`

```
data(fruits)
names(fruits)
[1] "type" "jug"  "var"
```

```
pca1 <- dudi.pca(fruits$jug, scal = T, scan = F, nf = 4)
pca2 <- dudi.pca(fruits$var, scal = T, scan = F, nf = 4)
coif <- coinertia(pca1, pca2, scannf = F, nf = 3)
summary(coif)
```

```
Eigenvalues decomposition:
        eig    covar      sdX      sdY      corr
1 15.133835 3.890223 2.607581 1.864335 0.8002263
2  5.703734 2.388249 1.550666 1.776134 0.8671329
3  2.728231 1.651736 1.471356 1.433355 0.7831937
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

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

RV:
 0.4927474
```

```
par(mfrow = c(1, 3))
barplot(pca1$eig)
barplot(pca2$eig)
barplot(coif$eig)
```



Le coefficient RV qui est affiché est un véritable *coefficient de corrélation entre les deux typologies*. La co-inertie totale est en effet un produit scalaire entre opérateurs **D**-symétriques :

$$S = trace\left(\mathbf{XQX}^T\mathbf{DYRY}^T\mathbf{D}\right) = trace\left(\mathbf{W_X D W_Y D}\right) = \sum_{k=1}^{r}\omega_k$$

Le cosinus associé est :

$$\begin{aligned} RV &= \frac{trace\left(\mathbf{W_X D W_Y D}\right)}{\sqrt{trace\left(\left(\mathbf{W_X D}\right)^2\right)}\sqrt{trace\left(\left(\mathbf{W_Y D}\right)^2\right)}} \\ &= \frac{\sum_{k=1}^{r}\omega_k}{\sqrt{\sum_{i=1}^{r_\mathbf{X}}\lambda_i^2\left(\mathbf{X},\mathbf{Q},\mathbf{D}\right)}\sqrt{\sum_{j=1}^{r_\mathbf{Y}}\lambda_j^2\left(\mathbf{Y},\mathbf{R},\mathbf{D}\right)}} \end{aligned}$$

On peut écrire simplement :

$$RV = \frac{co\text{-}inertia\left(\mathbf{X},\mathbf{Y}\right)}{\sqrt{co\text{-}inertia\left(\mathbf{X},\mathbf{X}\right)}\sqrt{co\text{-}inertia\left(\mathbf{Y},\mathbf{Y}\right)}}$$

Il est compris entre 0 et 1. Sa définition vient de Escoufier (1973)[8]. 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 :

```
sum(cor(pca1$tab, pca2$tab)^2)/sqrt(sum(cor(pca1$tab, pca1$tab)^2) *
    sum(cor(pca2$tab, pca2$tab)^2))
```

```
[1] 0.4927474
```

L'édition de base, sous le titre *Eigenvalues decomposition* décompose les valeurs propres. Elle est adaptée aux cas de deux nuages centrés (le plus général). La valeur propre de co-inertie de rang k est un carré de covariance :

$$\begin{aligned}\omega_k &= \left(\mathbf{u}_k^T\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{Y}\mathbf{R}\mathbf{v}_k\right)^2 = cov^2\left(\mathbf{X}\mathbf{Q}\mathbf{u}_k, \mathbf{Y}\mathbf{R}\mathbf{v}_k\right) \\ \sqrt{\omega_k} &= corr\left(\mathbf{X}\mathbf{Q}\mathbf{u}_k, \mathbf{Y}\mathbf{R}\mathbf{v}_k\right) sdev\left(\mathbf{X}\mathbf{Q}\mathbf{u}_k\right) sdev\left(\mathbf{Y}\mathbf{R}\mathbf{v}_k\right)\end{aligned}$$

La première valeur propre vaut 15.134 ce qui correspond à une covariance de 3.890 qui est le produit d'une corrélation de 0.8002 par les deux écarts-types. Les axes de co-inertie ont une fonction de coordination de deux ACP lue par la partie corrélation et par la partie inertie projetée en-dessous. On projette sur le premier axe de co-inertie de $\mathbf{X}$ une inertie de 6.799, soit 92.9% de l'optimum 7.319 défini par l'inertie projetée sur le premier axe de l'ACP initiale. On retrouve cette quantité :

```
cumsum(redo.dudi(pca1)$eig[1:3])
```

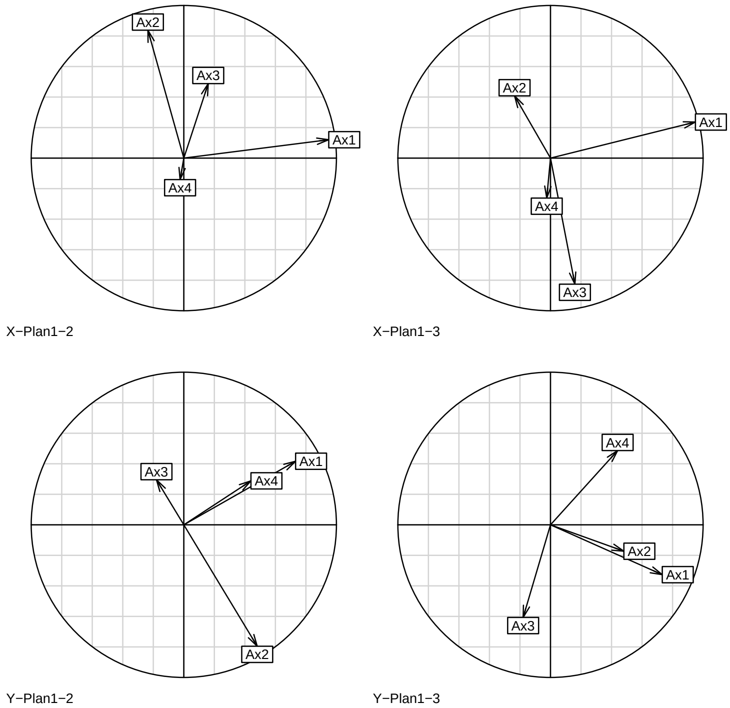
```
[1]  7.318882  9.930650 11.727775
```

```
cumsum(redo.dudi(pca2)$eig[1:3])
```

```
[1]  4.391663  7.620306 10.053072
```

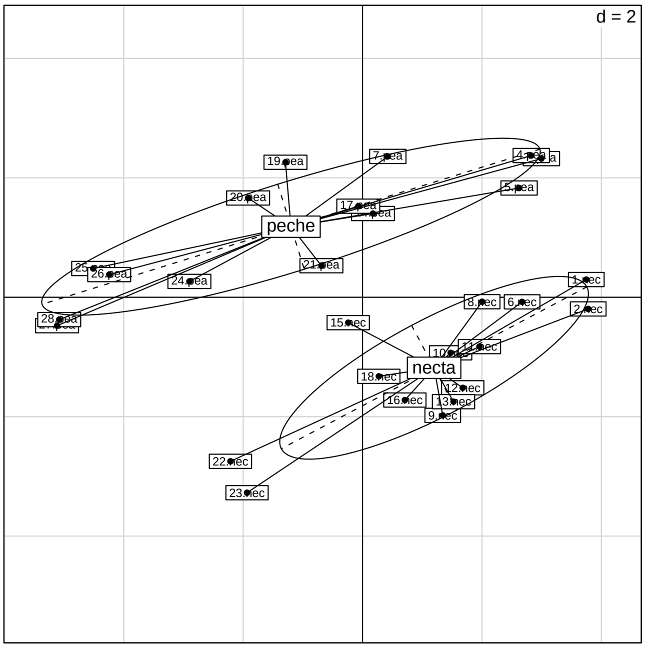
On compare ensuite le premier plan de co-inertie et le premier plan d'inertie, les premier sous-espaces de dimension 3, ... conformément à la théorie des axes principaux. Il est bon de replacer les deux systèmes d'axes :

```
par(mfrow = c(2, 2))
s.corcircle(coif$aX, sub = "X-Plan1-2")
s.corcircle(coif$aX, 1, 3, sub = "X-Plan1-3")
s.corcircle(coif$aY, sub = "Y-Plan1-2")
s.corcircle(coif$aY, 1, 3, sub = "Y-Plan1-3")
```



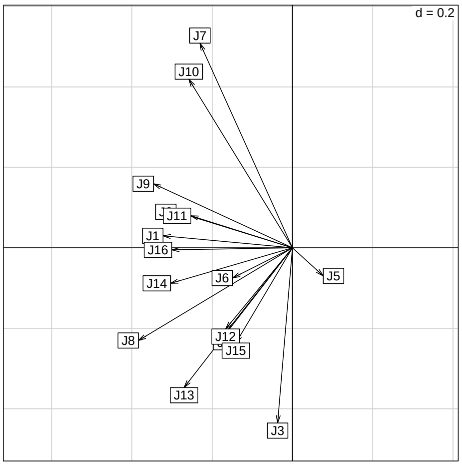
On voit ainsi comment les plans de co-inertie du premier tableau sont à une rotation près d'angle petit les plans d'inertie. Pour le second, la coordination a apporté des changements notables, en particulier une symétrie-rotation importante sur le plan 1-2 et un axe 3 "en travers" dans le champs des trois premiers axes principaux. On se contentera de discuter du premier plan.

```
s.label(coif$lX, clab = 0.7)
s.class(coif$lX, fruits$type, add.p = T)
```



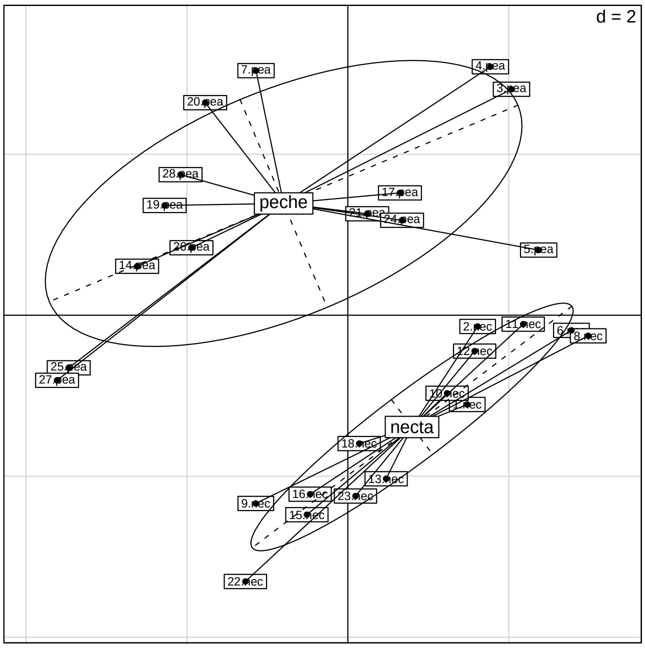
Observer que le plan d'ACP a légèrement "tourné".

```
s.arrow(coif$c1)
```



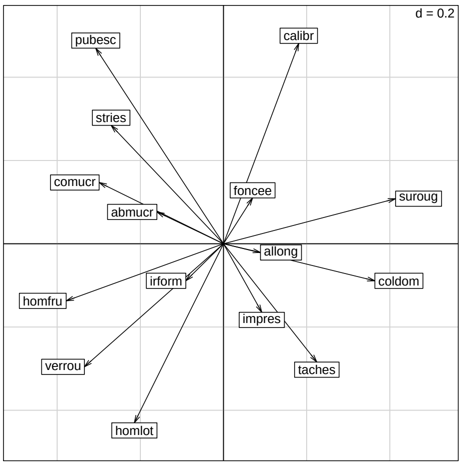
Interpréter en pensant qu'il s'agit de rangs (la préférence s'exprime par le rang 1, valeur faible, l'exclusion s'exprime par le rang 28, valeur forte). L'axe 1 exprime une opinion majoritaire, l'axe 2 une différence de choix entre juges. La séparation des types de fruits n'est pas nulle sur le premier et très forte sur le second. Trouver des illustrations dans les données.

```
s.label(coif$lY, clab = 0.7)
s.class(coif$lY, fruits$type, add.p = T)
```



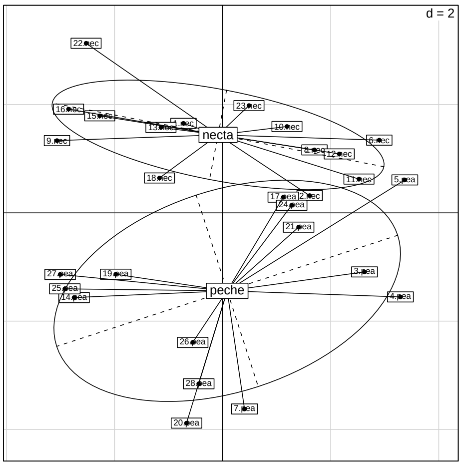
La structure est conservée alors qu'on utilise une ensemble de variables totalement différents.

```
s.arrow(coif$l1)
```



On a fait deux ACP simultanées. La seconde est cependant sensiblement perturbée pour se recadrer sur l'autre :

```
s.label(pca2$li, clab = 0.7)
s.class(pca2$li, fruits$type, add.p = T)
```



Pour approfondir l'interprétation, utiliser la signification des variables et la solution de :

`http://pbil.univ-lyon1.fr/R/exos/exo9.pdf`

5 Co-inertie et rotations procustéennes

La rotation procustéenne entretient des relations étroites avec l'analyse de co-inertie. Prendre le jeu de données `doubs`[36].

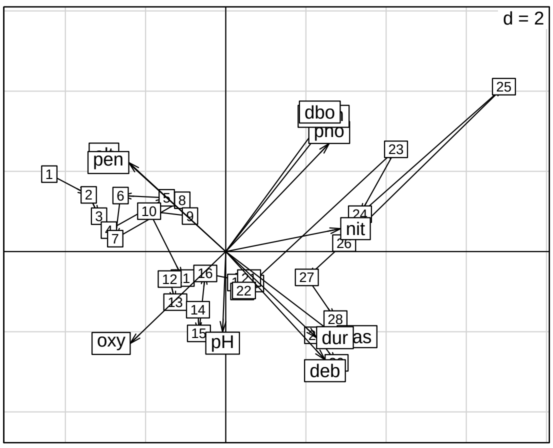
```
data(doubs)
names(doubs)
[1] "mil" "poi" "xy"
```

Faire l'ACP normée du tableau doubs$mil :

```
pca1 <- dudi.pca(doubs$mil, scal = T, scannf = F)
```

Un nuage de 30 points de $\mathbb{R}^{11}$ est projeté sur ses axes principaux et comme les stations sont dans l'ordre naturel sur la rivière on peut résumer par :

```
s.traject(pca1$li, clab = 0)
s.label(pca1$li, add.pl = T, clab = 0.75)
s.arrow(8 * pca1$c1, clab = 1, add.p = T)
```

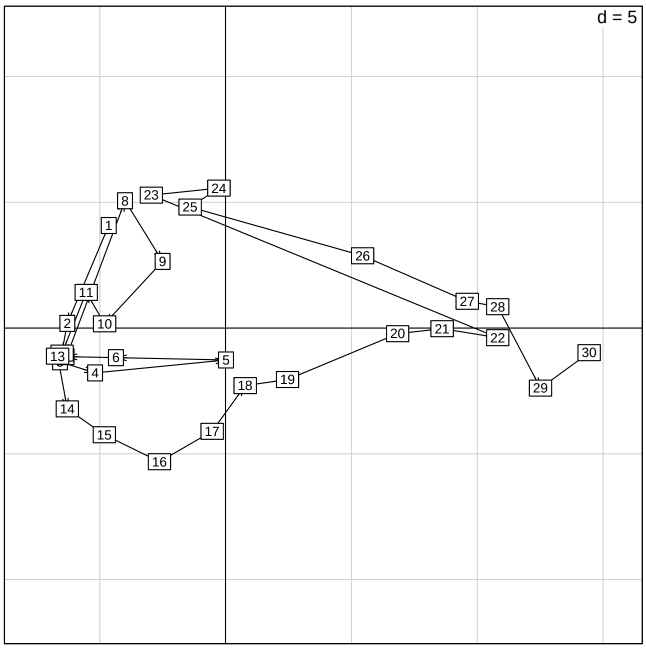


On a un gradient amont-aval (pente et altitude décroissantes, dureté, distance à la source et débit croissants) et on a des pollutions locales en particulier 23-25 (oxygène décroissant et charge organique croissante). L'augmentation simultanée en aval de la charge minérale et de la charge organique définit l'axe 1 comme compromis aval+pollution. Faire l'acp centrée du tableau faunistique :

```
pca2 <- dudi.pca(doubs$poi, scal = F, scannf = F)
```

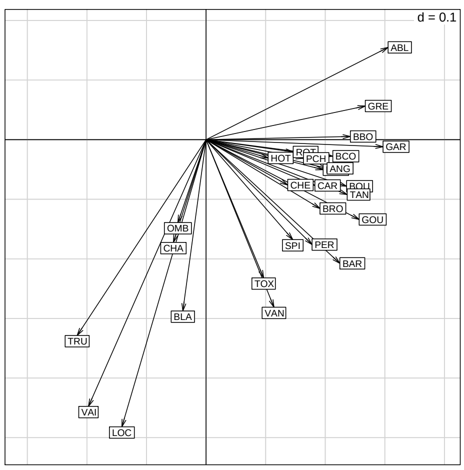
Un nuage de 30 points de $\mathbb{R}^{27}$ est projeté sur ses axes principaux et comme les stations sont dans l'ordre naturel sur la rivière, on peut résumer par :

```
s.traject(pca2$li, clab = 0)
s.label(pca2$li, add.pl = T, clab = 0.75)
```



et

```
s.arrow(pca2$c1, clab = 0.75)
```



On a un nouveau gradient amont-aval de la zone à salmonidés (Truite+Vairon+Loche, puis Ombre+Chabot+Blageon) à la zone à cyprinidés (de Toxostome+Vandoise à Gardon+Ablette). La zone polluée (23-25) rejoint des points en amont caractérisés par une grande pauvreté faunistique (pollution donc élimination). On voudrait recaler les deux analyses dans un cadre commun.

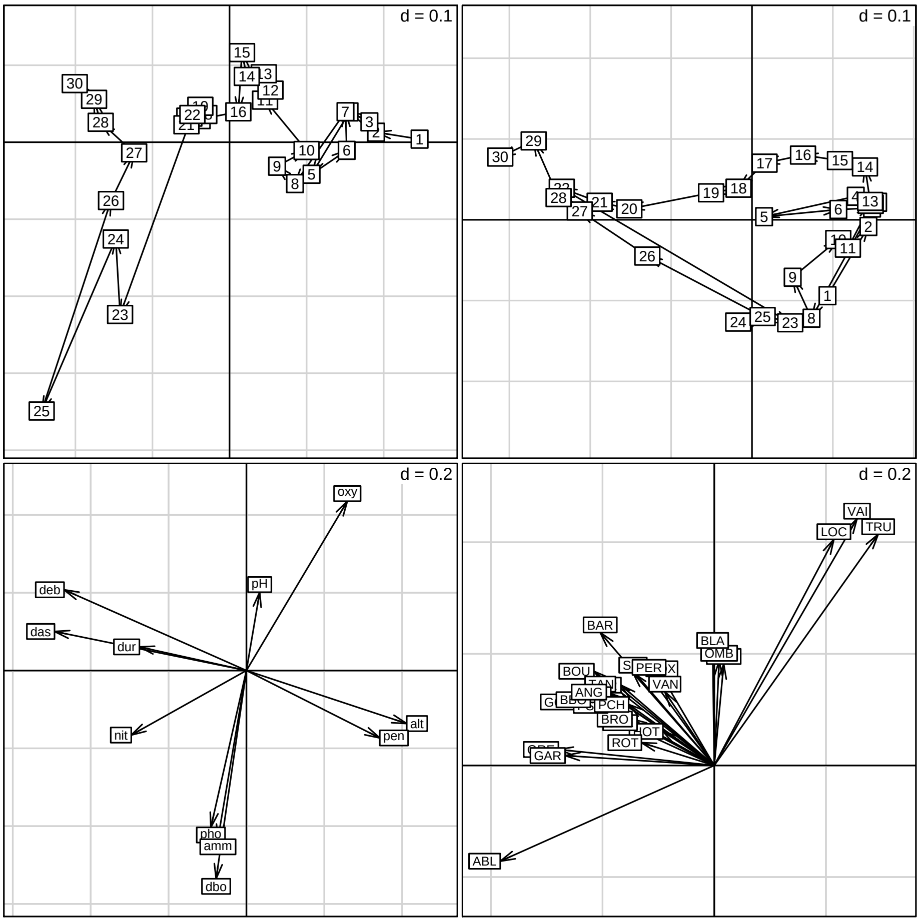
Exécuter et dépouiller l'analyse de co-inertie :

```
coiner1 <- coinertia(pca1, pca2, scannf = F)
summary(coiner1)
```

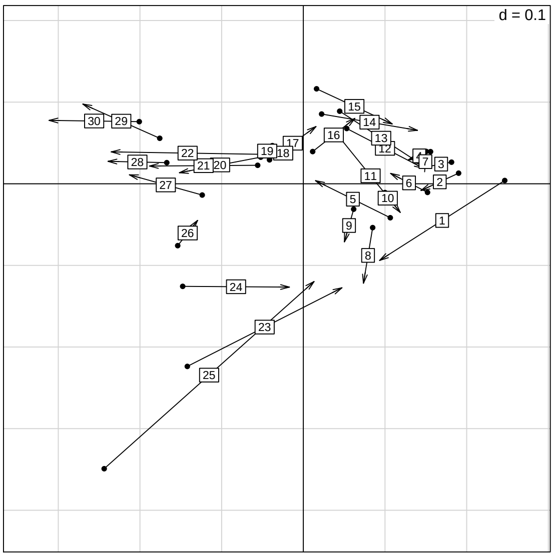
```
Eigenvalues decomposition:
        eig     covar      sdX      sdY      corr
1 119.01942 10.909602 2.326324 6.422570 0.7301798
2  13.87137  3.724429 1.685078 2.863743 0.7718017
Inertia & coinertia X:
    inertia      max     ratio
1  5.411785 6.321624 0.8560752
12 8.251272 8.553220 0.9646978

Inertia & coinertia Y:
    inertia      max     ratio
1  41.24940 42.74627 0.9649824
12 49.45042 50.90461 0.9714331

RV:
 0.4505569
```

Un tableau **X** a n lignes et p colonnes. Typiquement c'est un tableau de variables environnementales mesurées sur n sites. Un tableau **Y** a n lignes et q colonnes. Typiquement c'est un tableau de variables floro-faunistiques mesurées sur les mêmes n sites. L'analyse procustéenne permet une étude voisine du couple. Pour l'essentiel, la méthode est basée sur la rotation procustéenne dite *Procustes Rotation* présentée en écologie dans Digby et Kempton. (1987, Chapitre 4 *Methods for comparing ordinations*, § 4.1 *Procustes rotation*)[5].

```
pro1 <- procuste(pca1$tab, pca2$tab, nf = 2)
par(mfrow = c(2, 2))
s.traject(pro1$scor1, clab = 0)
s.label(pro1$scor1, clab = 0.8, add.p = T)
s.traject(pro1$scor2, clab = 0)
s.label(pro1$scor2, clab = 0.8, add.p = T)
s.arrow(pro1$load1, clab = 0.75)
s.arrow(pro1$load2, clab = 0.75)
```

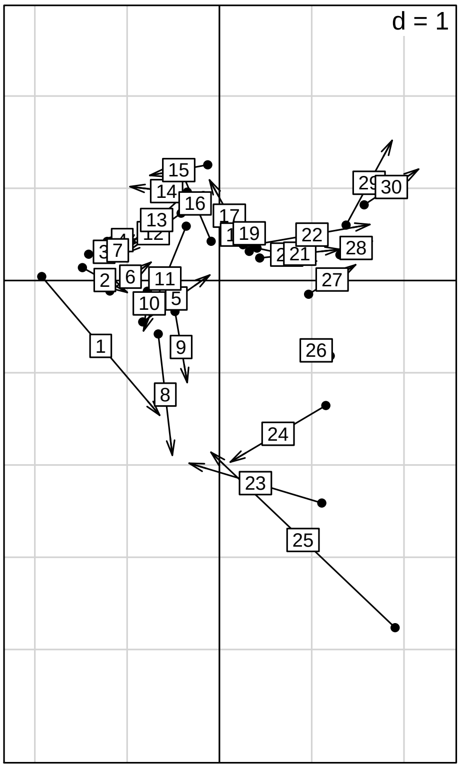



```
s.match(pro1$scor1, pro1$scor2, clab = 0.75)
```



Les deux nuages ont incontestablement des points communs quant à leur structure générale mais leur forme ne permet pas un ajustement complet. En particulier le nuage des points faunistiques est plus allongé que l'autre et la mise à l'échelle globale est trop simple.

```
s.match(coiner1$mX, coiner1$mY, clab = 0.75)
```



La remise à l'échelle (variance unité) par axe et par nuage change la présentation mais pas l'information acquise. En effet les calculs sont très voisins :

```
round(coiner1$eig/pro1$d^2, 1)
[1] 726.9 726.9 726.9 726.9 726.9 726.9 726.9 726.9 726.9 726.9 726.9
```

Au coefficient près de normalisation des tableaux, les valeurs propres de la co-inertie sont les carrés des valeurs singulières de la rotation procuste. Ce facteur vient de la pondération qu'on utilise dans la co-inertie ($1/n$) et de la mise à l'échelle qu'on utilise avant la rotation procuste.

Mais dans la rotation procuste, ce sont les racines carrées de ces valeurs propres qui sont importantes alors que dans la co-inertie ce sont les valeurs propres elles-mêmes. En effet, on cherche à minimiser :

$$d^2\left(\mathbf{X},\hat{\mathbf{Y}}\right) = \left\|\mathbf{X} - \mathbf{Y}\mathbf{R}^T\right\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \hat{y}_{ij}\right)^2$$

A l'optimum, cette distance sera :

$$\left\|\mathbf{Y} - \mathbf{X}\mathbf{R}\right\|^2 = 2\left(1 - \sum_{k=1}^{r}\theta_k\right)$$

La quantité $\sum_{k=1}^{r} \theta_k$ est caractéristique de la qualité d'ajustement de la rotation procuste. On appelle cette quantité $m^2_{\mathbf{XY}} = m^2_{\mathbf{YX}} = m^2$ quand les deux nuages ont reçu la mise à l'échelle initiale (Digby and Kempton *op.cit.* p. 114)[5] :

$$m^2 = \sum_{k=1}^{r} \theta_k$$

Le test `PROTEST` (Jackson 1995)[14] est le test de permutation entre les deux tableaux équivalent du test de permutation de la co-inertie qui a une constante près porte donc sur :

$$\alpha^2 = \sum_{k=1}^{r} \theta_k^2$$

PROTEST porte donc sur $\sum_{k=1}^{r} cov\left(\mathbf{Xu}_k, \mathbf{Yv}_k\right)$ alors que le test de la co-inertie porte sur $\sum_{k=1}^{r} cov^2\left(\mathbf{Xu}_k, \mathbf{Yv}_k\right)$. Les deux systèmes de vecteurs propres sont identiques en dépit de la mise à l'échelle initiale qui n'a pas lieu dans la co-inertie. On gardera donc les mêmes axes dans les deux analyses, mais leurs utilisation sera différente.

Notons cependant que la comparaison entre Procuste et Co-inertie ne tient que pour deux ACP centrées ou normées, la seconde méthode trouvant de multiples autres usages. Vérifier que les axes de co-inertie sont exactement les poids des variables de la seconde (en fait les axes de co-inertie des couples totalement appariés) :

```
cbind(coiner1$c1, pro1$load1)
```

```
            CS1         CS2         ax1         ax2
das  0.49281630  0.10036134 -0.49281630  0.10036134
alt -0.41165514 -0.13643154  0.41165514 -0.13643154
pen -0.34246684 -0.17335745  0.34246684 -0.17335745
deb  0.46832653  0.20752126 -0.46832653  0.20752126
pH  -0.03410717  0.20125773  0.03410717  0.20125773
dur  0.27484051  0.06073167 -0.27484051  0.06073167
pho  0.09126275 -0.40316684 -0.09126275 -0.40316684
nit  0.29574023 -0.16593741 -0.29574023 -0.16593741
amm  0.07317567 -0.43355309 -0.07317567 -0.43355309
oxy -0.25917016  0.43523863  0.25917016  0.43523863
dbo  0.07773827 -0.53527887 -0.07773827 -0.53527887
```

Utiliser ces valeurs pour représenter les variables des tableaux comme projections des vecteurs des bases canoniques sur les plans de références (observation qui ne semble pas être mentionnée dans la bibliographie).

Il y a donc une unité de fond entre l'analyse Procuste classique introduite en écologie par Digby et Kempton [5] qui l'attribuent à J.C. Gower[10] et l'analyse de co-inertie introduite dans [3] et qui trouve son origine dans l'article de Tucker (1958)[34]. La grande différence vient de la mise à l'échelle préalable dans le premier cas, suivie d'une rotation non déformante, suivie d'une projection alors que la seconde remet à l'échelle la projection sur chacun des axes.

La rotation Procuste trouve son plein intérêt quand on veut superposer une des deux analyses sur l'autre. C'est ce type d'exemple qui est proposé par Digby et Kempton ....(1987 p.116)[5]. Le chapitre 4 de cet ouvrage est fondamental. C'est le plus court et il ne contient pas de citations de la littérature écologique. Jackson ...(1995)[14] dit que : *Procustean methods are used infrequently in ecology. This lack of use likely reflects the proviously limited availability of the procedure.*

Mais ce chapitre 4 indique clairement que la comparaison de typologie sera fondamentale en écologie des communautés. Les analyses procustéennes généralisées y sont citées. Elles sont initiées par Gower ...(1975)[11]. Elles ont été l'objet d'une abondante bibliographie et de débats méthodologiques (ten Berge 1977[29], ten Berge and Bekker 1993[30]). Elles sont en concurrence avec d'autres méthodes K-tableaux.

6 Couplages d'Analayse des Correspondances Multiples

Si on prend deux ACM pour faire une analyse de co-inertie, on a un résultat théorique très particulier. On reprend les notations de la fiche tdr54. Le tableau disjonctif complet est associé à une pondération des individus et donne une pondération des modalités. Soient deux schémas d'ACM :

$$\begin{tikzcd} \boxed{l} \arrow[r, "\frac{1}{v}\mathbf{D}_l"] & \boxed{l} \arrow[d, "\mathbf{X_0}=\mathbf{XD}_{\mathrm{l}}^{-1}-\mathbf{1}_{\mathrm{nl}}"] \\ \boxed{n} \arrow[u, "\mathbf{X_0^T}"] & \boxed{n} \arrow[l, "\mathbf{D}"] \end{tikzcd} \qquad \begin{tikzcd} \boxed{m} \arrow[r, "\frac{1}{w}\mathbf{D}_m"] & \boxed{m} \arrow[d, "\mathbf{Y_0}=\mathbf{YD}_{\mathrm{m}}^{-1}-\mathbf{1}_{\mathrm{nm}}"] \\ \boxed{n} \arrow[u, "\mathbf{Y_0^T}"] & \boxed{n} \arrow[l, "\mathbf{D}"] \end{tikzcd}$$

Dans le premier, il y a l modalités réparties entre v variables. Dans le second, il y a m modalités réparties entre w variables. On croise les deux schémas :

$$\begin{tikzcd} \boxed{l} \arrow[r, "\frac{1}{v}\mathbf{D}_l"] & \boxed{l} \arrow[d, "\mathbf{Z}=\left(\mathbf{YD}_{\mathrm{m}}^{-1}-\mathbf{1}_{\mathrm{nm}}\right)^{\mathrm{T}}\mathbf{D}\left(\mathbf{XD}_{\mathrm{l}}^{-1}-\mathbf{1}_{\mathrm{nl}}\right)"] \\ \boxed{m} \arrow[u, "\mathbf{Z}^T"] & \boxed{m} \arrow[l, "\frac{1}{w}\mathbf{D}_m"] \end{tikzcd}$$

Alors :

$$\mathbf{Z} = \left(\mathbf{YD}_m^{-1} - \mathbf{1}_{nm}\right)^T \mathbf{D}\left(\mathbf{XD}_l^{-1} - \mathbf{1}_{nl}\right) = \mathbf{D}_m^{-1}\mathbf{Y}^T\mathbf{DXD}_l^{-1} - \mathbf{1}_{ml}$$

A une constante près, ceci est exactement le schéma de l'analyse des correspondances du tableau de Burt croisé entre les deux tableaux initiaux. Expérimenter cette propriété sur l'exemple `worksurv`. Source dans [23], reproduit dans [16]. Dans une enquête de 1970 auprès des travailleur français, on a extrait 4 questions :

1. pro (Professional elections). In professional elections in your firm, would you rather vote for a list supported by : "CGT" "CFDT" "FO" "CFTC" "Auton"(Autonomous) "Abst"(Abstention) "Nonaffi"(Nonaffiliated list) "NR" (No response)
2. una (Union affiliation). At the present time, are you affiliated to a Union, and in the affirmative, which one : "CGT" "CFDT" "FO" "CFTC" "Auton"(Autonomous) "CGC" "Notaffi" (Not affiliated) "NR" (No response)

3. pre (Presidential election). On the last presidential election (1969), can you tell me the candidate for whom you have voted ? ”Duclos” ”Deferre” ”Krivine” ”Rocard” ”Poher” ”Ducatel” ”Pompidou” ”NRAbs”
4. political sympathy. Which political party do you feel closest to, as a rule ? ”Communist” (PCF) ”Socialist” (SFIO+PSU+FGDS) ”Left” (Party of workers”,...) ”Center” (MRP+RAD.) ”RI” ”Right” (INDEP.+CNI) ”Gaullist” (UNR) ”NR” (No response)

```
data(worksurv)
eff <- attr(worksurv, "counts")
eff[1:20]
```

```
[1] 81  9  7  2  7  5  1  2  1  4  1  2  3  3  2  2  3  1  1  1
```

```
eff[301:319]
```

```
[1]  1  1  1  1  1  4  1  2  2 12  9  2  2  5  3 37  1  1  5
```

La proprié particulière de ces données est que chaque combinaison de réponses est assortie du nombre de personnes qui on répondu de cette manière. On a donc des variables qualitatives pondérées par des effectifs de réponse :

```
cbind(worksurv[1:10, ], eff[1:10])
```

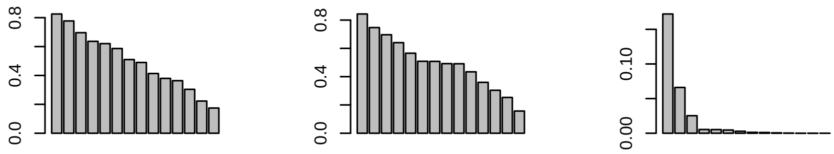
```
   pro una     pre       pol eff[1:10]
1  CGT CGT  Duclos Communist        81
2  CGT CGT  Duclos Socialist         9
3  CGT CGT  Duclos      Left         7
4  CGT CGT  Duclos    Center         2
5  CGT CGT  Duclos        NR         7
6  CGT CGT Deferre Socialist         5
7  CGT CGT Deferre     Right         1
8  CGT CGT Deferre        NR         2
9  CGT CGT Krivine Socialist         1
10 CGT CGT  Rocard Socialist         4
```

81 personnes ont répondu `CGT CGT Duclos Communist`, 9 personnes ont répondu `CGT CGT Duclos Socialist`, . . .Il y a en tout 309 types de réponse et 1049 personnes interrogées. Séparer le tableau en deux parties (opinions syndicales, opinions politiques), faire deux ACM et l’analyse du couple :

```
syndic <- worksurv[, 1:2]
politic <- worksurv[, 3:4]
par(mfrow = c(1, 3))
mca1 <- dudi.acm(syndic, eff, scannf = F, nf = 4)
mca2 <- dudi.acm(politic, eff, scannf = F, nf = 3)
coi <- coinertia(mca1, mca2, scannf = F, nf = 3)
summary(coi)
```

```
Eigenvalues decomposition:
         eig     covar       sdX       sdY      corr
1 0.17213619 0.4148930 0.8806157 0.9114774 0.5168967
2 0.06606594 0.2570329 0.8415122 0.8203781 0.3723182
3 0.02529015 0.1590288 0.7453460 0.8340557 0.2558131
Inertia & coinertia X:
     inertia       max     ratio
1   0.775484 0.8255839 0.9393158
12  1.483627 1.6028814 0.9255999
123 2.039167 2.2990180 0.8869732

Inertia & coinertia Y:
      inertia       max     ratio
1   0.8307911 0.8431065 0.9853929
12  1.5038114 1.5901795 0.9456866
123 2.1994603 2.2865250 0.9619227

RV:
 0.07112379
```

```
barplot(mca1$eig)
barplot(mca2$eig)
barplot(coi$eig)
```



Souligner le point remarquable de ce graphique. Pourquoi le RV est-il si faible ?

Construire le tableau de Burt.

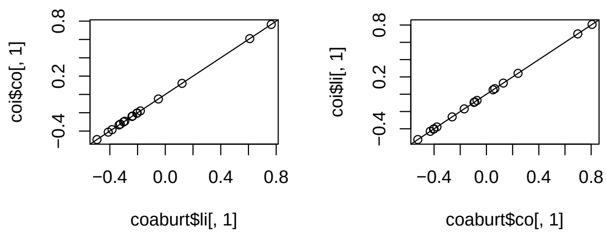
```
par(mfrow = c(1, 2))
burt1 <- acm.burt(syndic, politic, eff)
coaburt <- dudi.coa(burt1, scannf = F, nf = 3)
coaburt$eig
```

```
 [1] 1.721362e-01 6.606594e-02 2.529015e-02 5.367604e-03 5.265483e-03 4.639876e-03
 [7] 2.855106e-03 1.256895e-03 1.035097e-03 5.474492e-04 2.193185e-04 3.478312e-05
[13] 2.469734e-05 3.611902e-06
```

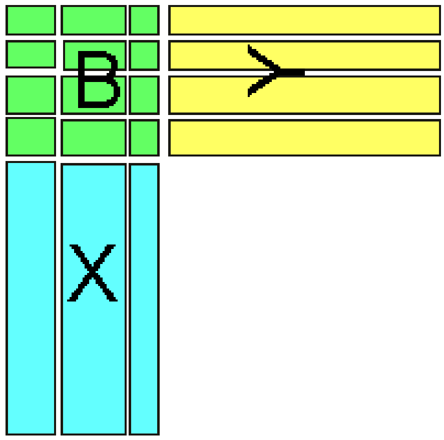
```
coi$eig
```

```
 [1] 1.721362e-01 6.606594e-02 2.529015e-02 5.367604e-03 5.265483e-03 4.639876e-03
 [7] 2.855106e-03 1.256895e-03 1.035097e-03 5.474492e-04 2.193185e-04 3.478312e-05
[13] 2.469734e-05 3.611902e-06
```

```
plot(coaburt$li[, 1], coi$co[, 1])
abline(0, 1)
plot(coaburt$co[, 1], coi$li[, 1])
abline(0, 1)
```



C'est pourquoi on dit souvent que l'Analyse des Correspondances Multiples *est* l'analyse des correspondances simples d'un tableau de Burt. En fait, c'est un point de vue pratique pour avoir un programme d'ACM avec simplement un programme d'AFC. Mais on a perdu la double représentation des individus (pattern de réponse). On la retrouve par le biais des individus supplémentaires. Les lignes des deux tableaux disjonctifs sont des lignes ou des colonnes supplémentaires du tableau de Burt :



La pratique de l'AFC des tableaux de Burt et de la projection des deux tableaux disjonctifs a été implantée par P. Cazes [1][2] comme **l'analyse canonique des variables qualitatives**. C'est en fait l'analyse de co-inertie des deux analyses des correspondances multiples [3].

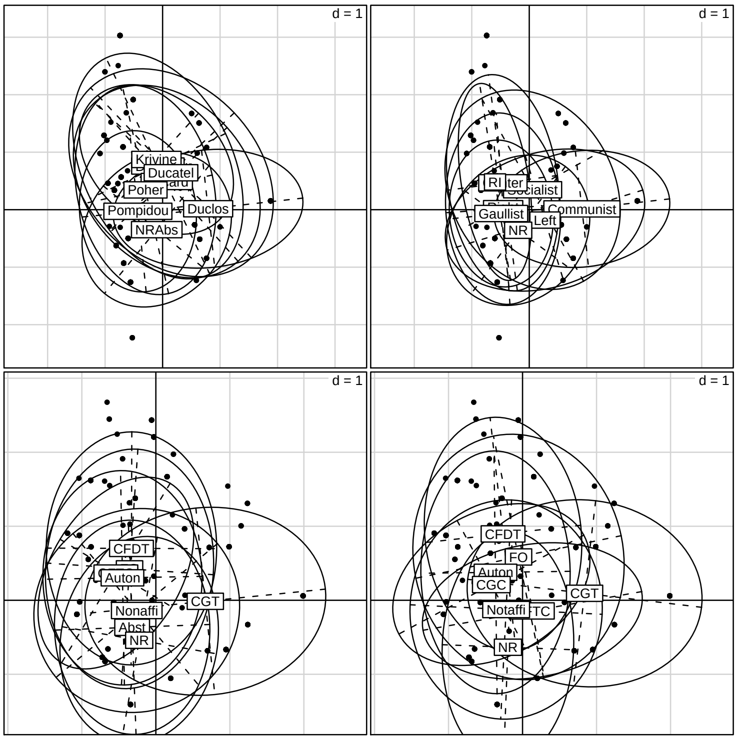
Le tableau de Burt contient 4 morceaux. Le premier croise élections professionnelles et élection présidentielle :

```
burt1[1:8, 1:8]
```

```
            pre.Duclos pre.Deferre pre.Krivine pre.Rocard pre.Poher pre.Ducatel
pro.CGT            155          15           2         10        31           2
pro.CFDT             8          12           4          6        24           1
pro.FO              10           8           1          2        19           1
pro.CFTC             1           1           1          2         3           1
pro.Auton           12           4           2          4        28           1
pro.Abst            16           4           0          2        18           0
pro.Nonaffi         17           3           0          2        15           0
pro.NR              14           2           0          2        11           1
            pre.Pompidou pre.NRAbs
pro.CGT               37        94
pro.CFDT              22        15
pro.FO                23        18
pro.CFTC              16         1
pro.Auton             39        23
pro.Abst              40        80
pro.Nonaffi           39        34
pro.NR                29        61
```

Tester l'association et vérifier la somme . Faire de même pour les 3 autres couples.

```
par(mfrow = c(2, 2))
s.class(coi$mX, politic[, 1], eff, cstar = 0)
s.class(coi$mX, politic[, 2], eff, cstar = 0)
s.class(coi$mY, syndic[, 1], eff, cstar = 0)
s.class(coi$mY, syndic[, 2], eff, cstar = 0)
```



Dans ce graphe, chaque variable d'un tableau est replacé dans le plan défini par l'autre. Cette représentation souligne que, même si la sympathie, le vote communiste, le vote CGT ou l'appartenance CGT définit un monde relativement à part, c'est l'ambiguïté qui domine dans ces données : l'élu est celui qui est au milieu. L'axe 2 souligne aussi le lien entre les non réponses (qui est aussi un type de réponse).

7 La théorie des profils écologiques

Elle est utilisée pour parler des relations des espèces avec leur milieu quand la description de ce dernier se fait avec des variables qualitatives (Godron et al. 1968)[9]. L'intérêt de cette approche est qu'elle est très accessible. Utiliser les données `tarentaise` décrites à :

`http://pbil.univ-lyon1.fr/R/pps/pps038.pdf`

```
data(tarentaise)
mil <- tarentaise$envir
fau <- tarentaise$ecol
table(fau$Mal)
```

```
  0   1
339  37
```

```
table(mil$alti)
```

```
A0600 A0750 A0900 A1050 A1200 A1350 A1500 A1650 A1800 A1950 A2100 A2250 A2400 A2550
   17    31    24    21    24    25    35    26    28    31    29    31    30    24
```

```
table(fau$Mal, mil$alti)
```

```
   A0600 A0750 A0900 A1050 A1200 A1350 A1500 A1650 A1800 A1950 A2100 A2250 A2400
 0    11    25    23    14    21    21    28    24    28    30    29    31    30
 1     6     6     1     7     3     4     7     2     0     1     0     0     0
   A2550
 0    24
 1     0
```

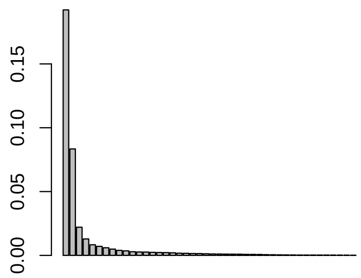
L'espèce Mal est présente 37 fois et absente 339 fois. Les 376 relevés sont répartis en 16 classes d'altitude. Dans la classe 600/750 m il y a 17 relevés et l'espèce Mal est présente 11 fois. Le profil écologique brut de l'espèce Mal sur la variable alti est le nombre de présences par classe de la variable.

```
table(mil$alti[fau$Mal == 1])
A0600 A0750 A0900 A1050 A1200 A1350 A1500 A1650 A1800 A1950 A2100 A2250 A2400 A2550
    6     6     1     7     3     4     7     2     0     1     0     0     0     0
chisq.test(fau$Mal, mil$alti, sim = T, B = 5000)
        Pearson's Chi-squared test with simulated p-value (based on 5000
        replicates)
data:  fau$Mal and mil$alti
X-squared = 52.0068, df = NA, p-value = 0.0002000
```

Le test Khi2 sur la table de contingence croisant la variable de milieu et la présence-absence de l'espèce indique si la répartition des presences de l'espèce est significativement différente de celle des absences, c'est-à-dire si l'espèce préfère significativement certaines modalités de milieu. En faisant cela, on arrive à 98*14 = 1372 tests pour les 98 espèces et les 14 variables de milieu.

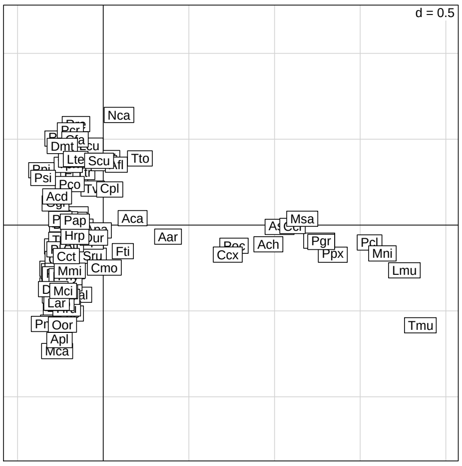
La nécessité de synthétiser cette multitude de relations binaires a conduit à essayer de représenter simultanément les espèces et les modalités de milieu pour rendre compte des facteurs écologiques essentiels. Romane [?] [22] a proposé empiriquement d'utiliser l'AFC sur le tableau qui regroupe toutes les espèces sur toutes les variables. On a ce tableau par une simple multiplication de matrice :

```
mil01 <- acm.disjonctif(mil)
dim(mil01)
[1] 376  58
prof <- data.frame(t(as.matrix(fau)) %*% as.matrix(mil01))
prof[1:5, 1:7]
    alti.A0600 alti.A0750 alti.A0900 alti.A1050 alti.A1200 alti.A1350 alti.A1500
Mal          6          6          1          7          3          4          7
Cca          8         12          9         15         11          5         10
Svu         13         17         11         12          5          1          1
Sat         14         27         18         12         13          9          9
Cbr          2          5          7          5          1          0          0
coaprof <- dudi.coa(prof, scannf = F)
```



Les lignes de ce tableau sont les espèces et les colonnes sont les modalités.

```
s.label(coaprof$li)
```



La carte des lignes donne la position des espèces. Elle est reliée à la carte des colonnes qu'on peut multi-fenêtrer pour plus de lisibilité.

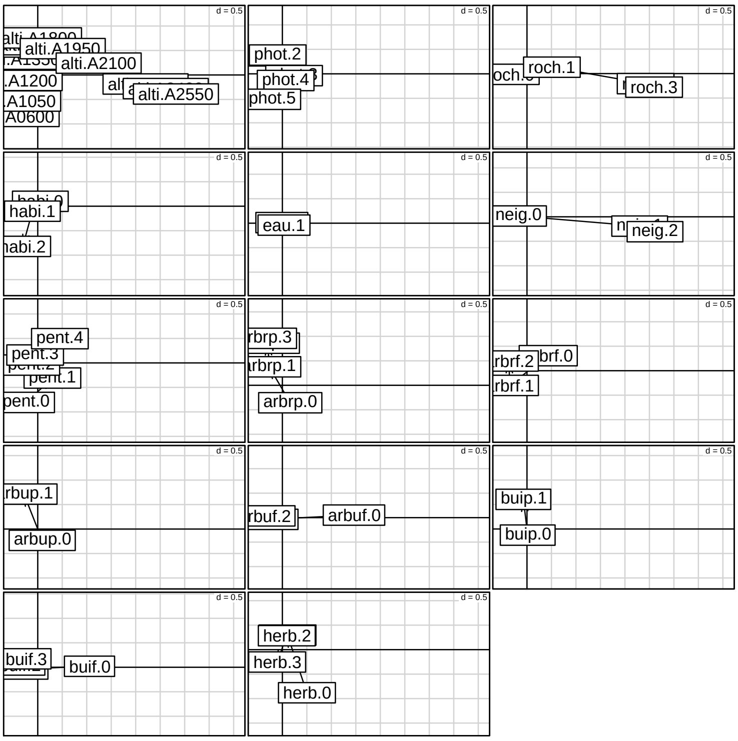
```
clanum <- factor(rep(1:14, unlist(apply(mil, 2, function(x) nlevels(factor(x))))))
clanum
```

```
 [1] 1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  3  3  3  3  4  4  4  5
[28] 5  6  6  6  7  7  7  7  7  8  8  8  8  9  9  9  10 10 11 11 11 12 12 13 13 13 13
[55] 14 14 14 14
Levels: 1 2 3 4 5 6 7 8 9 10 11 12 13 14
```

```
par(mfrow = c(5, 3))
f1 <- function(x, ...) {
    s.traject(x, clab = 0, ...)
    s.label(x, add.plot = T, clab = 2)
}
lapply(split(coaprof$co, clanum), f1, xlim = range(coaprof$co[,
    1]) * 1.5)
```

```
$`1`
s.label(dfxy = x, clabel = 2, add.plot = T)
$`2`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`3`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`4`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`5`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`6`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`7`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`8`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`9`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`10`
s.label(dfxy = x, clabel = 2, add.plot = T)

$`11`
s.label(dfxy = x, clabel = 2, add.plot = T)
```

```
$`12`
s.label(dfxy = x, clabel = 2, add.plot = T)
$`13`
s.label(dfxy = x, clabel = 2, add.plot = T)
$`14`
s.label(dfxy = x, clabel = 2, add.plot = T)
```



Le facteur 1 a une interprétation triviale. L'intérêt de cette analyse des correspondances est d'être une analyse de co-inertie entre l'AFC du tableau faunistique et l'ACM du tableau de milieu *après introduction dans l'ACM de la pondération des lignes de l'AFC* :

$$\begin{tikzcd} \boxed{p} \arrow[r, "\mathbf{D}_p"] & \boxed{p} \arrow[d, "\mathbf{X_0}=\mathbf{D}_\mathrm{n}^{-1}\mathbf{X}\mathbf{D}_\mathrm{p}^{-1}-\mathbf{1}_\mathrm{np}"] \\ \boxed{n} \arrow[u, "\mathbf{X}_\mathbf{0}^\mathrm{T}"] & \boxed{n} \arrow[l, "\mathbf{D}_n"] \end{tikzcd} \qquad \begin{tikzcd} \boxed{m} \arrow[r, "\frac{1}{v}\mathbf{D}_m"] & \boxed{m} \arrow[d, "\mathbf{Y_0}=\mathbf{Y}\mathbf{D}_m^{-1}-\mathbf{1}_{nm}"] \\ \boxed{n} \arrow[u, "\mathbf{Y}_\mathbf{0}^T"] & \boxed{n} \arrow[l, "\mathbf{D}_n"] \end{tikzcd}$$

Dans le premier, il y a p espèces (colonnes) et n lignes (sites). Dans le second, il y a m modalités réparties entre v variables. On croise les deux schémas :

$$dudipm\mathbf{D}_p\mathbf{Z} = \left(\mathbf{Y}\mathbf{D}_m^{-1} - \mathbf{1}_{nm}\right)^T \mathbf{D}_n \left(\mathbf{D}_n^{-1}\mathbf{X}\mathbf{D}_p^{-1} - \mathbf{1}_{np}\right)\frac{1}{v}\mathbf{D}_m\mathbf{Z}^T$$

Alors :

$$\mathbf{Z} = \mathbf{Z} = \left(\mathbf{Y}\mathbf{D}_m^{-1} - \mathbf{1}_{nm}\right)^T \mathbf{D}_n \left(\mathbf{D}_n^{-1}\mathbf{X}\mathbf{D}_p^{-1} - \mathbf{1}_{np}\right) = \mathbf{D}_m^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{D}_p^{-1} - \mathbf{1}_{mp}$$

On a exactement une AFC simple.

```
dudi1 <- dudi.coa(fau, scannf = F, nf = 3)
dudi2 <- dudi.acm(mil, dudi1$lw, scannf = F, nf = 3)
coi <- coinertia(dudi1, dudi2, scannf = F, nf = 3)
summary(coi)
```

```
Eigenvalues decomposition:
         eig     covar       sdX       sdY      corr
1 0.19229723 0.4385171 0.8768820 0.5300928 0.9433948
2 0.08338545 0.2887654 0.7007483 0.4601658 0.8955066
3 0.02207917 0.1485906 0.5510354 0.3609263 0.7471250
Inertia & coinertia X:
     inertia       max     ratio
1   0.768922 0.7769446 0.9896741
12  1.259970 1.2746702 0.9884676
123 1.563610 1.7126438 0.9129804

Inertia & coinertia Y:
      inertia       max     ratio
1   0.2809984 0.2830892 0.9926143
12  0.4927509 0.5030206 0.9795840
123 0.6230187 0.6450949 0.9657784

RV:
 0.4397868
```

```
coaprof$eig[1:14]
```

```
 [1] 0.192297232 0.083385451 0.022079171 0.012887575 0.008330700 0.007077041
 [7] 0.005995169 0.004916215 0.003928216 0.003559962 0.002893564 0.002685993
[13] 0.002599630 0.002383271
```

```
coi$eig[1:14]
```

```
 [1] 0.192297232 0.083385451 0.022079171 0.012887575 0.008330700 0.007077041
 [7] 0.005995169 0.004916215 0.003928216 0.003559962 0.002893564 0.002685993
[13] 0.002599630 0.002383271
```

Vérifier l'identité des coordonnées :

```
plot(coaprof$li[, 1], coi$co[, 1])
plot(coaprof$co[, 1], coi$li[, 1])
```

En bref, la co-inertie de deux ACP normées est l'inter-batterie de Tucker[34], la co-inertie de deux ACM est l'analyse canonique sur variables qualitatives de Cazes et la co-inertie d'une AFC et d'une ACM est l'AFC sur profils écologiques de Romane, retrouvée dans Montaña and Greig-Smith 1990[18] et explicitée dans [17]. L'intérêt est dans le principe de l'analyse. On peut y mettre deux triplets quelconques et obtenir des propriétés particulières par simple lecture du schéma croisé.

8 Nouvelles associations de tableaux : l'exemple de (`niche`)

En écologie, on pense souvent à la niche d'une espèce comme partie de l'espace multivarié qu'elle occupe. Utiliser la liste `trichometeo`[35] décrite dans :

`http://pbil.univ-lyon1.fr/R/pps/pps034.pdf`

On y trouve 49 unités de piégeages lumineux.

```
data(trichometeo)
fau <- trichometeo$fau
meteo <- trichometeo$meteo
dim(fau)
```

```
[1] 49 17
```

```
names(fau)
```

```
 [1] "Che" "Hyc" "Hym" "Hys" "Psy" "Aga" "Glo" "Ath" "Cea" "Ced" "Set" "All" "Han"
[14] "Hfo" "Hsp" "Hve" "Sta"
```

```
dim(meteo)
```

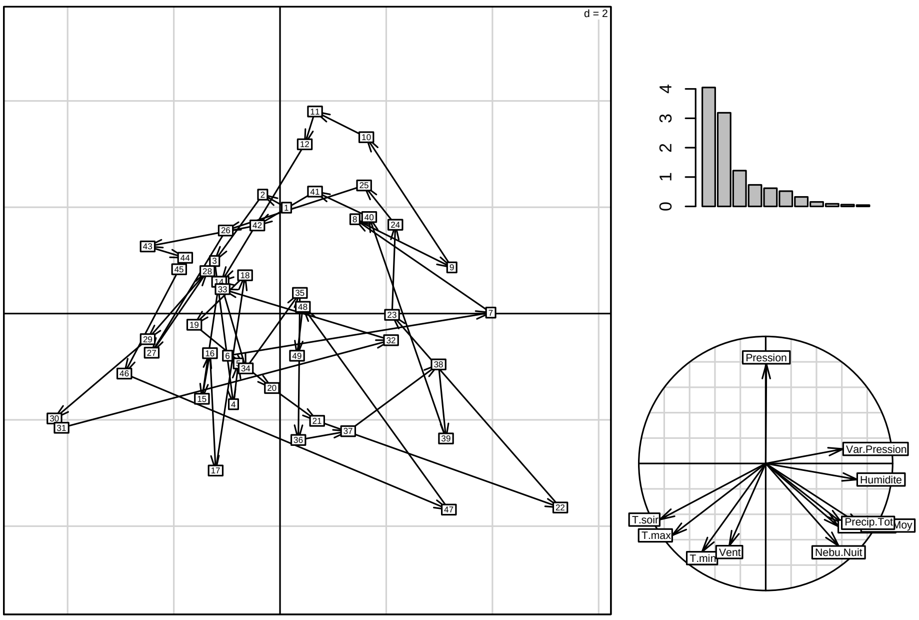
```
[1] 49 11
```

```
names(meteo)
```

```
 [1] "T.max"        "T.soir"       "T.min"        "Vent"         "Pression"
 [6] "Var.Pression" "Humidite"     "Nebu.Nuit"    "Precip.Nuit"  "Nebu.Moy"
[11] "Precip.Tot"
```

Transformer les données faunistiques :

```
faulog <- log(fau + 1)
pca <- dudi.pca(meteo, scannf = F, nf = 3)
layout(matrix(c(1, 1, 1, 1, 2, 3), nrow = 2))
s.traject(pca$li, clab = 0)
s.label(pca$li, add.p = T, clab = 0.75)
barplot(pca$eig)
s.corcircle(pca$co)
```



Les variables forment 3 groupes Pression / Température / Précipitations, le cycle beau temps -> chaleur -> orage se reproduit en été. On se demande si les conditions météorologiques influencent plusieurs espèces de la même manière (les pièges capturent les insectes adultes après leur émergence).

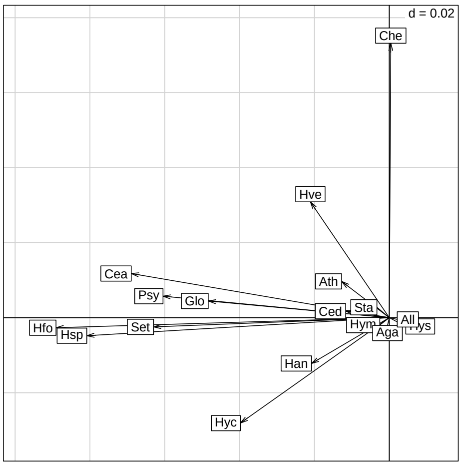
```
fauprof <- apply(fau, 2, function(x) x/sum(x))
```

On a calculé les profils par espèces (somme 1 par colonne). Quelles sont les moyennes par variables ? Que signifie le centrage d'un tel tableau ? Faire l'ACP centrée (sans normalisation) :

```
dim(fauprof)
```

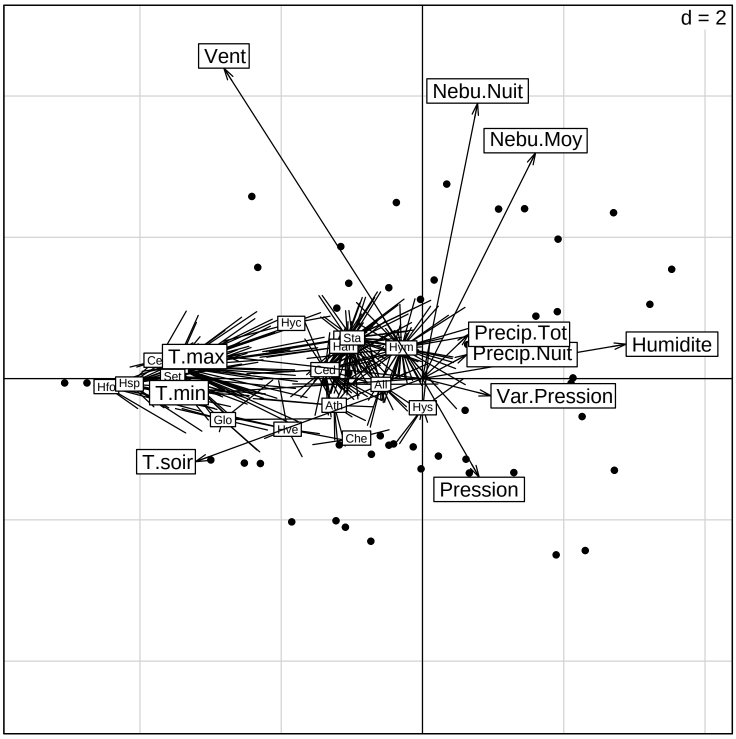
```
[1] 49 17
```

```
pcafau <- dudi.pca(fauprof, scal = F, scannf = F)
s.arrow(pcafau$co)
```



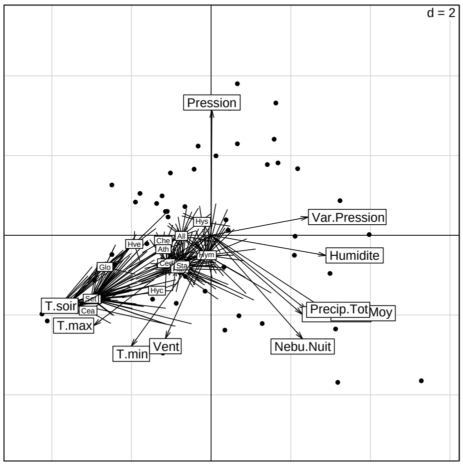
Interpréter. Construire le schéma de la co-inertie de ces deux ACP. Donner un sens au tableau croisé. Expliciter le critère maximisé. Exécuter et interpréter.

```
coi <- coinertia(pca, pcafau, scannf = F)
s.label(coi$lX, clab = 0)
s.distri(coi$lX, data.frame(fauprof), clab = 0.6, add.p = T, cell = 0,
    csta = 0.3)
s.arrow(7 * coi$c1, clab = 1, add.p = T)
```



Retrouver le résultat de l'analyse dans le plan de l'analyse simple :

```
s.label(pca$li, clab = 0)
s.distri(pca$li, data.frame(fauprof), clab = 0.6, add.p = T, cell = 0,
    csta = 0.3)
s.arrow(4 * pca$co, clab = 1, add.p = T)
```



Donner une légende. Expliquer pourquoi l'utilisation de la CCA serait regrettable sur ce type de données. Cette pratique est disponible dans la fonction `niche` [**?**].

9 Co-inertie et CCA

Une CCA (Canonical Correspondence Analysis) croise un tableau d'ACP normée et un tableau d'AFC [31] [32]. Son schéma est du type :

$$JII\mathbf{D}_J\mathbf{Y}_0 = \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ}\mathbf{P}_{\mathbf{X}_0} = \mathbf{X}_0\left(\mathbf{X}_0^T\mathbf{D}_I\mathbf{X}_0\right)^{-}\mathbf{X}_0^T\mathbf{D}_I\mathbf{D}_I\mathbf{P}_{\mathbf{X}_0}^T\mathbf{Y}_0^T$$

I sites contiennent J espèces et donnent p variables environnementales. Les variables de $\mathbf{X}$ sont centrées et normées pour la pondération issue de l'AFC ($\mathbf{D}_I$). En première approche, on a du mal à reconnaître les deux schémas initiaux :

$$\begin{tikzcd} \boxed{p} \arrow[r, "\mathbf{I}_p"] & \boxed{p} \arrow[d, "\mathbf{X}_0"] \\ \boxed{I} \arrow[u, "\mathbf{X}_0^T"] & \boxed{I} \arrow[l, "\mathbf{D}_I"] \end{tikzcd} \qquad \begin{tikzcd} \boxed{J} \arrow[r, "\mathbf{D}_J"] & \boxed{J} \arrow[d, "\mathbf{Y}_0=\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}-\mathbf{1}_{IJ}"] \\ \boxed{I} \arrow[u, "\mathbf{Y}_0^T"] & \boxed{I} \arrow[l, "\mathbf{D}_I"] \end{tikzcd}$$

Dans le premier, il y a p variables (colonnes) et I sites (lignes). Dans le second, il y a J espèces (colonnes) et I sites (lignes). Il suffit en fait de reconnaître :

$$\mathbf{P}_{\mathbf{X}_0}^T\mathbf{D}_I\mathbf{P}_{\mathbf{X}_0} = \mathbf{D}_I\mathbf{P}_{\mathbf{X}_0}$$

pour voir dans la CCA le schéma :

$$JIp\mathbf{D}_J\mathbf{Y}_0 = \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ}\mathbf{X}_0^T\mathbf{D}_I\left(\mathbf{X}_0^T\mathbf{D}_I\mathbf{X}_0\right)^{-}\mathbf{D}_I\mathbf{X}_0\mathbf{Y}_0^T$$

ou encore $(\mathbf{X}_0^T\mathbf{D}_I\mathbf{1}_{IJ} = \mathbf{0}_{pJ})$:

$$\begin{tikzcd} \boxed{p} \arrow[r, "\left(\mathbf{X}_0^T\mathbf{D}_I\mathbf{X}_0\right)^-"] & \boxed{p} \arrow[d, "\mathbf{Z}=\mathbf{D}_I^{-1}\mathbf{P}^T\mathbf{X}_0"] \\ \boxed{J} \arrow[u, "\mathbf{Z}^T"] & \boxed{J} \arrow[l, "\mathbf{D}_J"] \end{tikzcd}$$

voisin de celui de la co-inertie :

$$\begin{tikzcd} \boxed{p} \arrow[r, "\mathbf{I}_p"] & \boxed{p} \arrow[d, "\mathbf{Z}=\mathbf{D}_I^{-1}\mathbf{P}^T\mathbf{X}_0"] \\ \boxed{J} \arrow[u, "\mathbf{Z}^T"] & \boxed{J} \arrow[l, "\mathbf{D}_J"] \end{tikzcd}$$

Les deux méthodes sont directement comparables. La première donne, pour un facteur $\mathbf{f}$ vérifiant :

$$\|\mathbf{f}\|^2_{\left(\mathbf{X}_0^T\mathbf{D}_I\mathbf{X}_0\right)} = 1 \Rightarrow \mathbf{f}^T\mathbf{X}_0^T\mathbf{D}_I\mathbf{X}_0\mathbf{f} = \|\mathbf{X}_0\mathbf{f}\|^2_{\mathbf{D}_I} = 1$$

une combinaison linéaire de variables $\mathbf{X}_0\mathbf{f}$centrée et réduite maximisant $\left\|\mathbf{D}_J^{-1}\mathbf{P}^T\mathbf{X}_0\mathbf{f}\right\|^2_{\mathbf{D}_J}$.

On reconnaît $\mathbf{D}_J^{-1}\mathbf{P}^T\mathbf{X}_0\mathbf{f}$ comme étant le vecteur des moyennes par espèce du score $\mathbf{X}_0\mathbf{f}$ lui-même $\mathbf{D}_J$-centré car :

$$\mathbf{1}_J\mathbf{D}_J\mathbf{D}_J^{-1}\mathbf{P}^T\mathbf{X}_0\mathbf{f} = \mathbf{1}_I\mathbf{D}_I\mathbf{X}_0\mathbf{f} = 0$$

.

La CCA donne des combinaisons de variables de milieu de variance unité maximisant la variance des positions moyennes des espèces. La seconde donne, pour un facteur $\mathbf{f}$ vérifiant $\|\mathbf{f}\|^2_{\mathbf{I}_p} = 1 \Rightarrow \mathbf{f}^T\mathbf{f} = 1$ une combinaison linéaire de variables $\mathbf{X}_0\mathbf{f}$centrée et réduite maximisant $\left\|\mathbf{D}_J^{-1}\mathbf{P}^T\mathbf{X}_0\mathbf{f}\right\|^2_{\mathbf{D}_J}$.

On reconnaît $\mathbf{D}_J^{-1}\mathbf{P}^T\mathbf{X}_0\mathbf{f}$ comme étant le vecteur des moyennes par espèce du score $\mathbf{X}_0\mathbf{f}$ lui-même $\mathbf{D}_J$-centré car :

$$\mathbf{1}_j\mathbf{D}_J\mathbf{D}_J^{-1}\mathbf{P}^T\mathbf{X}_0\mathbf{f} = \mathbf{1}_I\mathbf{D}_I\mathbf{X}_0\mathbf{f} = 0$$

.

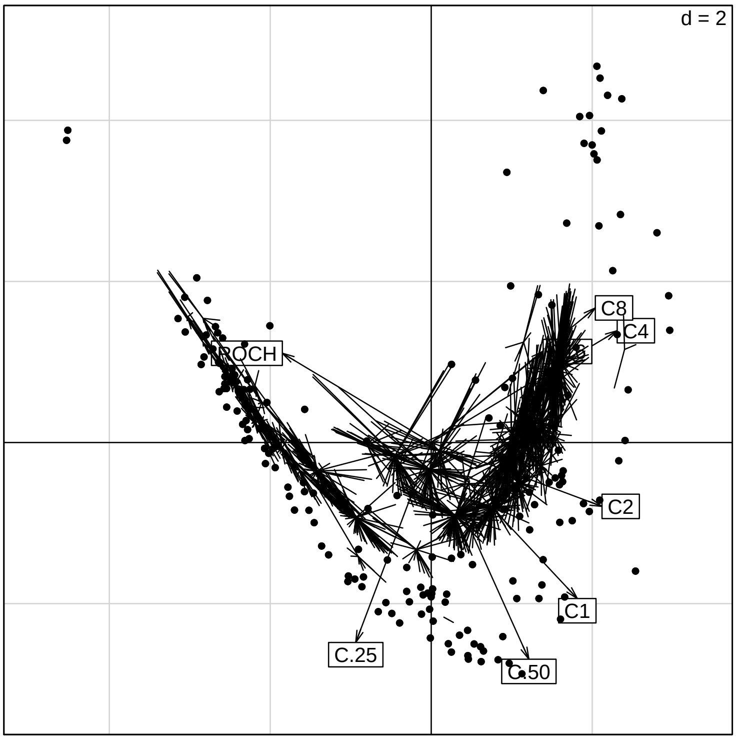
L'analyse de co-inertie donne des combinaisons de variables de milieu maximisant la variance des positions moyennes des espèces. Dans le premier cas on maximise la variance des moyennes sous la contrainte que la variance totale vaut 1. Dans la seconde on maximise la variance des moyennes en misant sur la variance (critère d'ACP) et la variance des moyennes (critère d'AFC).

Comparer les deux méthodes sur les données `rpjdl`.

```
data(rpjdl)
coafau <- dudi.coa(rpjdl$fau, scannf = F, nf = 4)
pcamil <- dudi.pca(rpjdl$mil, row.w = coafau$lw, scannf = F, nf = 2)
coi <- coinertia(pcamil, coafau, scannf = F, nf = 2)
summary(coi)
```

```
Eigenvalues decomposition:
        eig    covar      sdX       sdY      corr
1 2.0578677 1.434527 1.774003 0.8642081 0.9356988
2 0.4433456 0.665842 1.727995 0.5110553 0.7539817
Inertia & coinertia X:
    inertia      max     ratio
1  3.147086 3.150520 0.9989102
12 6.133053 6.211566 0.9873601

Inertia & coinertia Y:
     inertia       max    ratio
1  0.7468556 0.7532461 0.991516
12 1.0080332 1.0461518 0.963563

RV:
 0.5681413
```

```
s.label(coi$lX, clab = 0)
s.arrow(5 * coi$c1, add.plot = T)
s.distri(coi$lX, rpjdl$fau, cst = 0.25, add.plot = T, cell = 0)
```



On obtient un *triplot* de co-inertie. Les flèches indiquent le poids des variables. Les points donnent la position des sites par combinaison linéaire des variables. Les étoiles représentent la dispersion des espèces dans l'espace écologique. On a le gradient en milieu ouvert, le gradient en milieu fermé et l'articulation des deux par les espèces tolérantes.

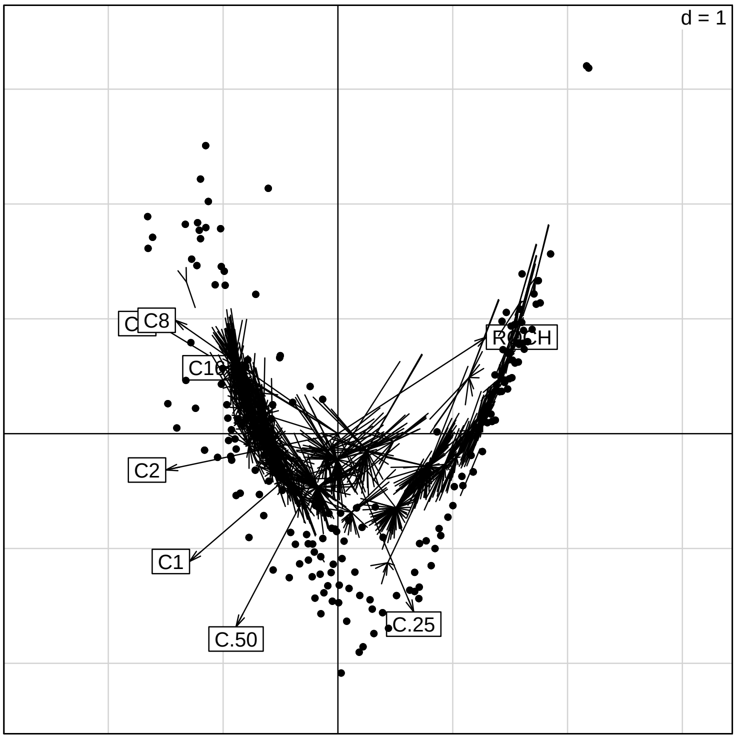
```
cca1 <- pcaiv(coafau, pcamil$tab, scannf = F, nf = 2)
cca1
```

```
Principal Component Analysis with Instrumental Variables
call: pcaiv(dudi = coafau, df = pcamil$tab, scannf = F, nf = 2)
class: pcaiv dudi
$rank (rank)     : 8
$nf (axis saved) : 2

eigen values: 0.6617 0.1773 0.07054 0.03891 0.03616 ...

 vector length mode    content
 $eig   8      numeric eigen values
 $lw    182    numeric row weigths (from dudi)
 $cw    51     numeric col weigths (from dudi)

 data.frame nrow ncol content
 $Y         182  51   Dependant variables
 $X         182  8    Explanatory variables
 $tab       182  51   modified array (projected variables)

 data.frame nrow ncol content
 $c1        51   2    PPA Pseudo Principal Axes
 $as        4    2    Principal axis of dudi$tab on PAP
 $ls        182  2    projection of lines of dudi$tab on PPA
 $li        182  2    $ls predicted by X

 data.frame nrow ncol content
 $fa        9    2    Loadings (CPC as linear combinations of X
 $l1        182  2    CPC Constraint Principal Components
 $co        51   2    inner product CPC - Y
 $cor       9    2    correlation CPC - X

 iner  inercum inerC inercumC ratio R2    lambda
 0.753 0.753   0.747 0.747    0.992 0.885 0.662
 0.293 1.05    0.275 1.02     0.977 0.645 0.177
```

```
w1 <- cor(cca1$l1, pcamil$tab)
w1
```

```
         ROCH        C.25       C.50         C1         C2         C4         C8
RS1 0.6464899  0.3301096 -0.4436866 -0.6462082 -0.7499994 -0.7926132 -0.7080341
RS2 0.4203057 -0.7762804 -0.8410546 -0.5564833 -0.1583008  0.4793813  0.4936859
           C16
RS1 -0.4637386
RS2  0.2867292
```

```
w1 <- cor(pcamil$tab, cca1$l1)
s.label(cca1$l1, clab = 0)
s.arrow(2 * w1, add.plot = T)
s.distri(cca1$l1, rpjdl$fau, cst = 0.25, add.plot = T, cell = 0)
```



On obtient un *triplot* de cca. Les points donnent la position des sites par combinaison linéaire (scores) des variables de **variance unité**. Les flèches indiquent les corrélations entre variables et scores des sites. Les étoiles représentent la dispersion des espèces dans l'espace écologique. On a le gradient en milieu ouvert, le gradient en milieu fermé et l'articulation des deux par les espèces tolérantes.

Quand le nombre des variables est limité (ici 8 variables pour 182 points) la CCA est meilleure. Quand l'équilibre entre nombre de sites et nombre de variables rend les régressions numériquement instables, préférer la co-inertie.

On retrouvera la notion de co-inertie dans son extension à K-tableaux. Pour citer la méthode, on peut utiliser les mises au point récentes qui reprennent ces éléments [7] [6].

Références

[1] P. Cazes. L'analyse de certains tableaux rectangulaires décomposé en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. i. définitions et applications à l'analyse canonique des variables qualitatives. ii questionnaires : variantes des codages et nouveaux calculs de contributions. *Les Cahiers de l'Analyse des Données*, 5 :145–161 & 387–406, 1980.

[2] P. Cazes. L'analyse de certains tableaux rectangulaires décomposé en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. iii codage simultanné de variables qualitatives et quantitatives. iv cas modèles. *Les Cahiers de l'Analyse des Données*, 6 :9–18 & 135–143, 1981.

[3] D. Chessel and P. Mercier. Couplage de triplets statistiques et liaisons espèces-environnement. In J.D. Lebreton and B. Asselain, editors, *Biométrie et Environnement*, pages 15–44. Masson, Paris, 1993.

[4] N. Cliff. Orthogonal rotation to congruence. *Psychometrika*, 31 :33–42, 1966.

[5] P. G. N. Digby and R. A. . Kempton. *Multivariate Analysis of Ecological Communities*. Chapman and Hall, Population and Community Biology Series, London, 1987.

[6] S. Dray, D. Chessel, and J. Thioulouse. Co-inertia analysis and the linking of ecological tables. *Ecology*, 84(11) :3078–3089, 2003.

[7] S. Dray, D. Chessel, and J. Thioulouse. Procustean co-inertia analysis for the linking of multivariate datasets. *Ecoscience*, 10 :110–119, 2003.

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

[9] M. Godron, P. Daget, L. Emberger, E. Le Floch, J. Poissonet, C. Sauvage, and J.P. Wacquant. *Relevé méthodique de la végétation et du milieu*. Editions du CNRS, Paris, 1968.

[10] J.C. Gower. Statistical methods of comparing different multivariate analyses of the same data. In F.R Hodson, D.G. Kendall, and P. Tautu, editors, *Mathematics in the archaeological and historical sciences*, pages 138–149. University Press, Edinburgh, 1971.

[11] J.C. Gower. Generalized procustes analysis. *Psychometrika*, 40 :33–51, 1975.

[12] B.F. Green. The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, 17 :429–440, 1952.

[13] J.R. Hurley and R.B. Cattel. Producing direct rotation to test a hypothetized factor structure. *Behavourial Science*, 7 :258–262, 1962.

[14] D.A. Jackson. Protest : a procustean randomization test of community environment concordance. *Ecosciences*, 2 :297–303, 1995.

[15] C.P. Klingenberg and G.S. McIntyre. Geometric morphometrics of developmental instability : Analyzing patterns of fluctuating asymmetry with procrustes. *Evolution*, 52 :1363–1375, 1998.

[16] B. Le Roux and H. Rouanet. Interpreting axes in multiple correspondence analysis : method of the contributions of points and deviation. In Blasius J. and M. Greenacre, editors, *Visualization of categorical data*, pages 197–220. Acamedic Press, London, 1997.

[17] P Mercier, D. Chessel, and S. Dolédec. Complete correspondence analysis of an ecological profile data table : a central ordination method. *Acta Œcologica*, 13 :25–44, 1992.

[18] C. Montaña and P. Greig-Smith. Correspondence analysis of species by environmental variable matrices. *Journal of Vegetation Science*, 1 :453–460, 1990.

[19] F. Mouttet. *Comparaison de tableaux par la méthode Procuste*. Thèse de 3° cycle, Paris VI, 1981.

[20] A.F. Olshan, A.F. Siegel, and D.R. Swindler. Robust and least-squares orthogonal mapping : Methods for the study of cephalofacial form and growth. *American Journal of Physical Anthropology*, 59 :131–137, 1982.

[21] F.J. Rohlf and D. Slice. Extensions of the procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, 39 :40–59, 1990.

[22] F. Romane. Un exemple d'utilisation de l'analyse factorielle des correspondances en écologie végétale. In E. van der Maarel and R. Tùxen, editors, *Grundfagen und Methoden in der Pflanzensoziologie*, pages 155–162. Dr. W. Junk b.v., The Hague, 1972.

[23] H. Rouanet and B. Le Roux. *Analyse des données multidimensionnelles*. Dunod, paris, 1993.

[24] P.H. Schönemann. A generalized solution solution of the orthogonal procustes problem. *Psychometrika*, 31 :1–10, 1966.

[25] P.H. Schönemann. On two-sided procustes problems. *Psychometrika*, 33 :19–34, 1968.

[26] P.H. Schönemann and R.M. Carrol. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 35 :245–256, 1970.

[27] R. Sibson. Studies in the robustness of multidimensional scaling. *Journal of the Royal Statistical Society, B*, 40 :234–238, 1978.

[28] P.H.A. Sneath. Trend-surface analysis of transformation grids. *Journal of Zoology*, 151 :65–122, 1967.

[29] J.M.F. ten Berge. Orthogonal procustes rotation for two or more matrices. *Psychometrika*, 42 :267–276, 1977.

[30] J.M.F. ten Berge and P.A. Bekker. The isotropic scaling problem in generalized procustes analysis. *Computational Statistics and Data Analysis*, 16 :201–204, 1993.

[31] C.J.F. Ter Braak. Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 :1167–1179, 1986.

[32] C.J.F. Ter Braak. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69 :69–77, 1987.

[33] F. Torre and D. Chessel. Co-structure de deux tableaux totalement appariés. *Revue de Statistique Appliquée*, 43 :109–121, 1994.

[34] L.R. . Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23 :111–136, 1958.

[35] P. Usseglio-Polatera and Y. Auda. Influence des facteurs météorologiques sur les résultats de piégeage lumineux. *Annales de Limnologie*, 23 :65–79, 1987.

[36] J. Verneaux. Cours d'eau de franche-comté (massif du jura). recherches écologiques sur le réseau hydrographique du doubs. essai de biotypologie. Thèse d'état, besançon, 1973.