ANALYSE CANONIQUE

Définition

- Méthode d'analyse multidimensionnelle qui présente des analogies à la fois avec l'analyse en composantes principales (A.C.P.) et avec la régression linéaire.
- Intérêt de l'AC essentiellement théorique : plusieurs méthodes d'ADD soit des cas particuliers. En pratique, les interprétations sont délicates

Objectif général de l'A.C. : Explorer les relations pouvant exister entre deux groupes de variables quantitatives observées sur le même ensemble d'individus, afin d'expliquer un groupe avec l'autres.

ANALYSE CANONIQUE

Les données

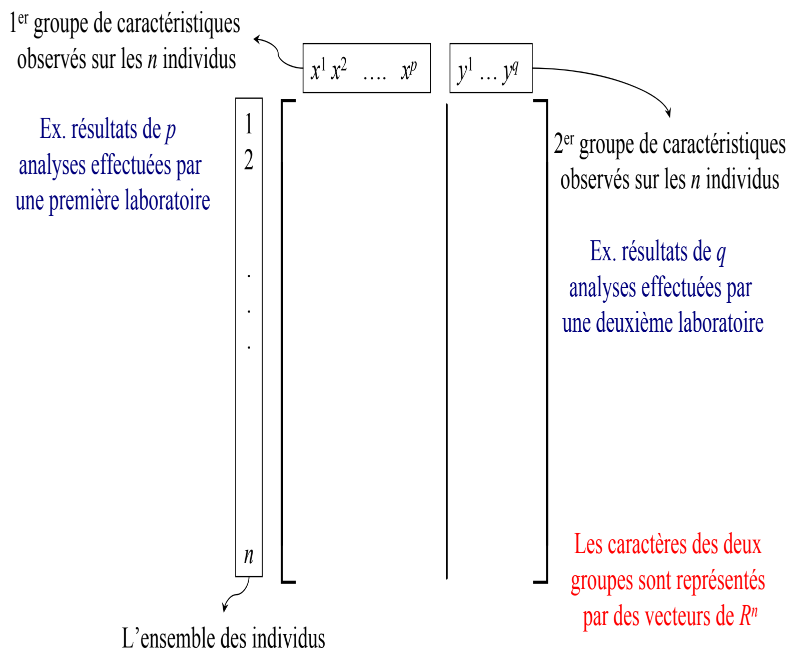
- On observe sur N individus sur P (<=N) variables quantitatives (X1,…,XP) plus Q (<=N) autres variables quantitatives (Y1, …,YQ). Si l'on se place dans une optique de prévision, les premières sont considérées comme explicatives les autres comme étant à expliquer par les premières.

- On appelle X (dim N*P) le tableau des variables explicatives et Y (dim N*Q) celui des variables à expliquer ou variables dépendantes.

- On suppose que toutes les variables X et Y sont centrées (sinon on les centre) et que les individus ont tous un poids 1/N : $P=I_N/N$

RQ : s'applique aussi sur des variables qualitatives dichotomiques. Par extension, aux autres variables qualitatives, en construisant autant de variables dichotomiques que de modalités

Schématiquement



- Matrice de variance-covariance du tableau complet:

$$\boxed{V = C'PC = \begin{pmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{pmatrix}}$$

Chacun des 2 tableaux de données décrit un nuage pour les mêmes N individus.

- Les variables de X déterminent dans R^N un espace E_X de dim P. Chaque point-individu i, i=1,…,N est repéré dans cet espace par un vecteur

$$\boxed{e_{iX} = (X_{i1},\ldots.,X_{ip})'}$$

- Les variables de Y déterminent dans R^N un espace E_Y de dim Q. Chaque point-individu i, i=1,…,N est repéré dans cet espace par un vecteur

$$\boxed{e_{iY} = (Y_{i1},\ldots.,Y_{iQ})'}$$

But de l'AC

- On cherche à expliquer le groupe de variables Y par le groupe de variables X, ou juste à décrire les ressemblances entre ces deux groupes de variables. Pour cela, on cherche à synthétiser ces ressemblances.

Méthode

- On recherche les combinaisons linéaires (variables canoniques) des variables de X et de Y qui soient les plus corrélées possible.
- On interprète ensuite ces variables canoniques pour donner un sens aux corrélations calculées.

- On visualise ces ressemblances sur des graphiques

ANALYSE CANONIQUE

Description de la méthode

- ✓ On construit

$$\boxed{U_1 = a_{11}X_1 + a_{21}X_2 + \ldots + a_{P1}X_P = Xa_1} \qquad \boxed{V_1 = b_{11}Y_1 + b_{21}Y_2 + \ldots + b_{Q1}Y_Q = Yb_1}$$

des combinaisons linéaires des variables X et Y choisies de telle manière que U1 et V1 soient le plus corrélés possible.

RQ : Les vecteurs a1 et b1 (appelés premiers facteurs) ne sont pas uniques. Pour assurer leur unicité , on impose à U1 et V1 d'être de variance unité.

- ✓ La corrélation r1 entre U1 et V1 est appelée première corrélation canonique

U1 et V1 sont appelées les premières variables canoniques.

- ✓ En général, U1 et V1 n'expliquent pas l'ensemble des liaisons entre les X et les Y. On cherche alors 2 nouvelles variables non corrélées avec U1 et V1 de corrélation maximale (après U1 et V1) et de variance unité.

$$U_2 = a_{12}X_1 + a_{22}X_2 + \ldots + a_{P2}X_P = Xa_2 \qquad V_2 = b_{12}Y_1 + b_{22}Y_2 + \ldots + b_{Q2}Y_Q = Yb_2$$

- ✓ On continue le procédé et on définit ainsi s couples de variables canoniques et une suite de corrélations canoniques décroissantes:

$$\boxed{r_1 \geq r_2 \geq r_3 \geq \ldots \geq r_s} \qquad s = \min(P,Q)$$

ANALYSE CANONIQUE

Rq : Lorsque Q=1, on peut montrer facilement que la première et seule corrélation canonique est égale au coefficient de corrélation multiple entre les variables explicatives et la variable dépendante. Ce cas particulier permet de voir la première corrélation canonique comme une généralisation au cas de plusieurs variables dépendantes de la RLM (sans intercept) de Y sur le tableau X.

Résolution du problème:

- Recherche du premier couple de variables canoniques

$$\boxed{P: \left| \begin{array}{c} r(U_1,V_1) \text{ maximal} \\ Var(U_1) = Var(V_1) = 1 \end{array} \right.}$$

$$\boxed{r(U_1,V_1) = \frac{Cov(U_1,V_1)}{\sigma_{U_1}\sigma_{V_1}} = \frac{U_1'V_1}{\sqrt{U_1'U_1V_1'V_1}} = \frac{\langle U_1,V_1\rangle}{\|U_1\|\|V_1\|} = \langle U_1,V_1\rangle = \cos(U_1,V_1)}$$

Interprétation géométrique : On recherche des directions de Ex et Ey les plus proches possibles (d'angle minimal) , ie tq $Xa_1 \approx Yb_1$

Rappel d'algèbre linéaire

Définition : Le projecteur orthogonal sur l'espace E engendré par les colonnes de X est l'application linéaire qui fait correspondre à u sa projection orthogonale sur E. Ce projecteur s'écrit

$\boxed{Pu = \hat{u}}$

$$P = X(X'X)^{-1}X'.$$

- On note P_X (resp. P_Y) le projecteur orthogonal sur E_X (resp. E_Y)

$$\boxed{P_X = X(X'X)^{-1}X' \quad \text{et} \quad P_Y = Y(Y'Y)^{-1}Y'}$$

$$\boxed{P : \left| \begin{array}{c} \|U_1 - P_Y U_1\| \text{ minimal, } \|V_1 - P_X V_1\| \\ \|U_1\| = \|V_1\| = 1 \end{array} \right.}$$

Solution

- U1 est le vecteur propre associé à la plus grande valeur propre de $P_X P_Y$ commune avec les v.p. de $P_Y P_X$
- V1 est le vecteur propre associé à la plus grande valeur propre de $P_Y P_X$ commune avec celles de $P_X P_Y$
- On montre que $\boxed{\lambda_1 = r^2(U_1, V_1)}$

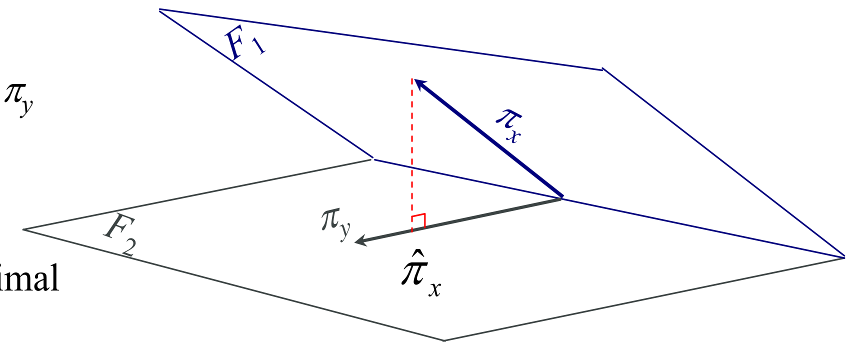
- Principe : on cherche les CL qui minimisent l'angle entre U1 et V1.

  - Soient F_1 et F_2 les sous-espaces vectoriels engendrés par combinaison linéaire des caractéristiques ($x^1, \quad, x^p$) et ($y^1, \quad, y^q$).

$$F_1 = \left\{ \pi_x \in R^n \middle| \pi_x = (x^1,\dots,x^p)\hat{w}_x, \hat{w}_x \in R^p \right\}$$
$$F_2 = \left\{ \pi_y \in R^n \middle| \pi_y = (y^1,\dots,y^q)\hat{w}_y, \hat{w}_y \in R^q \right\}$$

On cherche π_x tel que sa distance avec π_y soit minimale.

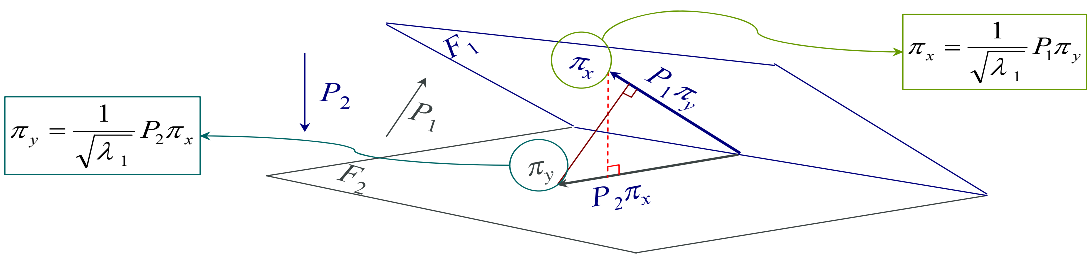
Si $\hat{\pi}_x$ le projeté orthogonal de π_x sur F_2, $\|\pi_x - \hat{\pi}_x\|$ minimal $\Leftrightarrow$ $\hat{\pi}_x$ maximal



- Soit P_1 le projeté orthogonal des éléments de F_2 sur F_1 et P_2 le projeté orthogonal des éléments de F_1 sur F_2.

$$P_1\pi_y = \cos(\pi_x, \pi_y)\pi_x \text{ et } P_2\pi_x = \cos(\pi_x, \pi_y)\,\pi_y \text{ Soit,}$$

$$P_2P_1\pi_y = \cos(\pi_x, \pi_y)\pi_y \text{ et } P_1P_2\pi_x = \cos(\pi_x, \pi_y)\pi_x$$



Les caractères canoniques sont des Vp de P_1P_2 et P_2P_1 rangés dans l'ordre décroissants des vp.

Propriétés :

- $U1=Xa_1$et $V1=Yb_1$: ce sont des CL de X et Y respectivement
- $Var(U1)=Var(V1)=1$
- $\lambda_1=r^2(U1,V1)$

$$\sqrt{\lambda_1}V_1 = P_Y U_{11}, \quad \sqrt{\lambda_1}U_1 = P_X V_{11}$$

K° (k<=s) Couple de variables canoniques (Uk,Vk):

- Uk est le vecteur propre associé à la valeur propre de rang k de $P_X P_Y$ commune avec les v.p. de $P_Y P_X$
- Vk est le vecteur propre associé à la valeur propre de rang k de $P_Y P_X$ commune avec celles de $P_X P_Y$

On a :

- $Uk=Xa_k$ et $Vk=Yb_k$: ce sont des CL de X et Y respectivement
- Var(Uk)=Var(Vk)=1
- $\lambda_k=r^2(Uk,Vk)$

$$\sqrt{\lambda_k} V_k = P_Y U_k , \quad \sqrt{\lambda_k} U_k = P_X V_k$$

Propriétés :

- Les Uk sont non corrélées
- Les Vk sont non corrélées
- Les Uk sont non corrélées aux Vj pour j≠k

Propriétés des facteurs ak et bk :

$$\boxed{\begin{aligned} R_X &= (X'X)^{-1}X'Y(Y'Y)^{-1}Y'X = V_{XX}^{-1}V_{XY}V_{YY}^{-1}V_{YX} \\ R_Y &= V_{YY}^{-1}V_{YX}V_{XX}^{-1}V_{XY} \end{aligned}}$$

- a_k est vecteur propre d'ordre k de R_X avec la vp $r^2(U_k,V_k)$
- b_k est vecteur propre de R_y d'ordre k de R_Y avec la vp $r^2(U_k,V_k)$

Remarque :

L'AC détemine Uk et Vk telles qu'en moyenne les deux variables soient le plus proches possibles :

$$\boxed{\left\| U_k - V_k \right\|^2 = \frac{1}{n}\sum_{i=1}^{n}(U_{ik} - V_{ik})^2}$$

Est minimum. Elles sont telles qu'en moyenne elles soient le plus proches possibles pour les n individus.

En pratique :

- On calcule C et les matrices VXX, VYY, VXY et VYX

- On calcule les facteurs ak et bk qui sont les vecteurs propres associés aux valeurs propres des matrices RX et RY tels que Xak et Ybk soient de norme 1 (si X est normé, il s’agit du vecteur propre unitaire de la matrice).

- On calcule les premiers composantes canoniques Uk=Xak et bk=Ybk

- **<u>Les cas particuliers de l'AC</u>**

- L'AC présente un grand intérêt d'un point de vue théorique car plusieurs techniques statistiques très utilisées en sont des cas particuliers :
- Si Y décrit une seule variable quantitative, l'AC se ramène à :
  – la RLS si X est constitué d'une seule variable quantitative
  – la RLM si X est constitué par plusieurs variables quantitatives
  – le modèle d'analyse de la variance si X est constituée d'une ou plusieurs variables qualitatives.
  – le modèle d'analyse de la covariance si X est un mélange de variables quantitatives et qualitatives.

Pour les 4 méthodes citées ci-dessus, le problème est de maximiser le coefficient de corrélation entre une variable quantitative Y et un ensemble de variables X. C'est donc bien un problème d'AC.

- L'analyse factorielle des correspondances est le cas particulier de l'AC pour lequel X et Y décrivent chacun les modalités d'une variable qualitative.

- L'analyse factorielle discriminante, est le cas particulier de l'AC pour lequel X décrit un ensemble de variables quantitatives et Y une variable qualitative.

- L’AC généralisée, généralisation de l’AC simple à m tableaux de données. Elle généralise l’analyse canonique, l’ACP et l’ACM.

Critiques de l'AC

- L'AC décrit les relations linéaires existant entre 2 ensembles de variables : les premières étapes mettent en évidence les directions de l'espace des variables selon lesquelles les deux ensembles sont les plus proches.
- Mais il est possible que les variables canoniques soient faiblement corrélées aux variables des tableaux X et Y. Donc elles sont dificilement interprétables.

  En effet, les variables d'origine n'interviennent pas dans les calculs de détermination des composantes canoniques, seuls interviennent les projecteurs sur les espaces engendrés par ces variables.

Interprétation :

- En AC, l'interprétation des résultats (corrélations canoniques et variables canoniques) n'est pas aussi simple que dans les autres méthodes d'ADD

- c'est probablement une des raisons qui fait que l'analyse des corrélations canoniques n'a pas été très utilisée par le passé.

- **Choix du nombre de composantes:** on coupe à une chute de la corrélation (comme en ACP)

- **Interprétation des variables canoniques** :

Pour donner un sens aux variables canoniques,

- on calcule la corrélation des nouvelles variables U et V avec les anciennes X et Y (ce sont les coefficients les plus forts qui donnent un sens les variables)

- Repérage des proximités entre variables sur un graphique comme en ACP : On fait figurer sur un même graphique l'ensemble des variables de départ (X et Y).

- On trace un cercle des corrélations
- L'axe correspondant à la k-ième étape est un compromis entre Uk et Vk.

$$\boxed{C_k = \frac{1}{2}(U_k + V_k)}$$

- La variable Xj sur l'axe k a pour coordonnées: r(Xj, Ck)

- Représentation des individus

- La représentation des individus (2 nuages de points) permet de cerner ce qui caractérise les directions pour lesquelles les nuages sont les plus ressemblants possibles, et a contrario voir les individus qui ont un comportement particulier, ie, qui se comportent différemment sur les deux tableaux de données.

- A la k-ième étape, On porte les coordonnées des individus pour chaque tableau (chaque individu est représenté deux fois) relativement aux axes de l'autre espace de l'AC .

- il s'agit de comparer la description des individus donnée par la variable canonique Uk à la description des individus donnée par la variable canonique Vk. Le graphique fait apparaitre les individus pour lesquels les variables canoniques sont proches et ceux pour lesquels elles sont éloignées.

- L'écart résiduel quantifie cet éloignement $|z_{1i}^k - z_{2i}^k|.$
-

ANALYSE CANONIQUE

Il s'agit d'une étude portant sur la liaison entre les caractéristiques de chauffage et la pollution dans différents arrondissements parisiens. On a choisi 8 variables explicatives X: le nombre d'habitations avec chauffage

- urbain,
- collectif au charbon,
- collectif au fuel,
- collectif au gaz,
- individuel au charbon,
- individuel au fuel,
- individuel au gaz,
- et le nombre d'habitations sans chauffage central.

Les 5 variables dépendantes Y retenues sont:

- la concentration de SO2 pendant l'hiver 1966-1967,
- la concentration de SO2 pour l'année 1967,
- la concentration de fumées noires pendant le même hiver,
- cette même concentration pour l'année 1967,
- la concentration d'oxyde de carbonne pendant l'année 1967.

ANALYSE CANONIQUE

La question posée est: "*les variables de chauffage suffisent-elles à rendre compte du phénomène de pollution mesuré par les concentrations retenues ?*"

- Un calcul explicite donne les corrélations canoniques suivantes: r1=0.96,r2= 0.92, r3=0.84, r4=0.78, r5=0.53. Les deux premières valeurs proches de 1.00 montrent la forte liaison qui existe entre les 2 groupes de variables X et Y. On peut donc dire qu'en première approximation, c'est le même phénomène qui est rendu par les 2 groupes de variables, ou encore que les variables de chauffage retenues expliquent l'essentiel de la pollution mesurée par les variables de concentration retenues.

ANALYSE CANONIQUE

- Peut-on aller plus loin qu'un simple constat de bonne liaison entre les 2 groupes de variables ? Il est utile pour cela de calculer les corrélations entre les variables canoniques U1 et V1 et les variables de chaque groupe X et Y (ou de les représenter sur un graphique)

Interprétons U1:

- les corrélations de U1 avec X montrent une opposition entre d'une part les logements à chauffage individuel ou les logements sans chauffage et d'autre part les logements à chauffage collectif (corrélations négatives pour les uns et positives pour les autres);

- les corrélations de U1 avec Y montrent une opposition entre les concentrations de fumée d'une part et d'oxyde de carbonne ou de SO2 d'autre part.

- L'analyse des corrélations de V1 avec X et Y conduit à la même interprétation.

La liaison entre les deux groupes de variables la mieux expliquée (corrélation de 0.96) est donc une liaison du type:

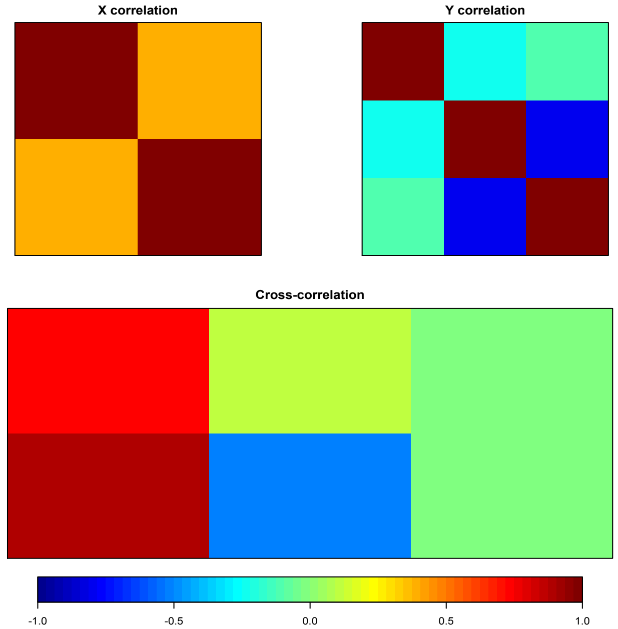
chauffage individuel, sans chauffage ←---------------> collectif,

oxyde de carbonne ←--------------> SO2.

- L'interprétation peut s'affiner en représentant les individus (ici les arrondissements) dans le plan formé par U1 et U2 , ou dans le plan formé par V1 et V2 . On peut ainsi y observer que les arrondissements de la périphérie se portent à gauche du graphique et ceux du centre à droite.

Sous R

package CCA et fda



library(CCA)

- <u>Visualisation des matrices de corrélation</u>:

correl=matcor(X, Y)
img.matcor(correl, type = 2)

Increasing values are translated into colors from blue (negative correlation) to red (positive correlation). If images obtained are uniformly in light green color, this corresponds to quasi-null correlations and the analysis is useless.

- <u>Analyse canonique:</u>

```
res.cc=cc(X,Y)
names(res.cc)
```

  - cor : canonical correlations
  - names: a list containing the names to be used for individuals and variables for graphical outputs
  - xcoef : estimated coefficients for the 'X' variables (facteurs canoniques)
  - ycoef estimated coefficients for the 'Y' variables (facteurs canoniques)
  - scores : a list returned by the internal function comput() containing individuals and variables coordinates on the canonical variates basis.

- <u>Choix du nombre de corrélations canoniques:</u>

```
barplot(res.cc$cor, xlab = "Dimension",ylab = "Canonical correlations", names.arg = 1:10, ylim = c(0,1))
```

Permet de choir le nombre de corrélations canoniques (analogue à l'éboulis des valeurs propres)

- Graphiques:

`plt.cc(res.cc)` dessine les graphqiues

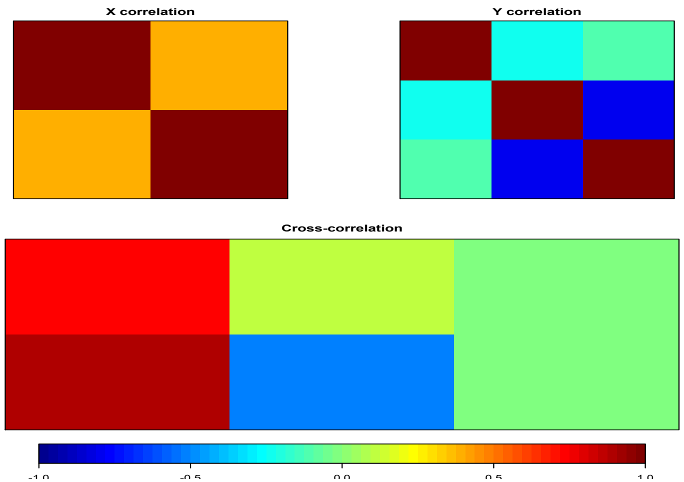
L’AC à la main

On observe les résultats de 5 individus à un premier groupe d’épreuves (2 épreuves contenues dans X), puis à un second (3 épreuves contenues dans Y)

data

```
   X1   X2   Y1   Y2   Y3
1  100  100  200    0 -107
2  200  400  600 -300  212
3 -400 -200 -600 -200  233
4  200 -300 -200  200   92
5 -100    0    0  300 -430
```

```
X=as.matrix(data[,1:2])
Y=as.matrix(data[,3:5])
```




```
Correl=matcor(X,Y)
$Xcor
          X1        X2
X1 1.0000000 0.3938632
X2 0.3938632 1.0000000

$Ycor
           Y1         Y2         Y3
Y1  1.0000000 -0.2192645 -0.1062470
Y2 -0.2192645  1.0000000 -0.7853129
Y3 -0.1062470 -0.7853129  1.0000000

$XYcor
              X1           X2         Y1         Y2            Y3
X1  1.0000000000  0.393863181  0.7454993  0.1153846 -0.0003556671
X2  0.3938631807  1.000000000  0.8981462 -0.5012804 -0.0003311080
Y1  0.7454993164  0.898146239  1.0000000 -0.2192645 -0.1062469927
Y2  0.1153846154 -0.501280412 -0.2192645  1.0000000 -0.7853129298
Y3 -0.0003556671 -0.000331108 -0.1062470 -0.7853129  1.0000000000
```

```
img.matcor(correl, type = 2)
```

Il existe des corrélations élevées entre X et Y donc on peut continuer l’analyse

- Calcul des matrices Rx et Ry

Vxx=correl$Xcor

	X1	X2
X1	1.0000000	0.3938632
X2	0.3938632	1.0000000

Vyy=correl$Ycor

	Y1	Y2	Y3
Y1	1.0000000	-0.2192645	-0.1062470
Y2	-0.2192645	1.0000000	-0.7853129
Y3	-0.1062470	-0.7853129	1.0000000

Vxy=correl$XYcor[1:2,3:5]

	Y1	Y2	Y3
X1	0.7454993	0.1153846	-0.0003556671
X2	0.8981462	-0.5012804	-0.0003311080

Vyx=correl$XYcor[3:5, 1:2]

	X1	X2
Y1	0.7454993164	0.898146239
Y2	0.1153846154	-0.501280412
Y3	-0.0003556671	-0.000331108

Rx=solve(Vxx)%*%Vxy%*%solve(Vyy)%*%Vyx

	X1	X2
X1	0.95447455	0.0219809
X2	0.03353452	0.9838086

Ry=solve(Vyy)%*%Vyx%*%solve(Vxx)%*%Vxy

	Y1	Y2	Y3
Y1	0.99647821	-0.02026026	-0.0004488666
Y2	-0.01005411	0.94196143	-0.0001387071
Y3	0.09757536	0.73766447	-0.0001564495

- <u>Calcul des facteurs et corrélations canoniques</u>

r²(U2,V2)

r²(U1,V1)

```
eigen(Rx)
$values
[1] 1.0000000 0.9382832
```

aa1 aa2

```
$vectors
          [,1]       [,2]
[1,] -0.4347988 -0.8051449
[2,] -0.9005276  0.5930781
```

r²(U2,V2)

r²(U1,V1)

```
eigen(Ry)
$values
[1] 1.000000e+00 9.382832e-01 7.261734e-18
```

bb1 bb2

```
$vectors
           [,1]      [,2]         [,3]
[1,] -0.98490222 0.2640135 0.0004535453
[2,]  0.17054511 0.7447666 0.0001520945
[3,]  0.02969824 0.6128781 0.9999998856
```

```
aa=eigen(Rx)$vectors ; bb=eigen(Ry)$vectors
aa1=as.vector(aa[,1]); aa2=as.vector(aa[,2])
a1=aa1/apply(X%*%aa1,2,sd)
a2=aa2/apply(X%*%aa2,2,sd)
bb1=as.vector(bb[,1])
bb2=as.vector(bb[,2])
 a=cbind(a1,a2); b=cbind(b1,b2)
              a1           a2
[1,] -0.001413319 -0.003916872
[2,] -0.002927177  0.002785209
```

ca=n=cc(X,Y)

ca$cor (corrélations canoniques)

```
[1] 1.0000000 0.9686502
```

ca$xcoef (facteurs canoniques)

```
          [,1]         [,2]
X1 0.001490926  0.003998318
X2 0.002874688 -0.002741835
```

ca$ycoef

```
            [,1]        [,2]
Y1  0.0021271440 0.001464594
Y2 -0.0006461024 0.007247189
Y3 -0.0001040423 0.005514936
```

- **Calcul des composantes canoniques**

Il n’y a que deux couples de composantes canoniques:

U1=-0.0014X1-0.0029X2 V1=-0.002Y1-0.0006Y2-0.0001Y3
U2=-0.0039X1+0.0028X2 V2=0.0014Y1+0.007Y2+0.005Y3

Leurs valeurs sur les individus valent :

```
U1=X%*%a1 ; U2=X%*%a2 ; U=cbind(U1,U2)
          [,1]       [,2]
1 -0.4350496 -0.1251663
2 -1.4535347  0.3007091
3  1.1707632  1.05197069
4  0.5654892 -1.6289369
5 0.1493319  0.3916872

apply(U,2,sd)
[1] 1 1
```

```
U=ca$scores$xscores
       [,1]       [,2]
1  0.4365613  0.1256483
2  1.4480602 -0.2970703
3 -1.1713078 -1.0509604
4 -0.5642212  1.6222141
5 -0.1490926 -0.3998318

apply(ca$scores$xscores,2,sd)
[1] 1 1
```

```
V=ca$scores$yscores
       [,1]       [,2]
1  0.4365613 -0.2971795
2  1.4480602 -0.1262340
3 -1.1713078 -1.0432138
4 -0.5642212  1.6638932
5 -0.1490926 -0.1972659
```

On vérifie : U=X%*%ca$xcoef
V=Y%*%ca$ycoef
cor(U,V)=ca$cor

- Corrélation des variables avec les composantes canoniques

cor(data,U)

```
          [,1]          [,2]
X1 -0.6760638078 -0.6874004601
X2 -0.9435600186  0.3968311016


Y1 -0.9886139788 -0.0847957757
Y2  0.3602704265 -0.5113094232
Y3  0.0003935861  0.0000935488
```

cor(data,V)

ca$scores$corr.X.xscores

```
        [,1]      [,2]

X1 0.6901879  0.7236302
X2 0.9369781 -0.3493881
```

ca$scores$corr.Y.xscores

```
           [,1]        [,2]
Y1  0.9904535925  0.085541282
Y2 -0.3507816541  0.494022843
Y3 -0.0003958636 -0.000113935
```

ca$scores$corr.X.yscores

```
        [,1]      [,2]
X1 0.6901879  0.7009446
X2 0.9369781 -0.3384348
```

ca$scores$corr.Y.yscores

```
           [,1]         [,2]
Y1  0.9904535925  0.0883097766
Y2 -0.3507816541  0.5100116099
Y3 -0.0003958636 -0.0001176224
```

- **Représentation des variables sur les axes**

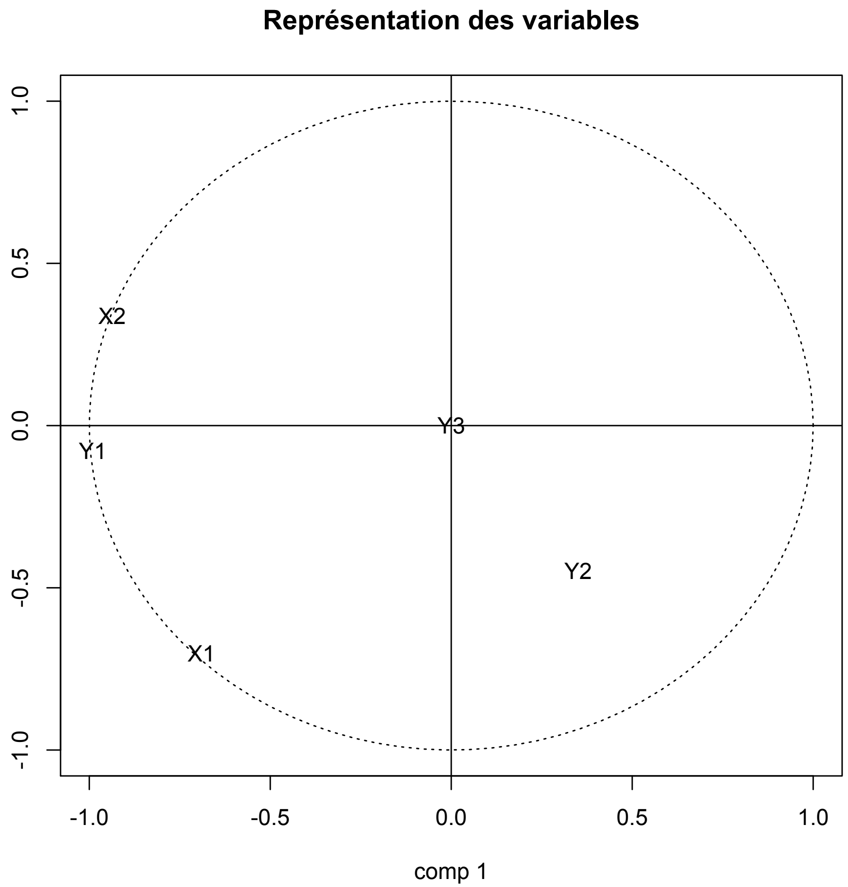
Position des axes :

C1=0.5*(U1+V1); C2=0.5*(U2+V2)

Coordonnées :

```
coord=cbind(cor(data,C1),cor(data,C2))
          [,1]      [,2]
X1 -0.69594109  0.4391335
X2 -0.93381200  0.8985711
Y1 -0.99337552  0.8435690
Y2  0.32807191 -0.4560219
Y3  0.02324971  0.1136613
```

```
a=seq(0,2*pi,length=100)
plot( cos(a), sin(a), type='l', lty=3,xlab='comp 1', ylab='comp 2',
main="Représentation des variables" ) text(coord[,1],coord[,
2],label=row.names(coord))
abline(h=0)
abline(v=0)
```

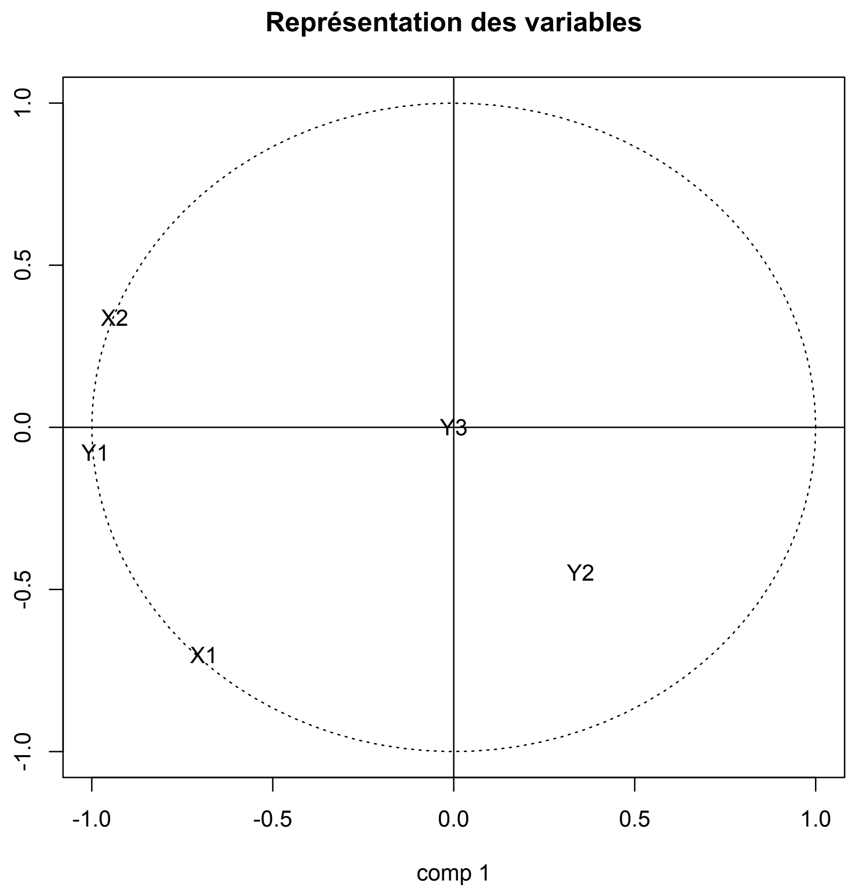


- Y3 a un comportement qui ne peut pas être prévu par aucune des vraibles du tableau X

- Y2 n'est pas assez proche du bord du cercle pour être interprétable (pas assez bien représenté).

- X1 et dans une moindre mesure X2 sont liés à l'axe 1, de même que Y1. En revanche Y2 et Y3 ne le sont pas. Ainsi, l'axe 1 montre une forte corrélation entre la réussite aux deux épreuves de X et celle de Y1. En terme de prévision, cela veut que le resultat à l'épreuve Y1 peut être prédit par les résultats aux épreuves de X.

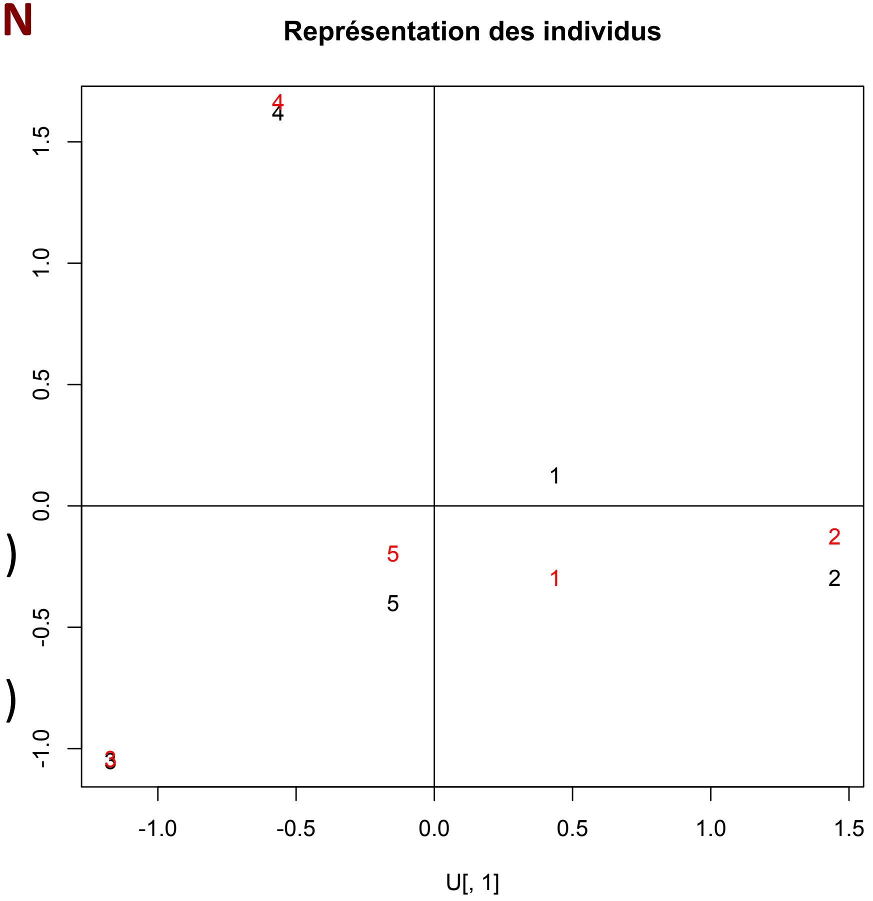
- L'axe 2 isole la réussite à X1 du reste



- **Représentation des Individus (2*N points)**

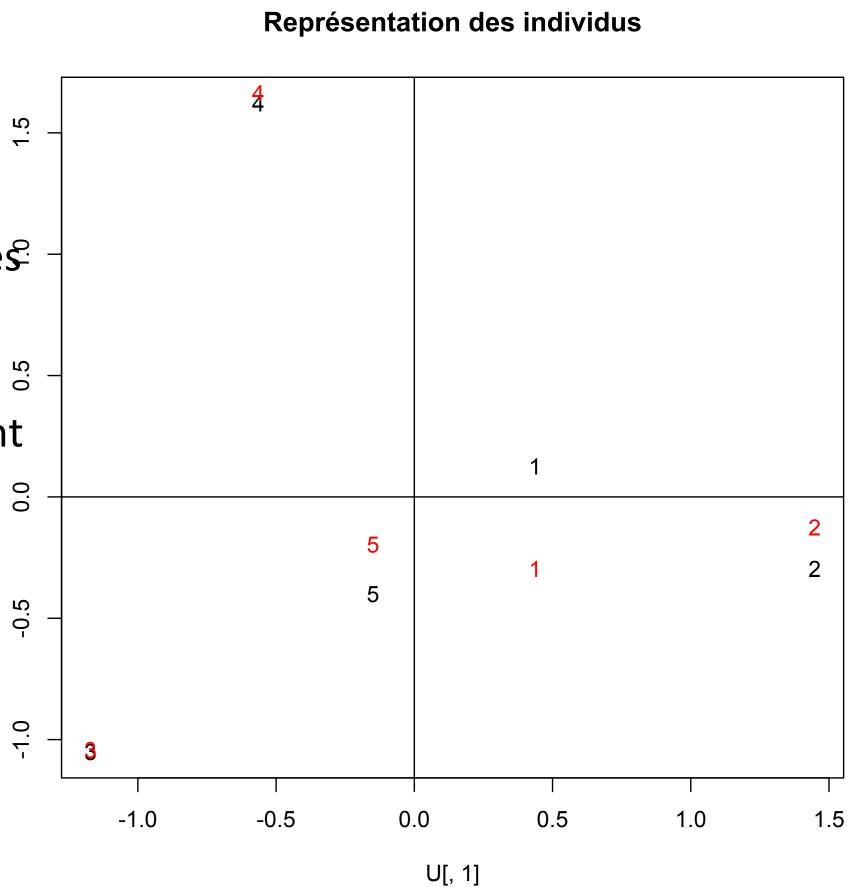
Chaque individu est représenté deux fois: une fois par le couple (Ui1,Ui2), une autre fois par le couple (Vi1,Vi2).

```
plot(U1,U2,col=0,
main="Représentation des individus")
text(U1,U2,label=row.names(X))
text(V1,V2,col=2,label=row.names(Y))
abline(h=0)
abline(v=0)
```

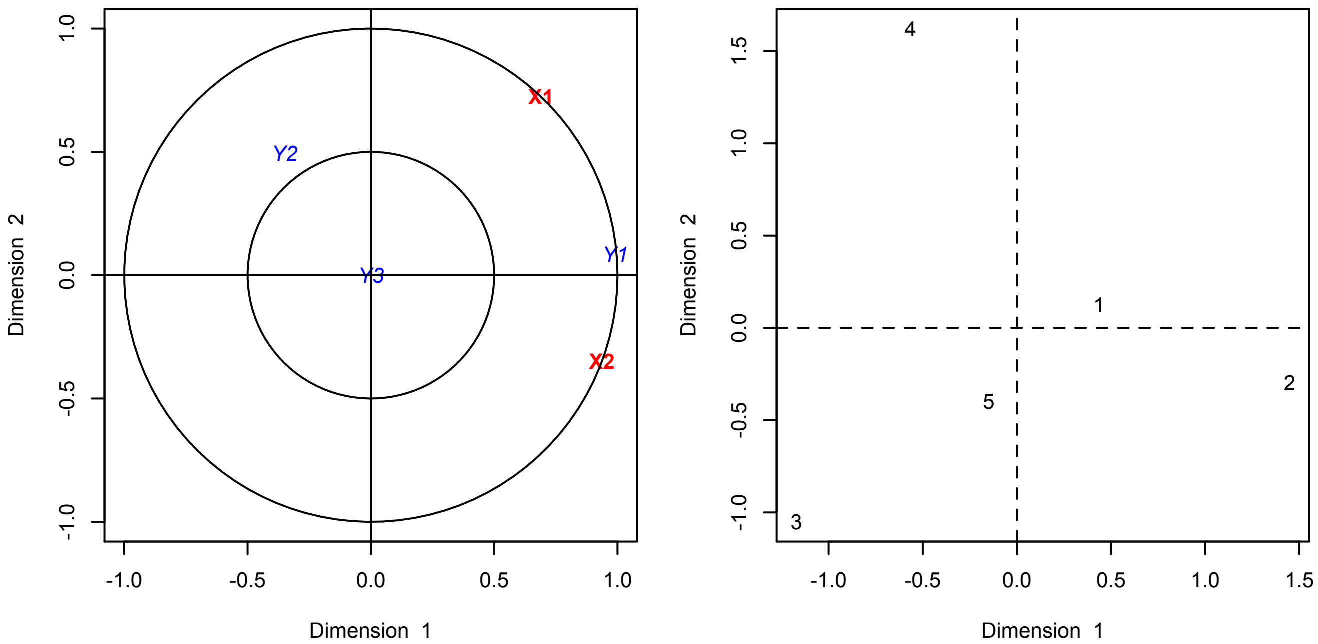


- L’écart résiduel entre les individus sur le premier axe est très faible voir nul, pour les 5 individus : il y une homogénéité des individus qui se comportent tous suivant ce qui a été remarqué sur l’axe 1 (prévisibilité de l’épreuve 1 du deuxième groupe par les épreuves du premier groupe), ce qui confirme la forte cohérence du phénomène mis en évidence su l’axe 1 L’axe oppose les individus 4 et 5 qui ont réussi aux épreuves de X et Y1 à l’individu 2 qui a mal réussi.

- Sur l’axe 2 les écarts sont faibles quoi que plus importants, . 4 se distingue avec une mauvaise réussite à X2 alors que 3 a bien réussi. L’ écart résiduel le plus important concerne l’individu 1.



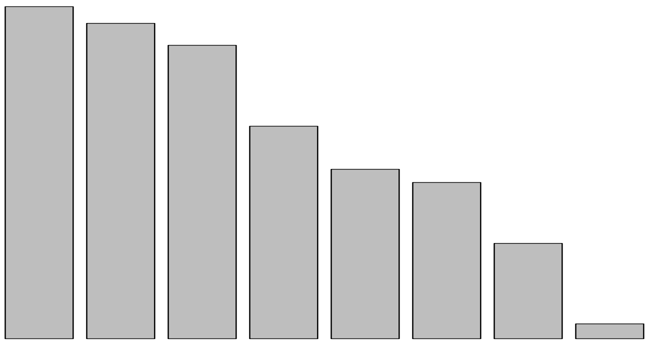
plt.cc(ca, var.label=T, ind.names=rownames(X))



```
CSP=read.table("CSPregions.txt", header=T, row.names=1)
EQ=read.table("Equipregions.txt", header=T, row.names=1)
ca=cc(CSP,EQ)
```

Nombres d'axes :

```
ca$cor
[1] 0.9754993 0.9265434 0.8620452 0.6243216 0.4977799 0.4590493 0.2801003 0.0437050
barplot(ca$cor, xlab = "Dimension",ylab = "Canonical correlations", names.arg = 1:10, ylim = c(0,1))
```

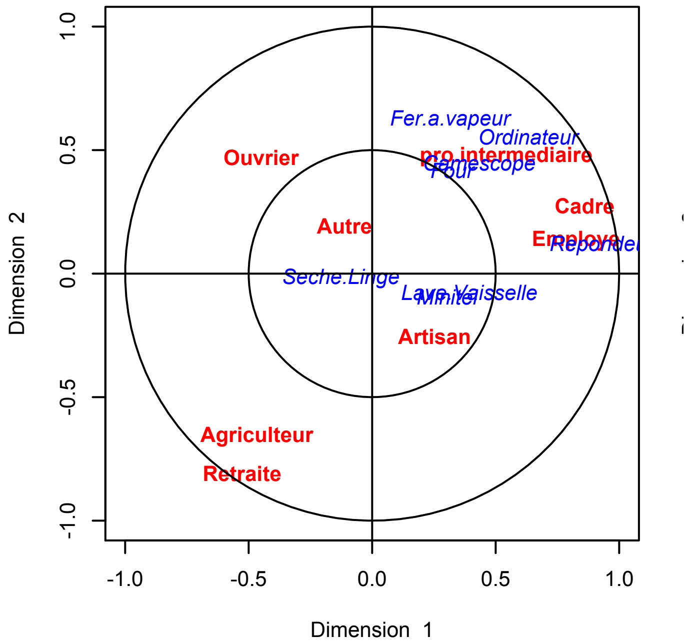


On choisit 3 dimensions

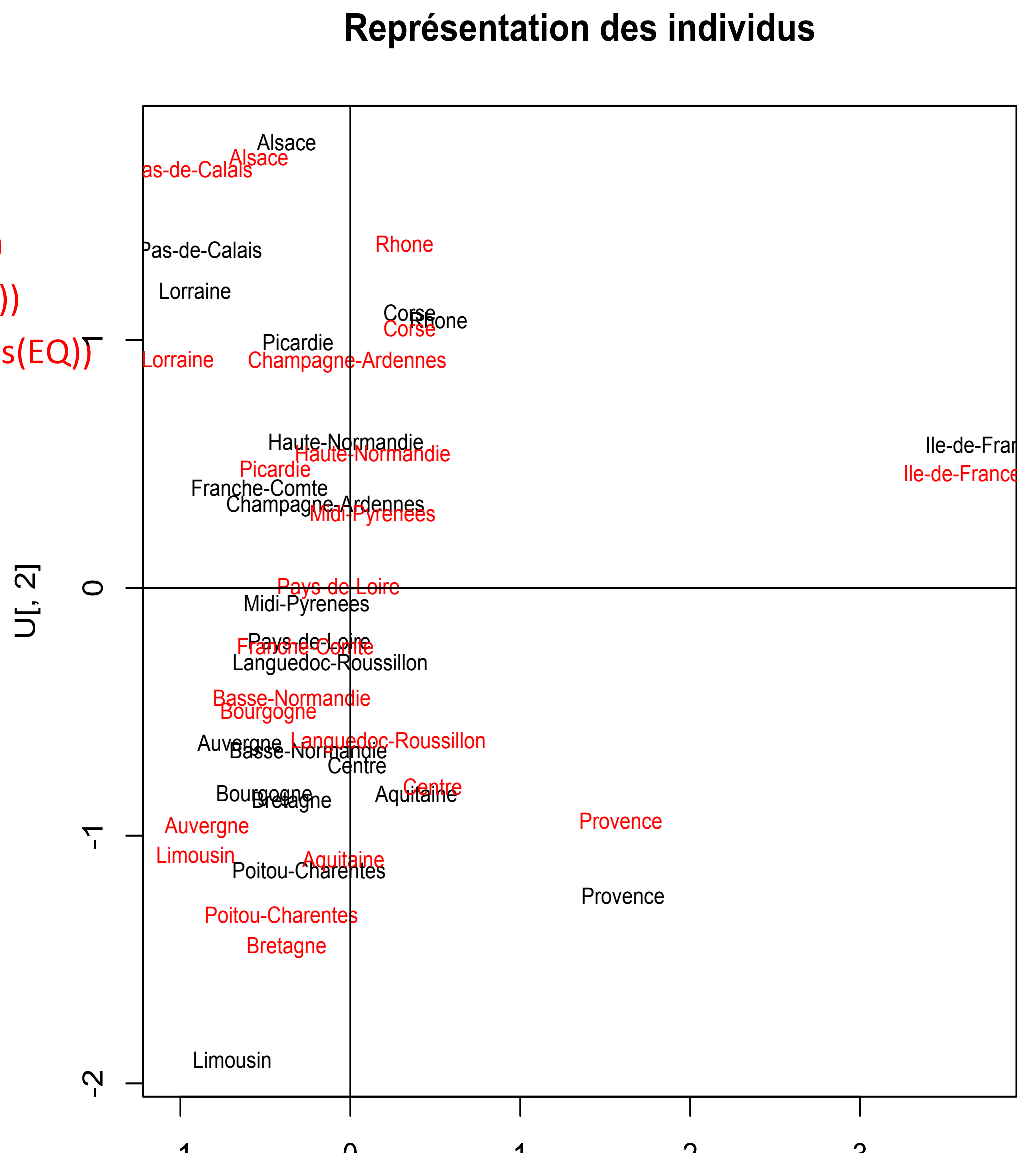
- Corrélations entre les variables canoniques et les variables de départ

```
> ca$scores$corr.X.xscores[,1:2]
                         [,1]       [,2]
Agriculteur        -0.4667456 -0.6596709
Artisan             0.2523507 -0.2537378
Cadre               0.8604098  0.2740927
pro.intermediaire   0.5427713  0.4734119
Employe             0.8253953  0.1338721
Ouvrier            -0.4491671  0.4709990
Retraite           -0.5264663 -0.8095468
Autre              -0.1117938  0.1925093
```

```
> ca$scores$corr.Y.yscores[,1:2]
                      [,1]        [,2]
Lave.Vaisselle   0.4023508 -0.08341658
Four             0.3310922  0.45090244
Minitel          0.3103581 -0.10130615
Seche.Linge     -0.1306188 -0.02221298
Repondeur        0.9529218  0.12274835
Camescope        0.4433023  0.47518447
Ordinateur       0.6469143  0.59656821
Fer.a.vapeur     0.3227889  0.67034431
```



```
U=ca$scores$xscores
 V=ca$scores$yscores
plot(U[,1],U[,2],col=0,
main="Représentation des individus")
text(U[,1],U[,2],label=row.names(CSP))
text(V[,1],V[,2],col=2,label=row.names(EQ))
abline(h=0)
abline(v=0)
```



Il esiste une corrélation forte positivee ntre l'axe 1 et employé cadre et le taux d'équipement en répondeurs

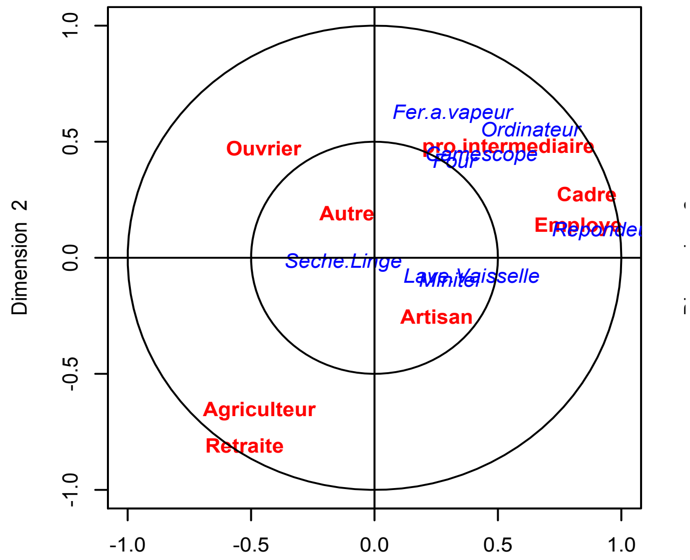
Du point de vue des individus, l'axe 1 oppose le nord et la loraine (peu de cadres, d'ouvriers et de répondeurs) à la provence et l'ile de France où ces quantités sont élevées.

Le coefficient de corrélation canonique étant élevé, Les écarts résiduels sont faibles sur l'ensemble des régions ce qui montre que les groupes-individus sont homogènes

Le deuxième axe oppose les micro-ordinateurs et les professions intermédiaires aux retraités. De point de vue des individus, ces corrélations se traduisent par ne opposition entre l'ile de France et dans une moindre mesure le nord et la lorraine (beaucoup de pi peu de retraités et fort équipement en ordinateurs) au limousin et à l'auvergne.

L'écart résiduel de l'Auvergne est particulièrement élevé. L'examen des composantes canoniques pour cette région (-277 pour U et -406 pour V montre que l'auvergne a un taux d'équipement en ordis particulièrement bas même en tenant compte de sa structure sociale (beaucoup de retraités, peu de pi). La FC a aussi un écart élevé: bien quayant un profil social dans la moyenne elle est peu équipée en ordinateurs.

Notons enfin la faible corrélation entre artisan et ordi, et une opposition entre ouvriers et artisans, correspondant à une opposition entre le nord (peu d'artisans, beaucoup d'ouvriers) et la provence (beaucoup d'artisans, peu d'ouvriers)



Représentation des individus