Fiche TD avec le logiciel R : tdr65

Variables Instrumentales

A.B. Dufour, D. Chessel & J. Thioulouse

La fiche introduit à l'usage de l'ACP sur variables instrumentales et de l'analyse canonique des correspondances ou AFC sur variables instrumentales. Elle donne des points de comparaison avec l'analyse canonique et l'analyse inter-batteries.

Table des matières

1 Analyse canonique et analyse des correspondances

L'analyse canonique s'occupe de deux tableaux de variables quantitatives. On prend par exemple les données `monde84` se trouvant dans la librairie `ade4`.

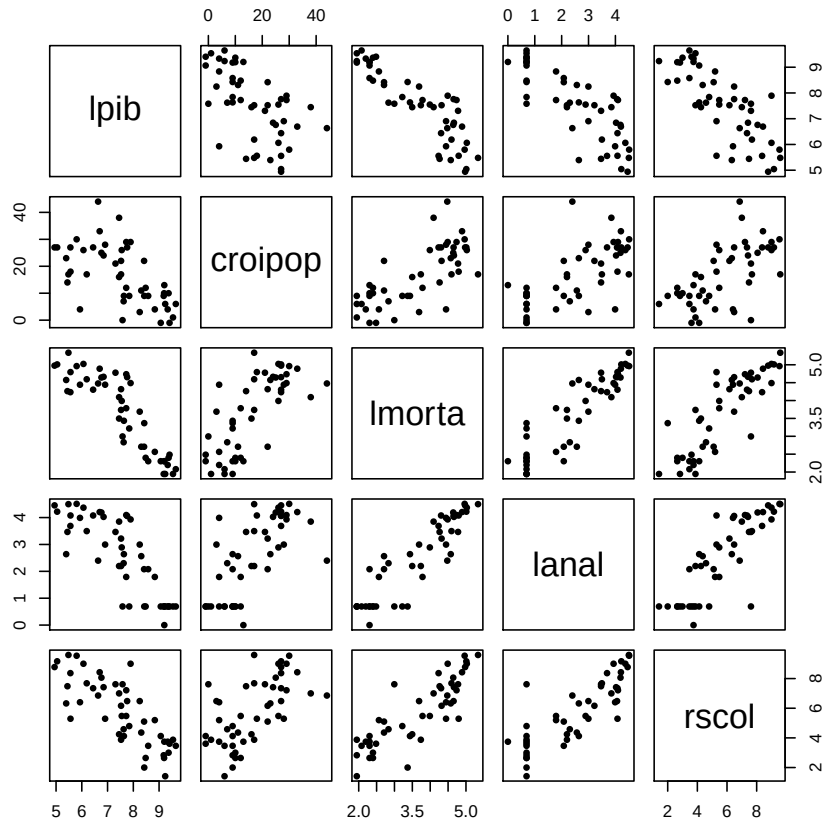
```
library(ade4)
data(monde84)
```

Le data frame contient 48 lignes et cinq colonnes. Ces données brutes font partie d'un ensemble de statistiques publiées dans "L'état du Monde 1984" (Editions La Découverte). Elles regroupent des informations concernant les années 1983, certaines sont associées au dernier recensement :
`pib` Produit intérieur brut par habitant (exprimé en dollars),
`croipop` taux de croissance de la population (exprimé en ‰),
`morta` taux de mortalité infantile, nombre de décés d'enfants âgés de moins d'un

an rapporté au nombre d'enfants nés vivants pendant l'année étudiée (exprimé en ‰),
`anal` Pourcentage d'analphabétisme dans la population des plus de 15 ans, `scol` Pourcentage global d'incription scolaire pour la catégorie des 11 -17 ans (approximative selon les pays)).

On étudie la distribution de chacune des variables et on forme deux tableaux :
X comprenant les variables `lpib`, le logarithme du PIB et `croipop`,
Y comprenant les variables `lmorta` ($log(morta)$), `lanal` ($log(anal + 1)$), et le taux de scolarisation `rscol` ($\sqrt{(}100 - rscol)$).

```
dfX <- cbind.data.frame(lpib = log(monde84$pib), croipop = monde84$croipop)
dfY <- cbind.data.frame(lmorta = log(monde84$morta), lanal = log(monde84$anal+1),
        rscol = sqrt(100-monde84$scol))
pairs(cbind.data.frame(dfX,dfY), pch=20)
```



On veut étudier la relation entre les variables de **X** et celles de **Y**. On regarde les coefficients des régressions multiples de chaque variable de **Y** avec les variables de **X** :

```
dfX0 <- scalewt(dfX)
dfY0 <- scalewt(dfY)
coefficients(lm(dfY0[,1]~-1+dfX0[,1]+dfX0[,2]))
```

```
 dfX0[, 1]  dfX0[, 2]
-0.6796325  0.3331534
```

```
coefficients(lm(dfY0[,2]~-1+dfX0[,1]+dfX0[,2]))
```

```
 dfX0[, 1]  dfX0[, 2]
-0.6171025  0.3321799
```

```
coefficients(lm(dfY0[,3]~-1+dfX0[,1]+dfX0[,2]))
```

```
 dfX0[, 1]  dfX0[, 2]
-0.6393950  0.2483069
```

Exercice. Vérifier la signification statistique de ces coefficients et commenter.

L'analyse canonique a pour but de trouver une combinaison $\mathbf{a}$ des variables de $\mathbf{X}_0$ et une combinaison $\mathbf{b}$ des variables de $\mathbf{Y}_0$ qui maximise :

$$corr^2\left(\mathbf{X}_0\mathbf{a}, \mathbf{Y}_0\mathbf{b}\right)$$

Pour ce faire, on utilise la fonction `cancor` :

```
cancor(dfX0,dfY0)
```

```
$cor
[1] 0.9188854 0.1014687
$xcoef
              [,1]       [,2]
lpib     0.10730757 -0.1426062
croipop -0.05221984 -0.1706591

$ycoef
             [,1]         [,2]       [,3]
lmorta -0.11765816 -0.002982152 -0.3526653
lanal  -0.01672338 -0.245644420  0.2638233
rscol  -0.01301075  0.262880725  0.1202336

$xcenter
       lpib      croipop
1.538121e-16 2.221169e-16

$ycenter
      lmorta         lanal         rscol
 1.950118e-16  2.636057e-16 -9.251859e-18
```

Les deux dernières informations (`xcenter` et `ycenter`) indiquent simplement que les données ont été préalablement centrées.

On sauvegarde les résultats de l'analyse dans un objet de R.

```
can1 <- cancor(dfX0, dfY0)
var(dfX0%*%can1$xcoef)*47
```

```
          [,1]      [,2]
[1,] 1.0000e+00 1.5169e-16
[2,] 1.5169e-16 1.0000e+00
```

```
var(dfY0%*%can1$ycoef)*47
```

```
              [,1]          [,2]          [,3]
[1,]  1.000000e+00 -8.815949e-16  1.059829e-15
[2,] -8.815949e-16  1.000000e+00 -7.503963e-17
[3,]  1.059829e-15 -7.503963e-17  1.000000e+00
```

```
cor(dfY0%*%can1$ycoef[,1],dfX0%*%can1$xcoef[,1])
```

```
          [,1]
[1,] 0.9188854
```

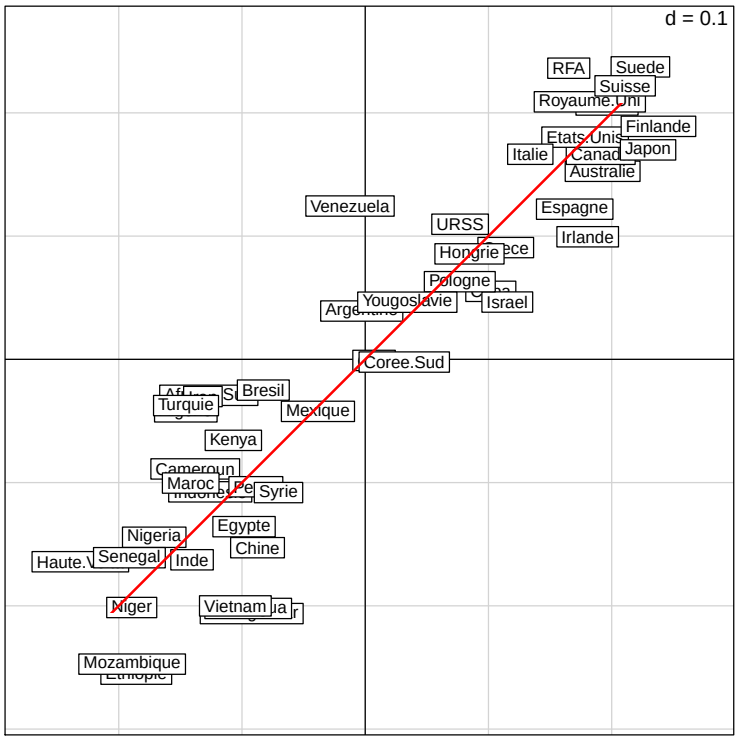
La procédure `cancor` est purement géométrique et utilise la métrique canonique de $\mathbb{R}^n$. Après centrage, ici inutile, les colonnes de $\mathbf{X}_0$ sont deux vecteurs de $\mathbb{R}^n$ et les colonnes de $\mathbf{Y}_0$ sont trois vecteurs de $\mathbb{R}^n$. Leur norme n'a aucun rôle :

```
 cancor(dfX, dfY)
$cor
[1] 0.9188854 0.1014687
$xcoef
                [,1]        [,2]
lpib     0.079908009 -0.10619359
croipop -0.004815451 -0.01573732

$ycoef
               [,1]         [,2]        [,3]
lmorta -0.110784542 -0.002807934 -0.33206248
lanal  -0.011699414 -0.171849016  0.18456670
rscol  -0.005972084  0.120665320  0.05518861

$xcenter
    lpib   croipop
 7.515621 16.666667

$ycenter
  lmorta    lanal    rscol
3.663344 2.551320 5.757566
```

On cherche un vecteur normé dans chacun des sous-espaces $\mathcal{H}$ et $\mathcal{K}$ (donné par les coefficients de la combinaison linéaire) qui optimise le cosinus et définit ainsi l'angle entre les deux sous-espaces par :

$$cos(\mathcal{H}, \mathcal{K}) = sup_{\mathbf{h}\in\mathcal{H},\mathbf{k}\in\mathcal{K}}\,(cos(\mathbf{h}, \mathbf{k}))$$

```
varcanoX <- dfX0%*%can1$xcoef[,1]
varcanoY <- dfY0%*%can1$ycoef[,1]
s.label(cbind(varcanoY,varcanoX),
                lab=row.names(monde84),clab=0.8)
abline(0,1, col="red", lwd=2)
```



```
 cor(varcanoX,dfX)
          lpib    croipop
[1,] 0.9562356 -0.7990498
 cor(varcanoX,dfY)
         lmorta      lanal      rscol
[1,] -0.9160949 -0.8555237 -0.8098218
 cor(varcanoY,dfY)
         lmorta      lanal      rscol
[1,] -0.9969633 -0.9310451 -0.8813089
 cor(varcanoY,dfX)
          lpib    croipop
[1,] 0.8786709 -0.7342351
 cor(dfX,dfY)
           lmorta      lanal      rscol
lpib    -0.875577 -0.8124745 -0.7854370
croipop  0.732880  0.6951294  0.6243677
```

On peut jouer sur la géométrie des 5 vecteurs de $\mathbb{R}^n$. On a des corrélations inter-classes et intraclasses (cosinus d'angles entre vecteurs), des corrélations multiples (cosinus d'angles entre vecteurs et sous-espaces) et des corrélations canoniques (cosinus d'angles entre sous-espaces). L'intérêt statistique de l'analyse canonique est limité mais son intérêt théorique est considérable. En particulier l'analyse des correspondances simple (AFC) est une analyse canonique entre deux paquets d'indicatrices de classes.

Exercice.
Cet exercice, purement mathématique, a pour but de vérifier cette proposition de lien entre AFC et analyse canonique.

1) Ouvrir le data frame `chats` de la librairie `ade4` et placer les effectifs dans un seul vecteur.

2) Transformer les noms des lignes et des colonnes en facteurs.

3) Répéter pour avoir dans deux tableaux les indicatrices de classe avec une ligne par individus.

4) Utiliser la procédure `corresp` de la librairie `MASS` qui prend, en entrée, soit deux variables qualitatives *i.e.* des facteurs,

   soit une table de contingence sous forme matricielle

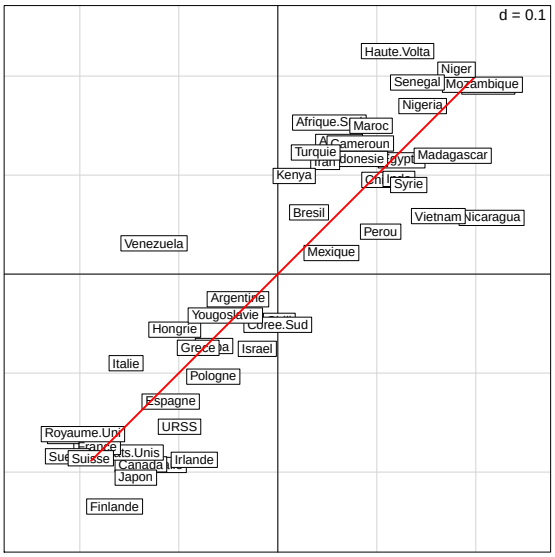
5) Vérifier que l'analyse canonique donne les résultats avec un facteur trivial surnuméraire. expliquer la présence de la corrélation 1.

6) Vérifier la procédure à l'aide de la fonction `dudi.coa` de la librairie `ade4`.

2 Analyse canonique et inter-batteries de Tucker

L'analyse canonique voit dans les deux tableaux deux paquets de variables dans le même espace. L'inter-batteries n'y voit que deux paquets d'individus appariés dans deux espaces différents comme la rotation procuste. On réalise la rotation procustéenne :

```
pro1 <- procuste(dfX0,dfY0)
s.label(cbind(pro1$scorX[,1],pro1$scorY[,1]),
                  lab=row.names(monde84),clab=0.8)
abline(0,1,col="red", lwd=2)
```



L'absence de déformation dans la rotation diminue fortement l'idée de corrélation. Entre analyse canonique et rotation procustéenne, l'analyse de coinertie trouve un compromis. Les axes maximisent le produit du critère optimisé en analyse canonique par le produit des critères optimisés dans les ACP simples. L'analyse inter-batteries est assurée par `coinertia` dans `ade4`.

```
pca1 <- dudi.pca(dfX0,scan=F)
pca2 <- dudi.pca(dfY0,scan=F)
coiner1 <- coinertia(pca1,pca2,scan=F)
summary(coiner1)
```

```
Coinertia analysis
Class: coinertia dudi
Call: coinertia(dudiX = pca1, dudiY = pca2, scannf = F)

Total inertia: 3.454

Eigenvalues:
      Ax1       Ax2
3.4530874 0.0007268

Projected inertia (%):
     Ax1      Ax2
99.97896  0.02104

Cumulative projected inertia (%):
    Ax1   Ax1:2
  99.98  100.00

Eigenvalues decomposition:
          eig      covar      sdX       sdY      corr
1 3.4530873684 1.85824847 1.256250 1.6517621 0.8955301
2 0.0007267686 0.02695865 0.649489 0.4144021 0.1001623

Inertia & coinertia X (pca1):
    inertia      max     ratio
1  1.578164 1.588151 0.9937116
12 2.000000 2.000000 1.0000000

Inertia & coinertia Y (pca2):
    inertia      max     ratio
1  2.728318 2.733045 0.9982705
```

```
12 2.900047 2.907450 0.9974538

RV:
 0.7682401
 cor(pca1$li[,1],pca2$li[,1])
[1] -0.8873048
 cor(dfY$lmorta,pca1$li[,1])
[1] 0.9025028
```

L'analyse inter-batteries, en tenant compte de la structure des nuages individus est très proche de l'analyse canonique sur les tableaux des premières coordonnées factorielles des ACP de départ. Seule une nuance dans l'intepétation la différencie d'une rotation procustéenne. *La morale est qu'on ne peut tout avoir en même temps (le beurre, ...)* (Chessel, 2003)

3 Composantes principales et variables instrumentales

Il y a encore une autre manière de voir le lien entre $\mathbf{X}$ et $\mathbf{Y}$, peut-être la plus complexe et la plus nuancée. Une première approche simple des ACPVI consiste à penser que chaque variable de $\mathbf{Y}$ a un modèle à partir des variables de $\mathbf{X}$. Ces modèles ont une composante principale qui est un modèle des modèles donc un modèle commun des données.

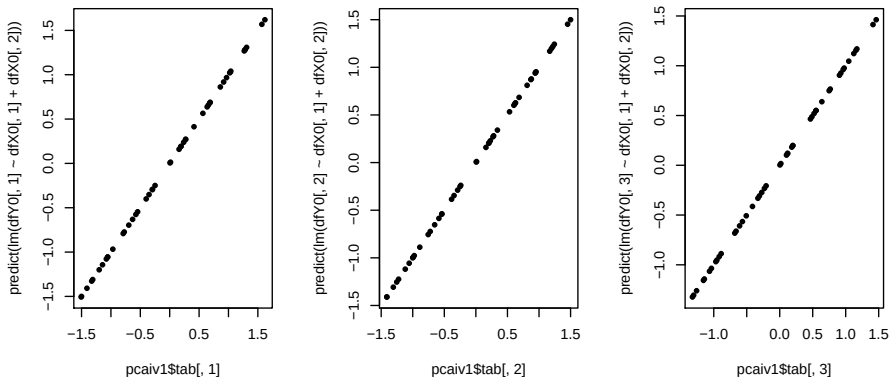
```
 pcaY <- dudi.pca(dfY, scannf=FALSE, nf=2)
 pcaiv1 <- pcaiv(pcaY, dfXO, scannf=FALSE, nf=2)
 pcaiv1
Principal Component Analysis with Instrumental Variables
call: pcaiv(dudi = pcaY, df = dfXO, scannf = FALSE, nf = 2)
class: pcaiv dudi
$rank (rank)     : 2
$nf (axis saved) : 2

eigen values: 2.227 0.001723

 vector length mode    content
 $eig   2      numeric eigen values
 $lw    48     numeric row weigths (from dudi)
 $cw    3      numeric col weigths (from dudi)

 data.frame nrow ncol content
 $Y         48   3    Dependant variables
 $X         48   2    Explanatory variables
 $tab       48   3    modified array (projected variables)

 data.frame nrow ncol content
 $c1        3    2    PPA Pseudo Principal Axes
 $as        2    2    Principal axis of dudi$tab on PAP
 $ls        48   2    projection of lines of dudi$tab on PPA
 $li        48   2    $ls predicted by X

 data.frame nrow ncol content
 $fa        3    2    Loadings (CPC as linear combinations of X
 $l1        48   2    CPC Constraint Principal Components
 $co        3    2    inner product CPC - Y
 $cor       2    2    correlation CPC - X
 summary(pcaiv1)
Principal component analysis with instrumental variables
Class: pcaiv dudi
Call: pcaiv(dudi = pcaY, df = dfXO, scannf = FALSE, nf = 2)

Total inertia: 2.229

Eigenvalues:
     Ax1      Ax2
```

```
2.227037 0.001723

Projected inertia (%):
    Ax1     Ax2
99.9227  0.0773

Cumulative projected inertia (%):
    Ax1   Ax1:2
  99.92  100.00

Total unconstrained inertia (pcaY): 3

Inertia of pcaY explained by dfX0 (%): 74.29

Decomposition per axis:
   iner inercum inerC inercumC ratio     R2  lambda
1 2.733    2.73 2.729     2.73 0.998 0.8161 2.22704
2 0.174    2.91 0.171     2.90 0.997 0.0101 0.00172
```

Cette procédure est dissymétrique. `cancor` prend deux tableaux. `procuste` prend deux tableaux. `coinertia` prend deux triplets. `pcaiv` prend un triplet et un tableau. Le tableau est celui des variables instrumentales. On suppose simplement que si $x_1, x_2, \ldots, x_p$ sont les variables de ce tableau, les modèles du type y en fonction de $x_1 + x_2 + \cdots + x_p$ ont un sens. Les x_i peuvent être qualitatives, quantitatives ou à modalités ordonnées.

L'ACPVI fait l'analyse du tableau projeté :

```
par(mfrow=c(1,3))
plot(pcaiv1$tab[,1],predict(lm(dfY0[,1]~dfX0[,1]+dfX0[,2])), pch=20)
plot(pcaiv1$tab[,2],predict(lm(dfY0[,2]~dfX0[,1]+dfX0[,2])), pch=20)
plot(pcaiv1$tab[,3],predict(lm(dfY0[,3]~dfX0[,1]+dfX0[,2])), pch=20)
```

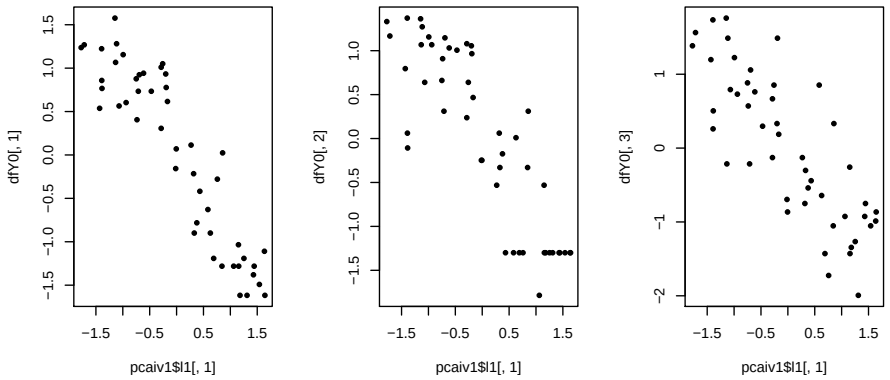


Les trois modèles jouent un rôle équivalent :

```
pcaiv1$c1
             CS1        CS2
lmorta -0.6138547 -0.1627103
lanal  -0.5731982 -0.5809546
rscol  -0.5427949  0.7975068
```

On a trouvé un prédicteur commun aux trois variables :

```
par(mfrow=c(1,3))
plot(pcaiv1$l1[,1],dfY0[,1], pch=20)
plot(pcaiv1$l1[,1],dfY0[,2], pch=20)
plot(pcaiv1$l1[,1],dfY0[,3], pch=20)
```



Ce prédicteur est une combinaison des variables de $\mathbf{X}$:

```
summary(lm(pcaiv1$l1[,1]~dfX0[,1]+dfX0[,2]))
Call:
lm(formula = pcaiv1$l1[, 1] ~ dfX0[, 1] + dfX0[, 2])
Residuals:
       Min         1Q     Median         3Q        Max
-4.745e-16 -1.106e-16 -3.520e-18  8.408e-17  1.384e-15

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) -8.905e-17  3.844e-17 -2.317e+00   0.0251 *
dfX0[, 1]    7.492e-01  4.753e-17  1.576e+16   <2e-16 ***
dfX0[, 2]   -3.549e-01  4.753e-17 -7.469e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.663e-16 on 45 degrees of freedom
Multiple R-squared:      1,        Adjusted R-squared:      1
F-statistic: 3.384e+32 on 2 and 45 DF,  p-value: < 2.2e-16
```

Ce prédicteur est une composante principale sous contrainte c'est-à-dire qu'elle appartient au sous-espace engendré par $\mathbf{X}_0$ (la contrainte) et maximise la somme des carrés des corrélations avec les variables de $\mathbf{Y}$ (composante principale) :

```
sum(cor(pcaiv1$l1[,1],dfY0)^2)
[1] 2.227037
pcaiv1$eig
[1] 2.227036523 0.001722847
```

En analyse canonique, on maximise

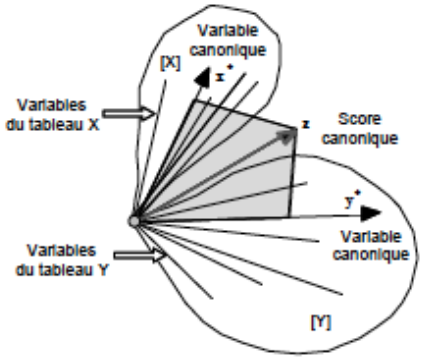
$$cos(\mathcal{H}, \mathcal{K}) = sup_{\mathbf{h}\in\mathcal{H},\mathbf{k}\in\mathcal{K}} \left(cos(\mathbf{h}, \mathbf{k})\right)$$

quand les vecteurs sont des variables normées.

En ACPVI, on maximise

$$\sum_{\mathbf{k}\in\mathcal{K}} corr^2_{\mathbf{h}\in\mathcal{H}}(\mathbf{h}, \mathbf{k})$$

En résumé, la rotation procustéenne envoie un nuage sur l'autre et est proche de l'analyse de coinertie qui étudie deux nuages de points dans une optique de couplage d'axes principaux. L'analyse canonique est symétrique et travaille sur le nuage des variables.



L'ACPVI s'apparente à une régression simultanée des varuables de $\mathbf{Y}$ par une variable de $\mathbf{X}$ adaptée. Ici, la nature des données (le sens du problème) impose la dernière comme pertinente.

4 Interprétation d'une ACPVI - ACP normée

```
 pcaiv1
Principal Component Analysis with Instrumental Variables
call: pcaiv(dudi = pcaY, df = dfXO, scannf = FALSE, nf = 2)
class: pcaiv dudi
$rank (rank)     : 2
$nf (axis saved) : 2

eigen values: 2.227 0.001723

 vector length mode    content
 $eig   2      numeric eigen values
 $lw    48     numeric row weigths (from dudi)
 $cw    3      numeric col weigths (from dudi)

 data.frame nrow ncol content
 $Y         48   3    Dependant variables
 $X         48   2    Explanatory variables
 $tab       48   3    modified array (projected variables)

 data.frame nrow ncol content
 $c1        3    2    PPA Pseudo Principal Axes
 $as        2    2    Principal axis of dudi$tab on PAP
 $ls        48   2    projection of lines of dudi$tab on PPA
 $li        48   2    $ls predicted by X

 data.frame nrow ncol content
 $fa        3    2    Loadings (CPC as linear combinations of X
 $l1        48   2    CPC Constraint Principal Components
 $co        3    2    inner product CPC - Y
 $cor       2    2    correlation CPC - X
```

`tab` : le tableau $\mathbf{P_X Y}$ des modèles linéaires des colonnes de $\mathbf{Y}$ par $\mathbf{X}$.

`cw` : le poids des colonnes (1 pour chacune des m colonnes de $\mathbf{Y}$).

`lw` : le poids des lignes ($1/n$ pour chacune des n colonnes de $\mathbf{Y}$).

`l1` : les composantes principales sous contrainte ou CPC (combinaisons linéaires de variables de $\mathbf{X}$ maximisant le critère de l'analyse).

```
var(pcaiv1$l1)*47/48
```

```
             RS1          RS2
RS1 1.000000e+00 1.158508e-15
RS2 1.158508e-15 1.000000e+00
```

`co` : les corrélations entre les CPC et les variables de $\mathbf{Y}$.

```
cor(dfY,pcaiv1$l1)
```

```
              RS1          RS2
lmorta -0.9160712 -0.006753649
lanal  -0.8553985 -0.024113791
rscol  -0.8100269  0.033102265
```

```
pcaiv1$co
```

```
            Comp1        Comp2
lmorta -0.9160712 -0.006753649
lanal  -0.8553985 -0.024113791
rscol  -0.8100269  0.033102265
```

`eig` : les valeurs propres, optimum du critère "somme des carrés des corrélations entre CPC et variables de $\mathbf{Y}$".

```
sum(pcaiv1$co[,1]^2)
```

```
[1] 2.227037
```

```
pcaiv1$eig[1]
```

```
[1] 2.227037
```

Quand on interprète l'analyse avec ce point de vue, on fait une analyse en composantes explicatives [1].
Mais dans un schéma de dualité, il y a toujours deux points de vue. Le second est formé :

`c1` : les pseudo-axes principaux ou PAP, vecteurs normés de $\mathbb{R}^n$.

```
t(as.matrix(pcaiv1$c1))%*%as.matrix(pcaiv1$c1)
```

```
              CS1           CS2
CS1  1.000000e+00 -5.551115e-17
CS2 -5.551115e-17  1.000000e+00
```

`ls` : les coordonnées des projections des lignes de $\mathbf{Y}$ sur les PAP.

```
(as.matrix(pcaY$tab)%*%as.matrix(pcaiv1$c1))[1:3,]
```

```
            CS1        CS2
[1,] -1.8384468  0.4994283
[2,] -1.6010038 -0.2594040
[3,]  0.3045826  0.1875690
```

```
pcaiv1$ls[1:3,]
```

```
      Axis1      Axis2
1 -1.8384468  0.4994283
2 -1.6010038 -0.2594040
3  0.3045826  0.1875690
```

`as` : les coordonnées des projections des axes principaux (AP) de $\mathbf{Y}$ sur les PAP. Ceci permet de comparer les AP et les PAP. Les PAP ont une propriété d'optimalité originale.

`li` : les prédictions des coordonnées des projections des lignes de $\mathbf{Y}$ sur les PAP par régressions multiples sur $\mathbf{X}$. Ces regressions définissent des carrés de corrélation multiple ou pourcentage de variance expliquée. Les PAP optimisent la variance expliquée c'est-à-dire le produit de la variance de la projection (critère

d'ACP) par le carré de corrélation multiple (critère d'analyse canonique). On peut superposer `ls` (projections sur les PAP) et les `li` (prédictions des positions).

```
lmprovi <- lm(pcaiv1$ls[,1]~dfX0[,1]+dfX0[,2])
predict(lmprovi)[1:5]
```

```
        1         2         3         4         5
-0.2877780 -0.4274749  0.4021648  1.5848058 -0.2530668
```

```
pcaiv1$li[1:5,1]
```

```
[1] -0.2877780 -0.4274749  0.4021648  1.5848058 -0.2530668
```

```
sum(predict(lmprovi)^2)/48
```

```
[1] 2.227037
```

```
summary(lmprovi)
```

```
Call:
lm(formula = pcaiv1$ls[, 1] ~ dfX0[, 1] + dfX0[, 2])
Residuals:
     Min       1Q   Median       3Q      Max
-1.6511 -0.4881  0.0203  0.5566  1.3871

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.296e-16  1.056e-01   0.000 1.000000
dfX0[, 1]    1.118e+00  1.306e-01   8.563  5.3e-11 ***
dfX0[, 2]   -5.297e-01  1.306e-01  -4.057 0.000195 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7315 on 45 degrees of freedom
Multiple R-squared:  0.8161,        Adjusted R-squared:  0.808
F-statistic: 99.88 on 2 and 45 DF,  p-value: < 2.2e-16
```

Quand on interprète l'analyse avec ce point de vue, on fait une analyse en composantes principales sur variables instrumentales au sens de Rao [2].

> *When a large number of measurements are available, it is natural to enquire whether they could be replaced by a fewer number of the measurements or of their functions,* ***without loss of much information****, for convenience in the analysis and in the intepretation of the data. Principal components, which are linear functions of the measurements, are suggested for this purpose. It is, therefore, relevant to examine in what sense principal components provide a reduction of the data without much loss of* ***information we are seeking from the data****.*

En ce sens, on peut dire qu'on a résumé les critères mortalité infantile, analphabétisme et scolarisation en un seul critère en conservant le maximum d'information prédictible par le PIB et le taux de croissance de la population.

Le dernier élément de sortie de la fonction donne une indcation sur le lien entre l'analyse de départ et l'analyse sous contrainte. Le premier PAP qui optimise :

$$Var_{expliquee} = Var \times R^2_{\mathbf{z},\mathbf{X}}$$

est inférieur au premier axe principal du point de vue de la variance projetée. De même, le plan des deux premiers PAP est moins bon que le plan des deux premiers AP, *etc.* La composante ratio donne les taux d'inertie conservés par rapport à l'optimum :

```
summary(pcaiv1)
```

```
Principal component analysis with instrumental variables
Class: pcaiv dudi
Call: pcaiv(dudi = pcaY, df = dfX0, scannf = FALSE, nf = 2)

Total inertia: 2.229

Eigenvalues:
     Ax1      Ax2
2.227037 0.001723

Projected inertia (%):
    Ax1     Ax2
99.9227  0.0773

Cumulative projected inertia (%):
    Ax1   Ax1:2
  99.92  100.00

Total unconstrained inertia (pcaY): 3

Inertia of pcaY explained by dfX0 (%): 74.29

Decomposition per axis:
   iner inercum inerC inercumC ratio     R2  lambda
1 2.733    2.73 2.729     2.73 0.998 0.8161 2.22704
2 0.174    2.91 0.171     2.90 0.997 0.0101 0.00172
```

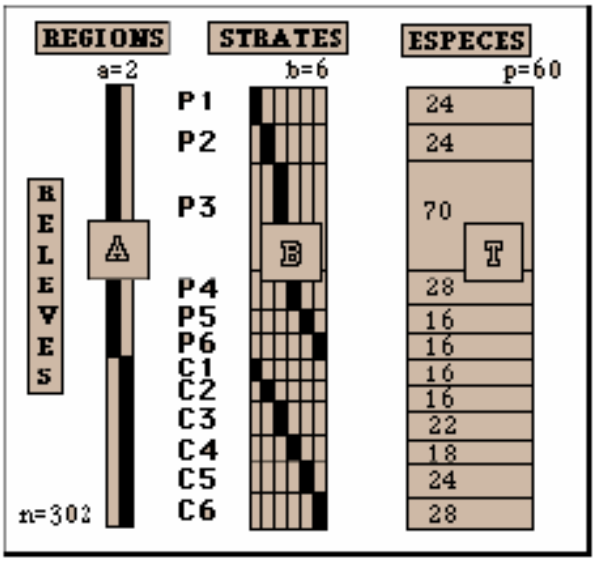
Donc la première composante principale simple était pratiquement la solution de l'ACPI :

```
cor(pcaY$li,pcaiv1$ls)
```

```
            Axis1       Axis2
Axis1 0.999960053 -0.08311219
Axis2 0.007123982  0.98249780
```

5 Analyse des correspondances sur variables instrumentales

Quand le schéma d'entrée dans une `pcaiv` est une analyse des correspondances, on parle d'AFCVI dont l'usage est très répandue en écologie sous le nom de "Canonical Correspondence Analysis" [4].
On prend par exemple les données *avimedi*. On a 302 relevés d'avifaune répartis dans deux régions (Provence et Corse) et 6 types de végétation ordonnés sur un gradient d'ouverture de la pelouse S1 à la forêt haute S6.



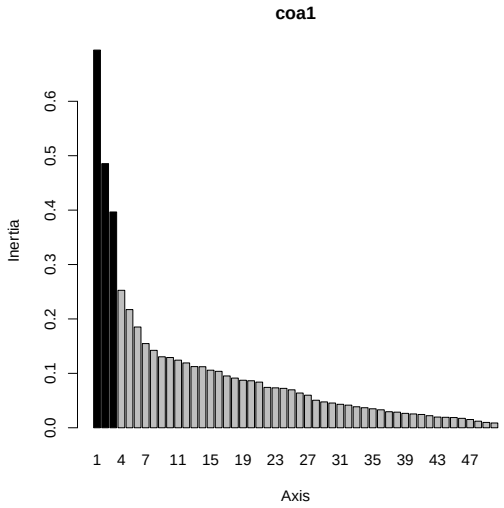

```
data(avimedi)
summary(avimedi$plan)
```

```
reg       str
Pr:178    S1:40
Co:124    S2:40
          S3:92
          S4:46
          S5:40
          S6:44
```

```
dim(avimedi$fau)
```

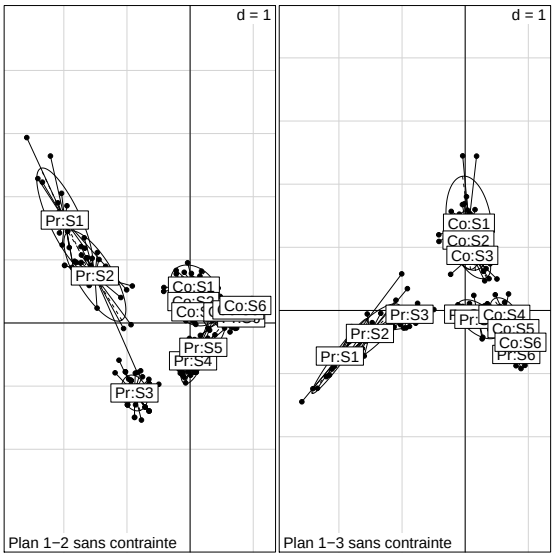
```
[1] 302  51
```

```
coa1 <- dudi.coa(avimedi$fau, scannf=FALSE, nf=3)
screeplot(coa1)
```



La première étape est d'interpréter l'analyse simple.

```
par(mfrow=c(1,2))
cla <- avimedi$plan$reg:avimedi$plan$str
s.class(coa1$li, cla ,sub="Plan 1-2 sans contrainte")
s.class(coa1$li,cla, xax=1, yax=3, sub="Plan 1-3 sans contrainte")
```




```
interAB <- pcaiv(coa1, cla, scannf=FALSE, nf=3)
bca(coa1, cla, scannf=FALSE, nf=3)$eig
```

```
[1] 0.66721649 0.43905367 0.35372966 0.21882347 0.16309322 0.05455513 0.04061491
[8] 0.03653980 0.02484898 0.01536691 0.01021233
```

```
interAB$eig
```

```
[1] 0.66721649 0.43905367 0.35372966 0.21882347 0.16309322 0.05455513 0.04061491
[8] 0.03653980 0.02484898 0.01536691 0.01021233
```

L'analyse inter-classes est une ACPVI sur une variable qualitative. On a, plus généralement, le moyen de décomposer l'inertie dans un plan d'expérience [3].

```
pcaiv(coa1, model.matrix(~avimedi$plan$reg:avimedi$plan$str), scannf=FALSE, nf=3)$eig
```

```
[1] 0.66721649 0.43905367 0.35372966 0.21882347 0.16309322 0.05455513 0.04061491
[8] 0.03653980 0.02484898 0.01536691 0.01021233
```

La structure du tableau est exclusivement une structure inter-classes :

```
summary(interAB)
```

```
Principal component analysis with instrumental variables
Class: pcaiv dudi
Call: pcaiv(dudi = coa1, df = cla, scannf = FALSE, nf = 3)

Total inertia: 2.024

Eigenvalues:
    Ax1     Ax2     Ax3     Ax4     Ax5
 0.6672  0.4391  0.3537  0.2188  0.1631

Projected inertia (%):
    Ax1     Ax2     Ax3     Ax4     Ax5
 32.964  21.692  17.476  10.811   8.058

Cumulative projected inertia (%):
    Ax1   Ax1:2   Ax1:3   Ax1:4   Ax1:5
  32.96   54.66   72.13   82.94   91.00

(Only 5 dimensions (out of 11) are shown)
```

```
Total unconstrained inertia (coa1): 4.971

Inertia of coa1 explained by cla (%): 40.71

Decomposition per axis:
   iner inercum inerC inercumC ratio    R2 lambda
1 0.694   0.694 0.693    0.693 0.999 0.963  0.667
2 0.485   1.179 0.484    1.177 0.998 0.908  0.439
3 0.397   1.576 0.392    1.569 0.995 0.902  0.354
```

```
sum(coa1$eig)
```

```
[1] 4.971324
```

```
sum(interAB$eig)
```

```
[1] 2.024055
```

```
2.024/4.971
```

```
[1] 0.4071615
```

On dit que l'inter-classe prend en compte 40.7% d'inertie. Mais l'inertie est faite de structures et d'aléas.

```
inertia.dudi(coa1)$TOT[1:3,]
```

```
   inertia       cum     ratio
1 0.6941041 0.6941041 0.1396216
2 0.4853867 1.1794908 0.2372589
3 0.3965725 1.5760633 0.3170309
```

```
inertia.dudi(interAB)$TOT[1:3,]
```

```
   inertia       cum     ratio
1 0.6672165 0.6672165 0.3296435
2 0.4390537 1.1062702 0.5465614
3 0.3537297 1.4599998 0.7213243
```

On observe donc que l'inertie inter-classe représente 40.7% de l'inertie totale mais qu'elle contient 72% de structure alors que l'inertie totale contient environ 32% de structure.

On peut vérifier que l'ordination directe et l'ordination sous contrainte sont très voisines.

```
cor(coa1$li,interAB$li)
```

```
           Axis1       Axis2      Axis3
Axis1  0.9789911  0.15511418  0.1674556
Axis2 -0.0873809 -0.94786871 -0.1042121
Axis3  0.1522985  0.09049953  0.9268011
```

Dans cet exemple, il est équivalent :

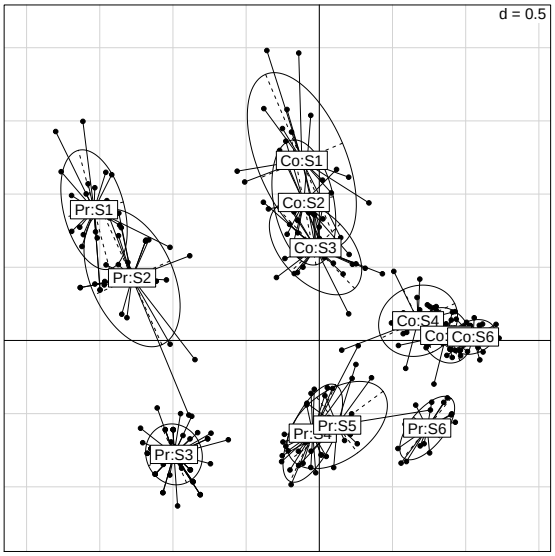
1. soit d'extraire du tableau des données sa structure directement,
2. soit d'extraire par une contrainte du type A :B une part de l'information puis une partie de cette part comme structure.

On a un outil performant de modélisation dans un contexte massivement multivarié.

```
interAplusB <- pcaiv(coa1, avimedi$plan, scannf=FALSE, nf=4)
inertia.dudi(interAplusB)
```

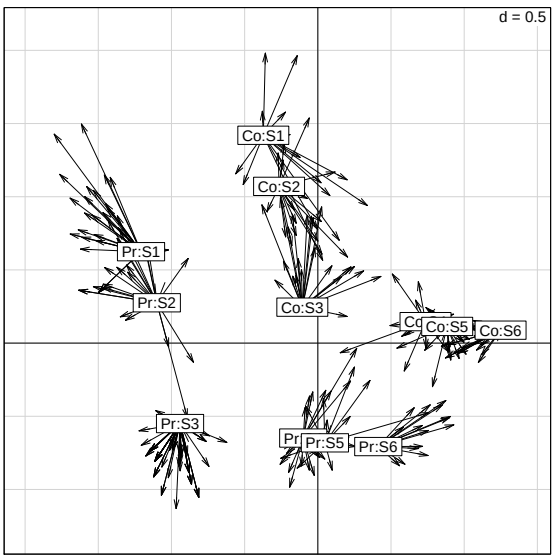
```
$TOT
     inertia       cum     ratio
1 0.61967550 0.6196755 0.4476468
2 0.36330481 0.9829803 0.7100943
3 0.23443992 1.2174202 0.8794511
4 0.10992219 1.3273424 0.9588577
5 0.03370855 1.3610510 0.9832084
6 0.02324455 1.3842955 1.0000000
```

```
s.class(interAplusB$ls,cla)
```



Le graphique est explicite. On a cherché une ordination à effet additif. Il vaut mieux, cependant, représenter une projection stricte sur les PAP (composantes `ls`) et son modèle (composante `li`) :

```
s.match(interAplusB$li,interAplusB$ls,clab=0)
s.class(interAplusB$li,cla,add.plot=T)
```

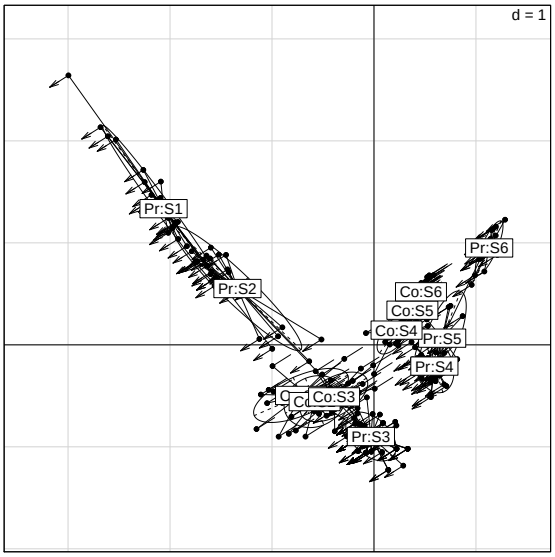


On a disposé les espèces dans un plan pour placer les relevés à la moyenne des espèces qu'ils contiennent (averaging), ceci pour obtenir les meilleurs modèles additifs possibles.

Ce n'est pas tout à fait possible car l'erreur est vers l'extérieur en provence et vers l'intérieur en Corse, les deux structures ayant des propriétés internes fortes.

Si on veut enlever l'effet région c'est-à-dire superposer au mieux les deux ordinations, on utilise une ACPVI orthogonale [2]. On voit immédiatement qu'une telle opération est presque possible.

```
pcaivnonA <- pcaivortho(coa1,avimedi$plan$reg, scannf=FALSE, nf=3)
s.match(pcaivnonA$li,pcaivnonA$ls,clab=0)
s.class(pcaivnonA$li,cla,add.plot=T)
```



Les petites flèches indiquent l'erreur de correction sur les vraies moyennes conditionnelles à opérer pour obtenir deux nuages ayant le même centre de gravité (dans un sens pour une classe et dans le sens opposé pour l'autre classe). Ce qui frappe, c'est l'insertion de la typologie Corse dans celle de la Provence et la mise en évidence du syndrôme d'insularité sur les structures des communautés.

Références

[1] J. Obadia. L'analyse en composantes explicatives. *Revue de Statistique Appliqu*
, 24 :5–28, 1978.

[2] C.R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya, A*, 26 :329–359, 1964.

[3] R. Sabatier, J.D. Lebreton, and D. Chessel. Principal component analysis with instrumental variables as a tool for modelling composition data. In R. Coppi and S. Bolasco, editors, *Multiway data analysis*, pages 341–352. Elsevier Science Publishers B.V., North-Holland, 1989.

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