T. D. n° II . L'ACP pratique.

Exercice n° 1.

Une étude sur des fournisseurs de matériel informatique a conduit à apprécier le service, la qualité et le prix de quatre fournisseurs. Pour cela un expert a noté ces entreprises avec des notes allant de -3 à 3. Les résultats sont consignés ci-dessous

Ent	Service	Qualité	Prix
E1	-2	3	-1
E2	-1	1	0
E3	2	-1	-1
E4	1	-3	2

1) Calculer le vecteur moyen des individus. Qu'en conclure?

2) Calculer la covariance entre $\mathbf{x}^1$ et $\mathbf{x}^1$. Que représente cette quantité?

3) Calculer la covariance entre $\mathbf{x}^1$ et $\mathbf{x}^2$.

4) Donner la matrice de corrélation.

On veut faire une ACP centrée avec des poids uniformes.

5) Sur quelle matrice faut-il travailler? Vérifier qu'elle admet une valeur propre nulle. Qu'est ce que cela implique?

6) On donne $\lambda_1 = 61/8$. En déduire λ_2.

7) Calculer les pourcentages d'inertie. Quelle dimension retenez-vous?

8) Soient les vecteurs propres $\mathbf{a}_1 = (1/2, -4/5, 3/10)'$ et $\mathbf{a}_2 = (0.65, 0.11, -0.75)'$. Calculer les composantes principales.

9) Représenter les individus et les variables dans le plan principal $(1, 2)$. Interpréter.

10) Calculer la corrélation entre les variables initiales et les composantes principales.

Exercice n° 2

Soit $\mathbf{X} = (\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3)$ tel que

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & -\rho \\ \rho & 1 & \rho \\ -\rho & \rho & 1 \end{bmatrix}$$

avec $-1 \leq \rho \leq 1$. On veut faire une ACP centrée réduite de $\mathbf{X}$.

1) Vérifier que $\mathbf{R}$ admet pour vecteur propre $\xi_1 = \frac{1}{\sqrt{3}}(1, -1, 1)'$.

2) Déterminer les autres éléments de la décomposition aux valeurs propres de $\mathbf{R}$.

3) Quels sont les % de variance expliquée? Quels axes retenir?

Corrections

1) Il suffit de faire la moyenne des colonnes du tableau, ce qui donne $\overline{x} = (\overline{x}_1, \overline{x}_2, \overline{x}_3)' = (0,0,0)'$. Le tableau est donc centré.

2) $cov\,(\mathbf{x}^1, \mathbf{x}^1) = \left\langle \mathbf{x}^1 - \overline{x}_1\mathbf{1}, \mathbf{x}^1 - \overline{x}_1\mathbf{1} \right\rangle_{D_p} = \frac{1}{n}\mathbf{x}^{1\prime}\mathbf{x}^1 = \|\mathbf{x}^1\|^2_{D_p}$ où D_p est la matrice associée à la métrique diagonale des poids des individus.

3) $cov\,(\mathbf{x}^1, \mathbf{x}^1) = \frac{1}{n}\mathbf{x}^{1\prime}\mathbf{x}^2$ puisque les données sont déjà centrées. On obtient $cov\,(\mathbf{x}^1, \mathbf{x}^1) = \frac{1}{4}(-2,-1,2,1)(3,1,-1,-3)' = -3$. On peut également calculer la matrice de variance

$$\mathbf{V} = \frac{1}{2}\begin{bmatrix} 5 & -6 & 1 \\ -6 & 10 & -4 \\ 1 & -4 & 3 \end{bmatrix}$$

4) La matrice de corrélation $\mathbf{R}$ a pour terme général $\dfrac{cov\left(\mathbf{x}^j, \mathbf{x}^k\right)}{\|\mathbf{x}^j\|_{D_p}\|\mathbf{x}^k\|_{D_p}}$. Elle est telle que $\mathbf{R} = \mathbf{M}^{1/2}\mathbf{V}\mathbf{M}^{1/2}$ où $\mathbf{M}$ est la matrice de terme général $\dfrac{1}{\|\mathbf{x}^j\|^2_{D_p}}$ et $\mathbf{V}$ la matrice de variance-covariance. Après calculs :

$$\mathbf{R} = \mathbf{M}^{1/2}\mathbf{V}\mathbf{M}^{1/2} = \begin{bmatrix} 1 & -\frac{6}{\sqrt{50}} & \frac{1}{\sqrt{15}} \\ -\frac{6}{\sqrt{50}} & 1 & -\frac{4}{\sqrt{30}} \\ \frac{1}{\sqrt{15}} & -\frac{4}{\sqrt{30}} & 1 \end{bmatrix}.$$

5) On travaille sur la matrice de covariance (tableau initial non réduit). Si elle admet une valeur propre nulle, alors son déterminant doit être nul puisque :

$$\det(\mathbf{V}) = \lambda_1 \times \lambda_2 \times \lambda_3.$$

Le calcul du déterminant donne $\det(\mathbf{V}) = \frac{1}{2^3}[5 \times (30-16) + 6 \times (-18+4) + 1 \times (24-10)] = 0$. Elle admet donc une valeur propre nulle. Cela implique qu'il existe une combinaison linéaire entre variables initiales dans le tableau $\mathbf{X}$. En effet, on vérifie aisément que $\mathbf{x}^3 = -(\mathbf{x}^1 + \mathbf{x}^2)$. Ceci indique qu'il n'est pas utile de mesurer l'une des trois variables : deux suffisent puisqu'on peut reconstituer la troisième à l'aide des deux autres.

6) On sait que l'inertie totale I_T est égale à la somme des valeurs propres. On a donc $I_T = \mathrm{Tr}(\mathbf{V}) = \lambda_1 + \lambda_2 + \lambda_3$. On a $\lambda_1 = 61/8$, $\lambda_3 = 0$ et $\mathrm{Tr}(\mathbf{V}) = 1/2 \times (5+10+3) = 9$. On en déduit que $\lambda_2 = 9 - 61/8 = 11/8$.

7) Le pourcentage d'inertie de la dimension k est donné par $p_k = \frac{\lambda_k}{I_T}$ (Fig. 1).
On retiendra naturellement deux axes pour représenter 100% de la variance.

8) Les composantes principales, qui donnent les coordonnées des individus dans la nouvelle base, sont données par $\mathbf{C} = \mathbf{X}_c\mathbf{M}\mathbf{A}$, où $\mathbf{X}_c = \mathbf{X}$ dans notre cas (le tableau est déjà centré), la métrique $\mathbf{M} = \mathbf{I}_3$ ici et $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]'$ est la matrice des vecteurs propres associées aux valeurs propres $(\lambda_1, \lambda_2, \lambda_3)$. On ne retiendra ici que deux composantes principales, la troisième valeur propre étant nulle. On obtient :

$$\mathbf{C} = \begin{bmatrix} -3.737 & 0.185 & 0 \\ -1.312 & 0.529 & 0 \\ 1.509 & -1.929 & 0 \\ 3.539 & 1.215 & 0 \end{bmatrix}.$$

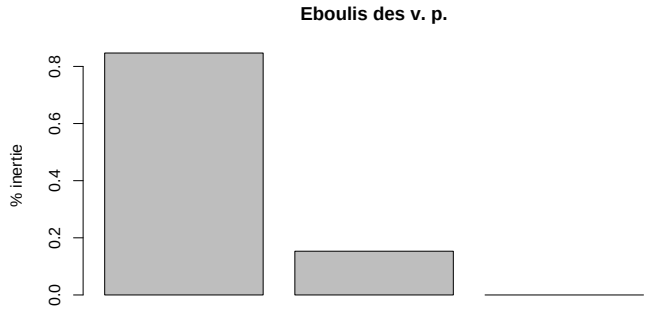

Figure 1: Eboulis des valeurs propres.

9)

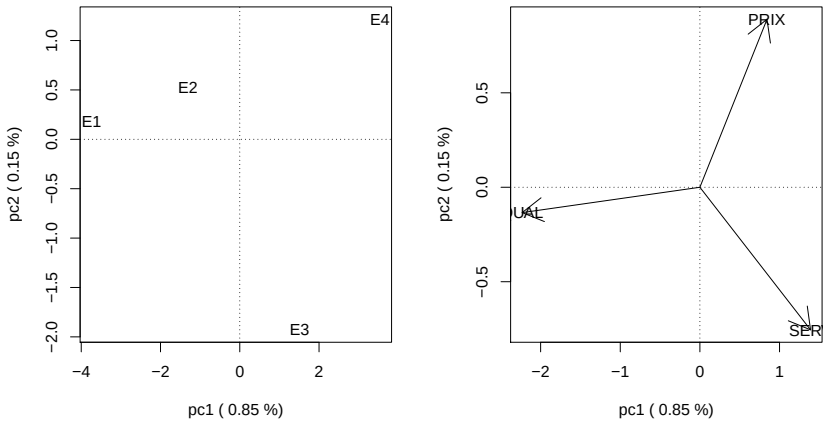

Figure 2: Représentation des individus et variables sur le premier plan factoriel.

10) La corrélation entre variables initiales et composantes principales permet d'interpréter les composantes principales en les reliant aux variables du tableau initial. On a :

$$\begin{bmatrix} 0.879 & -0.477 & 0.320 \\ -0.998 & -0.060 & -0.773 \\ 0.688 & 0.726 & 0.998 \end{bmatrix}.$$

En cadeau bonus voici le code R qui permet d'illustrer ce TD :

```
#TD2 EXO 1
library(ade4)
library(rgl)
```

```
X=cbind(c(-2,-1,2,1),c(3,1,-1,-3),c(-1,0,-1,2))
rownames(X)=paste("E",1:4,sep="")
colnames(X)=c("SERV","QUAL","PRIX")
#### METRIQUE
Dp=diag(1/4,4,4)
M=diag(1,3,3)
I=matrix(1,4,1)
###CALCUL DU VECTEUR MOYEN DES INDIVIDUS
g=t(I)%*%Dp%*%X

#### MATRICE DE VARIANCE COVARIANCE

V=t(X)%*%Dp%*%X

### MATRICE DE CORRELATION
### ICI LA METRIQUE M CHANGE
M=diag(1/diag(V),3,3)
R=sqrt(M)%*%V%*%sqrt(M)

### ACP CENTREE / POIDS UNIFORME
### ON TRAVAILLE SUR V
M=diag(1,3,3)

Xacp=eigen(V)
Xacp$val=round(Xacp$val,3)
colnames(Xacp$cp)=c("pc1","pc2","pc3")

### EBOULIS DES VP
Xacp$pinert=Xacp$val/sum(Xacp$val)
barplot(Xacp$pinert,main="Eboulis des v. p.",ylab="% inertie")

### COMP PRINC
Xacp$cp=X%*%M%*%Xacp$vect

### COORD VARIABLES (FACTEURS)
L=matrix(Xacp$val,3,3,byrow=TRUE)
###ATTENTION PRODUIT TERME A TERME ICI
Xacp$fac=sqrt(L)*Xacp$vect


### REPRESENTATIONS GRAPHIQUES
X11()
par(mfrow=c(1,2))
plot(Xacp$cp[,1:2],xlab=paste("pc1 (",round(Xacp$pinert[1],2),"%)"),
     ylab=paste("pc2 (",round(Xacp$pinert[2],2),"%)"),type="n")
```

```
text(Xacp$cp[,1:2],rownames(X))
abline(h=0,v=0,lty=3)
plot(Xacp$fac[,1:2],type="n",xlab=paste("pc1 (",round(Xacp$pinert[1],2),"%)"),
     ylab=paste("pc2 (",round(Xacp$pinert[2],2),"%)"))
arrows(0,0,Xacp$fac[,1],Xacp$fac[,2])
abline(h=0,v=0,lty=3)
text(Xacp$fac[,1:2],colnames(X))
par(mfrow=c(1,1))

### CORRELATION ENTRE VARIABLES INITIALES ET COMP. PRINC.

Y=cbind(X,Xacp$cp)
CORPCV=cor(Y)[,4:6]
CORPCV=CORPCV[1:3,]
CORPCV=round(CORPCV,3)

#### AVEC UN PACKAGE
x11()
Xdudi=dudi.pca(X,center=TRUE,scale = FALSE,scan = FALSE)
par(mfrow = c(1,2))
s.label(Xdudi$li)
s.arrow(Xdudi$co, lab = colnames(X))
par(mfrow = c(1,1))
```

Exercice n° 2.

1. Si $\mathbf{R}$ admet pour vecteur propre ξ_1 alors il vérifie $\mathbf{R}\xi_1 = \lambda_1\xi_1,\ \lambda_1 \in \mathbb{R}^+$. On calcule $\mathbf{R}\xi_1$:

$$\frac{1}{\sqrt{3}}\begin{bmatrix} 1 & \rho & -\rho \\ \rho & 1 & \rho \\ -\rho & \rho & 1 \end{bmatrix}\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{3}}\begin{pmatrix} 1-2\rho \\ 2\rho-1 \\ -2\rho+1 \end{pmatrix} = (1-2\rho)\,\xi_1$$

   donc, ξ_1 est bien vecteur propre de $\mathbf{R}$ pour la valeur propre $\lambda_1 = 1-2\rho$. Cette valeur propre étant positive (propriété de $\mathbf{R}$ matrice symétrique définie positive) on doit avoir $-1 \leq \rho \leq \frac{1}{2}$.

2. Pour déterminer les autres éléments propres de $\mathbf{R}$, on résout $\det(\mathbf{R}-\lambda\mathbf{I}) = 0$, ce qui équivaut à

$$(1-\lambda)\left((1-\lambda)^2-\rho^2\right) - 2\rho^2(1-\lambda+\rho) = 0$$

$$(1-\lambda+\rho)\left[(1-\lambda)(1-\lambda-\rho) - 2\rho^2\right] = 0$$

$$(1-\lambda+\rho)\left[\lambda^2 - \lambda(2-\rho) + 1 - \rho - 2\rho^2\right] = 0$$

   On sait que $\lambda_1 = 1-2\rho$ est valeur propre de $\mathbf{R}$. Ceci permet de calculer par identification la racine du polynôme ci-dessus. On montre que $\lambda = 1+\rho$ est racine double. On peut

maintenant déterminer les vecteurs propres pour cette valeur propre. Soit $\xi = (x, y, z)'$ un vecteur vérifiant $\mathbf{R}\xi = \lambda\xi$. En développant, on obtient le système

$$\begin{cases} -\rho x + \rho y - \rho z = 0 \\ \rho x - \rho y + \rho z = 0 \\ -\rho x + \rho y - \rho z = 0 \end{cases} .$$

Il nous faut maintenant trouver des valeurs arbitraires de x, y et z qui vérifient ce système. On en trouve facilement 2 tiercés avec $(1, 1, 0)$ et $(1, 0, -1)$ qui ne soient pas combinaison linéaire l'un de l'autre. En normalisant ces vecteurs, on obtient finalement les deux vecteurs propres $\xi_2 = \frac{1}{\sqrt{2}}(1, 1, 0)'$ et $\xi_3 = \frac{1}{\sqrt{2}}(1, 0, -1)'$. Finalement, la matrice des corrélations $\mathbf{R}$ peut être décomposée sous la forme $\mathbf{R} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$ avec $\mathbf{P} = [\xi_1, \xi_2, \xi_3]$, matrice des vecteurs propres et $\mathbf{\Lambda}$, matrice des valeurs propres de termes diagonaux $(1 - 2\rho, 1 + \rho, 1 + \rho)$.

3. Les pourcentages d'inertie expliquée sont donnés, dans chaque direction propre, par le rapport d'une valeur propre sur la somme totale des valeurs propres, égale dans ce cas à 3, puisqu'elle correspond à l'inertie totale calculée à partir de la matrice des corrélations (variables réduites). Nous avons déjà vu que les valeurs possibles de ρ sont $-1 \leq \rho \leq \frac{1}{2}$ pour assurer la positivité des valeurs propres. Supposons maintenant que $-1 < \rho < 0$. On peut ranger les valeurs propres par ordre décroissant avec $1 - 2\rho > 1 + \rho$. On se rend alors compte que l'espace initial à 3 variables peut être réduit à une seule variable, combinaison linéaire des 3 variables initiales. En effet, si l'on considère le sous-espace propre de dimension 2 associé à la valeur propre double, l'information du nuage de points résumé dans cet espace est identique dans les deux directions. Cela n'apporte rien de les conserver : on ne gardera qu'un axe. Dans ce cas, l'éboulis correspond au tracé, sur le même graphique, de barres de hauteur $(1 - 2\rho)/3$, $(1 + \rho)/3$ et $(1 + \rho)/3$. Voyons maintenant le cas où $\rho = 0$. Dans ce cas, la matrice de corrélation est diagonale : les variables sont non corrélées deux à deux. Les valeurs propres sont toutes égales à 1 : le nuage de point est sphérique. Si $0 < \rho < 0.5$, les valeurs propres sont telles que $1 + \rho > 1 - 2\rho$: les deux premiers sous-espaces sont à retenir, de même inertie. A vous de regarder les cas extrêmes $\rho = -1$ et $\rho = 0.5$ et d'en déduire les axes à retenir et les variances associées.