MAÎTRISE BPE - UV MIAB 2 - JUIN 2003

STATISTIQUES (2 HEURES)

Exercice 1 : Une analyse des correspondances 2-2

On utilise l'objet :

```
 minitab=as.data.frame(matrix(c(2,1,1,2),2,2))
 minitab
  V1 V2
1  2  1
2  1  2
 minidudi=dudi.coa(minitab, scan=F)
 names(minidudi)
 "tab"  "cw"   "lw"   "eig"  "rank" "nf"   "c1"   "l1"   "co"   "li"   "call" "N"
```

Quel est le contenu de chacune des 12 composantes de la liste `minidudi` ? On donnera une réponse exacte, sans approximation numérique, en notation mathématique, dans la place prévue à cet effet sur la feuille de réponse.

Exercice 2 : Inertie et Chi2 en analyse des correspondances

Pour introduire la question, une observation utile :

```
 data(chats)
 chats
     f0 f12 f34 f56 f78 f9a fbc fcd
age1  8  15  44  11   7   4   0   0
age2  6  12  36  21  11   6   1   1
age3  4   7  18  13  12   4   2   2
age4  2   8   7   3   7   5   1   0
age5  2   3   5   3   4   6   0   0
age6  1   0   5   3   2   2   1   1
age7  0   0   3   2   5   4   1   1
age8  2   2   5   1   7   4   1   0
 chisq.test(chats)$statistic/sum(chats)
X-squared
   0.2106
 sum(dudi.coa(chats,scannf=F)$eig)
[1] 0.2106
```

Une table de contingence forme un tableau $\mathbf{X} = \left[x_{ij} \right]$ de nombres positifs ou nuls. **X** a I lignes et J colonnes. n est la somme de tous les éléments du tableau **X**. **P** est le tableau de terme général $p_{ij} = x_{ij} / n$. Dans les notations habituelles, $p_{i.}$ est la somme par ligne, $p_{.j}$ est la somme par colonne . $\mathbf{D}_I$ et $\mathbf{D}_J$ sont les matrices diagonales associées. Exprimer le chi2 de la table de contingence **X** en fonction de I, J, p_{ij}, $p_{i.}$, $p_{.j}$ et n. En prenant une forme du triplet de l'analyse des correspondances, calculer l'inertie totale en fonction de I, J, p_{ij}, $p_{i.}$, $p_{.j}$ et n. En déduire une relation entre les valeurs propres de l'analyse des correspondances d'un tableau **X** et le chi2 associé à **X**.

Exercice 3 : Analyse en composantes principales et génétique

Le polymorphisme biochimique des bovins domestiques fait l'objet de nombreuses études. D. Laloë (INRA, Jouy-en-Josas) propose un extrait pédagogique comportant 2 races taurines africaines (Taurins N'Dama et Baoulé), 2 races de Zébus (Zébu Azawak du Niger et Zébu malgache) et 2 races bovines européennes (Charolais et Salers). Le tableau donne les fréquences alléliques de 4 systèmes génétiques ($\alpha_{s1}-\mathbf{Cn}$, $\beta-\mathbf{Cn}$, $\kappa-\mathbf{Cn}$ et $\beta-\mathbf{Lg}$) définis par le polymorphisme des protéines du lait :

	alpha		beta			kappa			beta-lacto	
Ndama	0.89	0.11	0.60	0.37	0.03	0.27	0.73	0.00	0.10	0.90
Baoule	0.92	0.08	0.63	0.36	0.01	0.34	0.64	0.02	0.09	0.91
Zebu_a	0.22	0.78	0.08	0.86	0.06	0.83	0.17	0.00	0.14	0.86
Zebu_m	0.17	0.83	0.10	0.90	0.00	0.75	0.25	0.00	0.27	0.73
Charolais	0.92	0.08	0.10	0.76	0.14	0.49	0.51	0.00	0.67	0.33
Salers	0.96	0.04	0.19	0.70	0.11	0.54	0.46	0.00	0.64	0.36

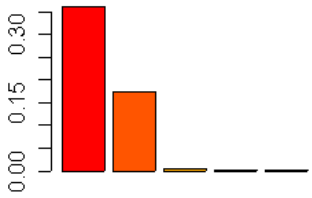
Le but de l'exercice est d'étudier les propriétés de l'analyse en composantes principales des tableaux de fréquences alléliques. On écrit le tableau traité $\mathbf{X}=\left[\mathbf{X}_1\middle|\mathbf{X}_2\middle|\mathbf{X}_3\middle|\mathbf{X}_4\right]$ et on note $\mathbf{X}_0=\left[\mathbf{X}_{01}\middle|\mathbf{X}_{02}\middle|\mathbf{X}_{03}\middle|\mathbf{X}_{04}\right]$ le tableau centré par variable.

Le fichier est lu dans R :

```
bovins
            a1   a2   b1   b2   b3   k1   k2   k3   l1   l2
Ndama     0.89 0.11 0.60 0.37 0.03 0.27 0.73 0.00 0.10 0.90
Baoule    0.92 0.08 0.63 0.36 0.01 0.34 0.64 0.02 0.09 0.91
Zebu_a    0.22 0.78 0.08 0.86 0.06 0.83 0.17 0.00 0.14 0.86
Zebu_m    0.17 0.83 0.10 0.90 0.00 0.75 0.25 0.00 0.27 0.73
Charolais 0.92 0.08 0.10 0.76 0.14 0.49 0.51 0.00 0.67 0.33
Salers    0.96 0.04 0.19 0.70 0.11 0.54 0.46 0.00 0.64 0.36
```

Le tableau est soumis à une analyse en composantes principales centrée :

```
pca1 = dudi.pca(bovins,scale=F)
Select the number of axes: 2
```



Quelques indications utiles :

```
names(pca1)
 [1] "tab"  "cw"   "lw"   "eig"  "rank" "nf"   "c1"   "l1"   "co"   "li"
[11] "call" "cent" "norm"

round(apply(bovins,2,mean),dig=3)
   a1    a2    b1    b2    b3    k1    k2    k3    w1    w2
0.680 0.320 0.283 0.658 0.058 0.537 0.460 0.003 0.318 0.682

pca1$eig
[1] 0.3609921 0.1754960 0.0049996 0.0014243 0.0000436

cumsum(pca1$eig)/sum(pca1$eig)
[1] 0.6649 0.9881 0.9973 0.9999 1.0000
```

Exercice 3 - Q1 Donner le plus simplement possible ce que vous attendez des ordres suivants:

```
apply(bovins,1,sum)
dim(pca1$tab)
length(pca1$lw)
pca1$cw
pca1$nf
is.data.frame(pca1$l1)
pca1$tab[1,1:2]
sum(pca1$tab[4,3:5])
apply(pca1$tab,1,sum)
apply(pca1$tab,2,sum)
```

Exercice 3 - Q2 Sachant que `pca1$l1[1,1]` contient la valeur `-0.9167`, que peut-on prévoir pour le contenu de `pca1$li[1,1]`. De manière symétrique sachant que `pca1$co[1,2]` contient la valeur `0.07267`, que peut-on prévoir pour `pca1$c1[1,2]` ?

On note $\mathbf{1}_m$ le vecteur à m composantes toutes égales à 1 et $\mathbf{0}_m$ est le vecteur à m composantes toutes égales à 0. $\mathbf{A}^t$ est la transposée de $\mathbf{A}$.

Exercice 3 - Q3 Une seule de ces égalités est fausse. Dites laquelle et dites pourquoi.

$$\mathbf{X}_{01}\mathbf{1}_2 = \mathbf{0}_6 \quad \mathbf{X}_{01}^t\mathbf{1}_6 = \mathbf{0}_2$$
$$\mathbf{X}_{02}\mathbf{1}_3 = \mathbf{0}_6 \quad \mathbf{X}_{02}^t\mathbf{1}_6 = \mathbf{0}_3$$
$$\mathbf{X}_{03}\mathbf{1}_3 = \mathbf{0}_6 \quad \mathbf{X}_{03}^t\mathbf{1}_6 = \mathbf{0}_3$$
$$\mathbf{X}_{04}\mathbf{1}_4 = \mathbf{0}_6 \quad \mathbf{X}_{04}^t\mathbf{1}_6 = \mathbf{0}_2$$

Exercice 3 - Q4 Quels sont les rangs des matrices $\mathbf{X}_{01}, \mathbf{X}_{02}, \mathbf{X}_{03}$ et $\mathbf{X}_{04}$?

Exercice 3 - Q5 Donner 4 vecteurs indépendants qui vérifient $\mathbf{X}_0\mathbf{u} = 0$.

Exercice 3 - Q6 Donner un vecteur qui vérifie $\mathbf{X}_0^t\mathbf{v} = 0$.

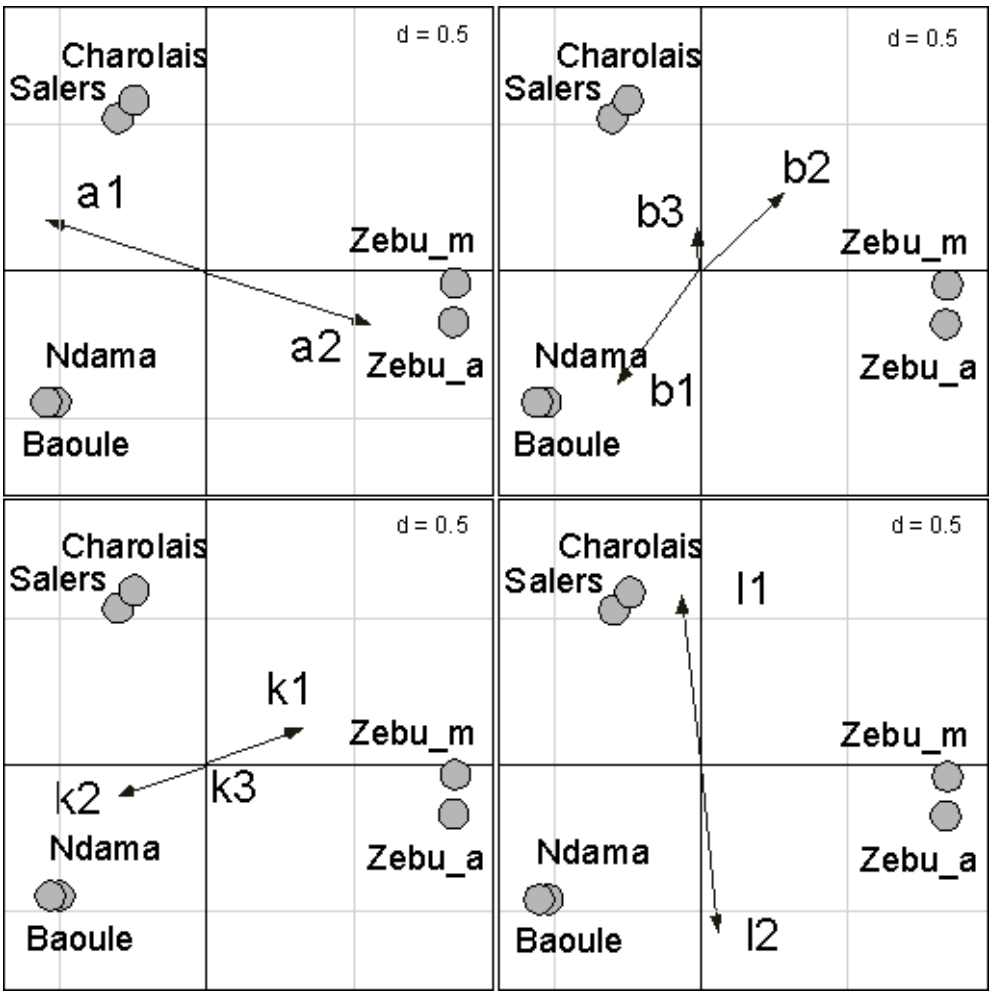
Exercice 3 - Q7 Quelles sont les dimensions de la matrice $\mathbf{C}$ de variances-covariances ? Donner 4 vecteurs indépendants qui vérifient $\mathbf{C}\mathbf{u} = 0$.

Exercice 3 - Q8 Quel est le rang de la matrice $\mathbf{C}$? Donner un argument numérique **et** un argument mathématique.

Exercice 3 - Q9 Quel est le taux d'inertie projetée sur le plan 1-2 de cette analyse en composantes principales.

Exercice 3 - Q10 Donner une légende à la figure, sachant qu'elle a été construite à partier de :

```
par(mfrow=c(2,2))
s.label(pca1$li)
s.arrow(pca1$c1[1:2,],add.p=T)
s.label(pca1$li)
s.arrow(pca1$c1[3:5,],add.p=T)
s.label(pca1$li)
s.arrow(pca1$c1[6:8,],add.p=T)
s.label(pca1$li)
s.arrow(pca1$c1[9:10,],add.p=T)
```


d = 0.5
Charolais
Salers
a1
Zebu_m
a2
Ndama
Zebu_a
Baoule
d = 0.5
Charolais
Salers
b2
b3
Zebu_m
Ndama
b1
Zebu_a
Baoule
d = 0.5
Charolais
Salers
k1
Zebu_m
k2
k3
Ndama
Zebu_a
Baoule
d = 0.5
Charolais
Salers
l1
Zebu_m
Ndama
Zebu_a
Baoule
l2

MAÎTRISE BPE - UV MIAB 2 -STATISTIQUES (2 HEURES)

NOM : PRÉNOM :

Exercice 1 : Une analyse des correspondances 2-2

composante	nature	valeur
tab		
cw		
lw		
eig		
rank		
c1		
l1		
co		
li		
N		

Justificatifs et commentaires :

Exercice 2 : Inertie et Chi2 en analyse des correspondances

Exercice 3 : Analyse en composantes principales et génétique

Exercice 3 - Q1 Qu'attendez-vous des ordres :

```
apply(bovins,1,sum)


dim(pca1$tab)


length(pca1$lw)

pca1$cw

pca1$nf

is.data.frame(pca1$l1)

pca1$tab[1,1:2]
```

```
sum(pca1$tab[4,3:5])


apply(pca1$tab,1,sum)


apply(pca1$tab,2,sum)
```

Exercice 3 - Q2 `pca1$li[1,1]` et `pca1$co[1,2]`

Exercice 3 - Q3 Laquelle est fausse et pourquoi ?

Exercice 3 - Q4 Les rangs des matrices

Exercice 3 - Q5 4 vecteurs indépendants qui vérifient $\mathbf{X}_0\mathbf{u} = 0$

Exercice 3 - Q6 Un vecteur qui vérifie $\mathbf{X}_0'\mathbf{v} = 0$

Exercice 3 - Q7 Dimensions de **C** et 4 vecteurs indépendants qui vérifient $\mathbf{Cu} = 0$

Exercice 3 - Q8 Quel est le rang de la matrice **C** ?

Exercice 3 - Q9 Quel est le taux d'inertie projetée sur le plan 1-2

Exercice 3 - Q10 Donner une légende à la figure

Éléments de réponse. Exercice 1 : Une analyse des correspondances 2-2

La matrice à traiter est $\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. La table de fréquence associée est $\mathbf{P} = \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{bmatrix}$. Les diagonales des poids sont $\mathbf{D}_I = \mathbf{D}_J = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$. Le tableau doublement centré s'écrit :

$$\mathbf{P}_0 = \left[\frac{p_{ij}}{p_{i.} p_{.j}} - 1 \right] = \begin{bmatrix} 1/3 & -1/3 \\ -1/3 & 1/3 \end{bmatrix}$$

Le vecteur $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ est propre pour la valeur propre 0. L'autre est un score centré et normé pour la pondération uniforme, donc au signe près le vecteur $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$. Le vecteur des moyennes conditionnelles par ligne de **P** est $\begin{pmatrix} -1/3 \\ 1/3 \end{pmatrix}$ de variance $\frac{1}{9}$. Donc la première (et unique) valeur propre vaut $\frac{1}{9}$. Les coordonnées des lignes sur l'axe unique sont (coordonnée = composante x racine de la valeur propre) $\begin{pmatrix} -1/3 \\ 1/3 \end{pmatrix}$. Les coordonnées des colonnes sont identiques.

`tab` est un data.frame 2-2 qui contient $\mathbf{P}_0$, `cw` est un vecteur qui contient les poids des colonnes soit $(1/2, 1/2)$, `lw` est un vecteur qui contient les poids des lignes $(1/2, 1/2)$, `eig` est un vecteur qui contient les valeurs propres non nulles soit donc l'unique valeur $1/9$. `c1` est un data.frame à un seule colonne contenant le vecteur $(-1, 1)$. `l1` est un data.frame à un seule colonne contenant le vecteur $(-1, 1)$. `co` est un data.frame à une seule colonne contenant le vecteur $(-1/3, 1/3)$. `li` est un data.frame à une seule colonne contenant le vecteur $(-1/3, 1/3)$. `N` contient le total du tableau traité donc la valeur 6.

```
unclass(minidudi)
$tab                        $cw           $lw
       V1      V2            V1  V2        1   2
1  0.3333 -0.3333          0.5 0.5       0.5 0.5
2 -0.3333  0.3333

$eig
[1] 0.1111

$rank
[1] 1

$c1           $l1           $co                  $li
   CS1           RS1             Comp1             Axis1
V1  -1        1  -1         V1 -0.3333         1 -0.3333
V2   1        2   1         V2  0.3333         2  0.3333

$N
[1] 6
```

Exercice 2 : Inertie et Chi2 en analyse des correspondances

Les effectifs observés sont $o_{ij} = x_{ij} = np_{ij}$. Les effectifs théoriques sont $c_{ij} = np_{i.}p_{.j}$. Le chi2 vaut :

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(o_{ij}-c_{ij}\right)^2}{c_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(np_{ij}-np_{i.}p_{.j}\right)^2}{np_{i.}p_{.j}} = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij}-p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}}$$

On peut prendre un schéma quelconque centré, par exemple :

$$\begin{array}{ccc} & \mathbf{D}_J & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{D}_I^{-1}-\mathbf{1}_{IJ}\uparrow & & \downarrow \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}-\mathbf{1}_{IJ} \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I & \end{array}$$

$$I_T = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i.}p_{.j}\left(\frac{p_{ij}}{p_{i.}p_{.j}}-1\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij}-p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}}$$

On en déduit :

$$n\left(\lambda_1+\lambda_2+\ldots+\lambda_r\right) = \chi^2$$

Exercice 3 : Analyse en composantes principales et génétique

Exercice 3 - Q1 Qu'attendez-vous des ordres :

```
apply(bovins,1,sum) # La somme par ligne du tableau initial
    Ndama    Baoule    Zebu_a    Zebu_m Charolais    Salers
        4         4         4         4         4         4
dim(pca1$tab) # Les dimensions du tableau traité
[1]  6 10

 length(pca1$lw) # Le nombre de lignes
[1] 6

 pca1$cw # Les poids des colonnes, unitaires en ACP
 [1] 1 1 1 1 1 1 1 1 1 1

 pca1$nf # Le nombre de facteurs conservés
[1] 2

 is.data.frame(pca1$l1)  # le tableau des composantes principales forment un
data.frame
[1] TRUE

pca1$tab[1,1:2] # la valeur observée 0.89 - la moyenne 0.68 donne 0.21
        a1    a2
Ndama 0.21 -0.21
```

```
 sum(pca1$tab[4,3:5]) # La somme par ligne et par bloc de colonne dans le tableau
centré vaut 0 : voir question 3 pour la notation mathématique
[1] -1.388e-17

 apply(pca1$tab,1,sum) # La somme par ligne fait la somme des sommes des blocs
toujours nulle. On attend 0 partout.

     Ndama     Baoule     Zebu_a     Zebu_m  Charolais     Salers
-3.331e-16 -2.498e-16 -4.163e-16 -2.567e-16 -2.220e-16 -3.886e-16

  apply(pca1$tab,2,sum) # La somme par colonne dans une ACP centrée est toujours
nulle et on attend 0 partout.
        a1         a2         b1         b2         b3         k1         k2
-2.220e-16  0.000e+00 -2.498e-16  0.000e+00  1.388e-17 -3.886e-16 -1.110e-16
        k3         l1         l2
-2.602e-18 -3.331e-16 -4.996e-16
```

Exercice 3 - Q2 `pca1$li[1,1] et pca1$co[1,2]`
On attend `-0.9167*sqrt(0.3609921)` (composante du vecteur x racine de la valeur propre)

```
pca1$l1[1,1]
[1] -0.9167
 pca1$li[1,1]
[1] -0.5508
 -0.9167*sqrt(0.3609921)
[1] -0.5508
```

On attend `0.07267/sqrt(0.1754960)` (coordonnée divisée par la racine de la valeur propre)

```
 pca1$co[1,2]
[1] 0.07267
 pca1$c1[1,2]
[1] 0.1735
 0.07267/sqrt(0.1754960)
[1] 0.1735
```

Exercice 3 - Q3 Laquelle est fausse et pourquoi ?

Chaque tableau élémentaire donne des sommes par ligne égales à 1. C'est encore vrai pour la somme des moyennes par colonne. Donc chaque tableau centré donne des somme par ligne égales à 0. D'autre part, la somme par colonne dans un tableau centré est toujours nulle. La première colonne d'égalité donne la première propriété et la seconde colonne exprime la seconde propriété. Mais l'équation $\mathbf{X}_{04}\mathbf{1}_4 = \mathbf{0}_6$ comporte une faute et n'a pas de sens. Il aurait fallu écrire $\mathbf{X}_{04}\mathbf{1}_2 = \mathbf{0}_6$.

Exercice 3 - Q4 Les rangs des matrices

Vue la question précédente on a toujours une combinaison linéaire de colonnes nulle et des rangs égaux à 1, 2, 2 et 1 pour les dimensions de 2, 3, 3 et 2 colonnes.

Exercice 3 - Q5 4 vecteurs indépendants qui vérifient $\mathbf{X}_0\mathbf{u} = 0$

Vue la question 3, il suffit de prendre :

$$\mathbf{u}_1 = \begin{pmatrix}1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}^t$$
$$\mathbf{u}_2 = \begin{pmatrix}0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0\end{pmatrix}^t$$
$$\mathbf{u}_3 = \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0\end{pmatrix}^t$$
$$\mathbf{u}_4 = \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1\end{pmatrix}^t$$

Exercice 3 - Q6 Un vecteur qui vérifie $\mathbf{X}_0^t\mathbf{v} = 0$

Il suffit de prendre $\mathbf{v} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}^t$

Exercice 3 - Q7 Dimensions de **C** et 4 vecteurs indépendants qui vérifient $\mathbf{Cu} = 0$

La matrice est de dimension 10-10. Les vecteurs **u** de la question 5 précédente conviennent.

Exercice 3 - Q8 Quel est le rang de la matrice **C** ?

La procédure n'a affiché que 5 valeurs propres. Le rang doit être de 5. Pourquoi ? Le rang de **C** est celui de $\mathbf{X}_0$ qui est une matrice 10-6. Le rang est au plus égal à 6 mais la somme des lignes est nulle. Le rang est au plus égal à 5. C'est la seule difficulté du problème. La présence de 4 combinaisons linéaires nulles (à coefficients non nuls) pour 10 vecteurs (colonnes) à 6 composantes impliquent un rang inférieur ou égal à 6 : on voit cet aspect dans `sum(pca1$tab[4,3:5])` et `apply(pca1$tab,1,sum)` de la question 1 et les 4 vecteurs des questions 5 et 7. Mais la présence des variables centrées font que le tableau est aussi formé de 6 lignes à 10 composantes dont la somme nulle implique un rang inférieur ou égal à 5. On voit cet aspect dans `apply(pca1$tab,2,sum)` de la question 1 et le vecteur de la question 6.

Exercice 3 - Q9 Quel est le taux d'inertie projetée sur le plan 1-2 ?

Il vaut la somme des deux premières valeurs propres divisée par la somme des variances (ou des valeurs propres) soit 0.9881 comme indiqué dans l'indication du sujet :

```
cumsum(pca1$eig)/sum(pca1$eig)
[1] 0.6649 0.9881 0.9973 0.9999 1.0000
```

Exercice 3 - Q10 Donner une légende à la figure.

La figure est formée de quatre fenêtres, chacune d'entre elles utilisant un seul locus. Dans une fenêtre on trouve la carte factorielle des lignes de l'ACP c'est-à-dire la projection des 6 points à 10 composantes sur un plan d'inertie qui contient toute l'information (99%). Les échelles sont communes aux deux axes et un carré a 0.5 de côté. Pour un locus donné, on trouve en outre la projection des vecteurs de la base canonique (10 composantes) associés aux colonnes d'un même locus.

Mathématiquement le point a1 de la première fenêtre est la projection sur le plan principal du vecteur $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ et le point a2 la projection du vecteur $\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$. La représentation globale est un biplot multi-fenêtré d'ACP.

Expérimentalement, on y lit directement l'apport que chaque locus fait dans la typologie globale. L'allèle a2 du système alpha isole les zébus, l'allèle l1 du système beta-lacto identifie les bovins européens, l'allèle k2 du système kappa isole les taurins africains. La forme de la figure pose le problème de la cohérence, ou de la redondance des typologies faites par différents marqueurs (par exemple dans Moazami-Goudarzi, K., and D. Laloe. 2002. Is a multivariate consensus representation of genetic relationships among populations always meaningful? *Genetics* **162**:473-484) et introduit ausx méthodes *K*-tableaux.