Fiche de Biostatistique

Problème d'analyse des données

Maîtrise BPE - MIAB

Lactoprotéines

D. Chessel

Plan

1. Questions

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 ($a_{s1}-\mathbf{Cn}$, $b-\mathbf{Cn}$, $k-\mathbf{Cn}$ et $b-\mathbf{Lg}$) 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 $\mathbf{X}=[\mathbf{X_1}, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4]$ et on note $\mathbf{X}_0=[\mathbf{X}_{01}, \mathbf{X}_{02}, \mathbf{X}_{03}, \mathbf{X}_{04}]$ le tableau centré par variable.

Le fichier est lu dans R :

```
> bovins
            a1   a2   b1   b2   b3   k1   k2   k3   w1   w2
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
```

Les moyennes des variables sont :

```
> 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
```

Les variances des variables sont :

```
> round(apply(bovins,2,var),dig=3)
   a1    a2    b1    b2    b3    k1    k2    k3    w1    w2
0.142 0.142 0.068 0.057 0.003 0.049 0.047 0.000 0.072 0.072
```

Question 1 Compléter (placer les étiquettes, le centre de gravité et la droite portée par le premier axe principal du nuage centré) et légender la figure.

Question 2 Quel résultat donnera la commande :

```
> apply(bovins,1,sum)
```

Question 3 Quelle est la propriété particulière du vecteur des moyennes du tableau étudié ?

Question 4 Calculer $\mathbf{X}_{01}$.

Question 5 Quel est le rang de $\mathbf{X}_{01}$? Quel sont les rangs de $\mathbf{X}_{02}, \mathbf{X}_{03}$ et $\mathbf{X}_{04}$?

Question 6 Donner 4 vecteurs indépendants qui vérifie $\mathbf{X}_0\mathbf{u} = 0$.

On considère 3 points du plan de coordonnées définies dans le logiciel R par:

```
A_c(-1/sqrt(2),-1/sqrt(6))
B_c(1/sqrt(2),-1/sqrt(6))
C_c(0,2/sqrt(6))
```

Question 7 Tracer le triangle ABC.

Question 8 Démontrer que le triangle ABC est équilatéral.

Question 9 O est l'origine. Démontrer que OA est perpendiculaire à BC.

Question 10 Placer sur le graphique le centre de gravité du nuage des 6 points en expliquant votre méthode.

Question 11 Quelle est l'inertie de ce nuage de 6 points autour de son centre de gravité ?

Question 12 Placer sur le graphique les axes principaux du nuage **centré** des 6 points (on ne demande aucune solution exacte mais juste une indication *au pif* !).

Question 13 Compléter et légender la figure obtenue par :

```
> plot(princomp(bovins[,3:5])$scores[,1:2],asp=1)
```

On fait maintenant l'ACP centrée du tableau tout entier.

Question 14 Quelles sont les dimensions de la matrice $\mathbf{C}$ de variances-covariances ? Donner 4 vecteurs indépendants qui vérifient $\mathbf{Cu} = 0$.

On a :

```
> pr1_princomp(bovins)
> pr1$sdev
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
6.008e-01 4.189e-01 7.071e-02 3.774e-02 6.603e-03 4.001e-09 3.551e-09 0.000e+00
   Comp.9   Comp.10
0.000e+00 0.000e+00
```

Question 15 Quel est le rang de la matrice $\mathbf{C}$? Donner un argument numérique **et** un argument mathématique.

```
> pr1$sdev^2
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
3.610e-01 1.755e-01 5.000e-03 1.424e-03 4.360e-05 1.601e-17 1.261e-17 0.000e+00
   Comp.9   Comp.10
0.000e+00 0.000e+00
```

Question 16 Donner une légende à la figure.

```
> pr1$loadings[,1:2]
     Comp.1    Comp.2
a1 -0.55797  0.173469
a2  0.55797 -0.173469
b1 -0.28101 -0.393408
b2  0.30093  0.286929
b3 -0.01992  0.106480
k1  0.31804  0.119716
k2 -0.31302 -0.111154
k3 -0.00502 -0.008563
w1 -0.06428  0.576464
w2  0.06428 -0.576464
```

Question 17 Quelle est la propriété particulière des vecteurs propres de cette analyse et d'où vient-elle ?

```
> pr1$scores[,1:2]
             Comp.1  Comp.2
  Ndama     -0.5508 -0.4511
  Baoule    -0.5437 -0.4508
  Zebu_a     0.8381 -0.1598
  Zebu_m     0.8343 -0.0485
  Charolais -0.2630  0.5876
  Salers    -0.3150  0.5226
```

Question 18 Compléter et légender la figure.

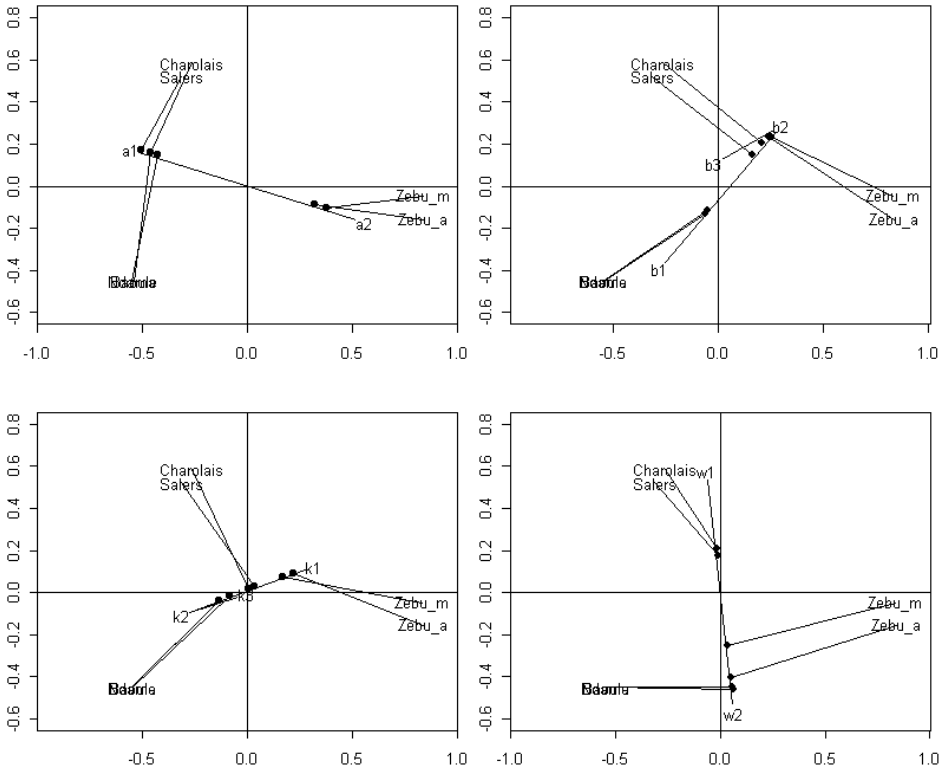
Question 19 Démontrer que pour tout matrice **A** de nombre réels $\mathbf{Au} = \mathbf{0} \Leftrightarrow \mathbf{A}^t\mathbf{Au} = \mathbf{0}$

Pour approfondir l'interprétation, on considère le tableau bov2 constitué à partir du précédent par le procédé très simple exprimé par :

0.89	0.11	0	0	0	0	0	0	0	0
0.92	0.08	0	0	0	0	0	0	0	0
0.22	0.78	0	0	0	0	0	0	0	0
0.17	0.83	0	0	0	0	0	0	0	0
0.92	0.08	0	0	0	0	0	0	0	0
0.96	0.04	0	0	0	0	0	0	0	0
0	0	0.6	0.37	0.03	0	0	0	0	0
0	0	0.63	0.36	0.01	0	0	0	0	0
0	0	0.08	0.86	0.06	0	0	0	0	0
0	0	0.1	0.9	0	0	0	0	0	0
0	0	0.1	0.76	0.14	0	0	0	0	0
0	0	0.19	0.7	0.11	0	0	0	0	0
0	0	0	0	0	0.27	0.73	0	0	0
0	0	0	0	0	0.34	0.64	0.02	0	0
0	0	0	0	0	0.83	0.17	0	0	0
0	0	0	0	0	0.75	0.25	0	0	0
0	0	0	0	0	0.49	0.51	0	0	0
0	0	0	0	0	0.54	0.46	0	0	0
0	0	0	0	0	0	0	0	0.1	0.9
0	0	0	0	0	0	0	0	0.09	0.91
0	0	0	0	0	0	0	0	0.14	0.86
0	0	0	0	0	0	0	0	0.27	0.73
0	0	0	0	0	0	0	0	0.67	0.33
0	0	0	0	0	0	0	0	0.64	0.36

Les lignes du nouveau tableau centré sont projetées en individus supplémentaires sur le plan 1-2 de l'analyse qui précéde. On montre facilement que la position d'un point sur la carte ordinaire est obtenue par la somme de ses 4 nouvelles positions.

Question 20 Expliquer en quoi la figure obtenue exprime les données et commenter le rôle des différents sous-tableaux dans le résultat global.



Annexe : princomp(mva) R Documentation

Principal Components Analysis

Description

princomp performs a principal components analysis on the given data matrix and returns the results as an object of class princomp.

Usage

```
princomp(x, cor = FALSE, scores = TRUE, covmat = NULL,
         subset = rep(TRUE, nrow(as.matrix(x))))
```

Arguments

x a matrix (or data frame) which provides the data for the principal components analysis.
cor a logical value indicating whether the calculation should use the correlation matrix or the covariance matrix.
scores a logical value indicating whether the score on each principal component should be calculated.
covmat a covariance matrix, or a covariance list as returned by cov.wt, cov.mve or cov.mcd. If supplied, this is used rather than the covariance matrix of x.
subset a vector used to select rows (observations) of the data matrix x.
x, object an object of class "princomp", as from princomp().
npcs the number of principal components to be plotted.
type the type of plot.
... graphics parameters.

Details

The calculation is done using eigen on the correlation or covariance matrix, as determined by cor. This is done for compatibility with the S-PLUS result. A preferred method of calculation is to use svd on x, as is done in prcomp.

Note that the default calculation uses divisor N for the covariance matrix.

Value

princomp returns a list with class "princomp" containing the following components:

sdev the standard deviations of the principal components.
loadings the matrix of variable loadings (i.e., a matrix whose columns contain the eigenvectors).
center the means that were subtracted.
scale the scalings applied to each variable.
n.obs the number of observations.
scores if scores = TRUE, the scores of the supplied data on the principal components.
call the matched call.

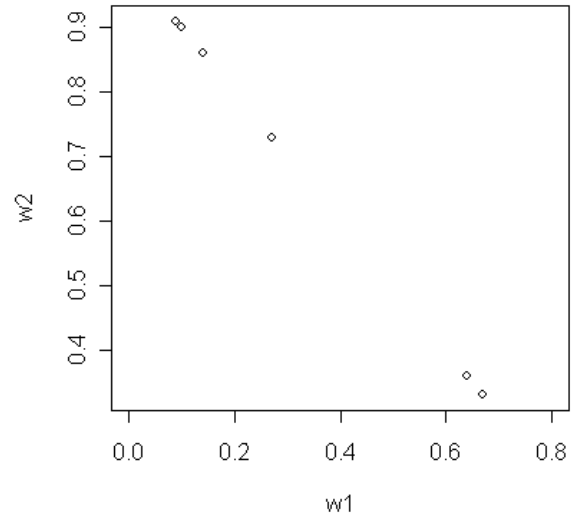
References

Mardia, K. V., J. T. Kent and J. M. Bibby (1979). Multivariate Analysis, London: Academic Press.

Venables, W. N. and B. D. Ripley (1997, 9). Modern Applied Statistics with S-PLUS, Springer-Verlag.

2. Feuille de réponse

Question 1 Compléter et légender la figure.



Question 2 Quel résultat donnera la commande :

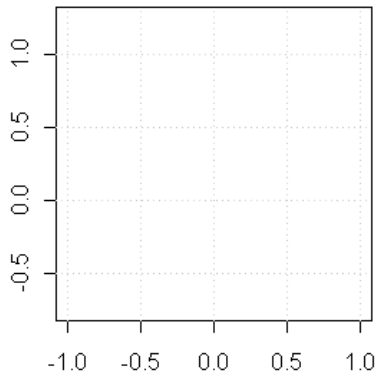
Question 3 Quelle est la propriété particulière du vecteur des moyennes du tableau étudié ?

Question 4 Calculer $\mathbf{X}_{01}$.

Question 5 Quel est le rang de $\mathbf{X}_{01}$? Quel sont les rangs de $\mathbf{X}_{02}, \mathbf{X}_{03}$ et $\mathbf{X}_{04}$?

Question 6 Donner 4 vecteurs indépendants qui vérifie $\mathbf{X}_0\mathbf{u} = 0$.

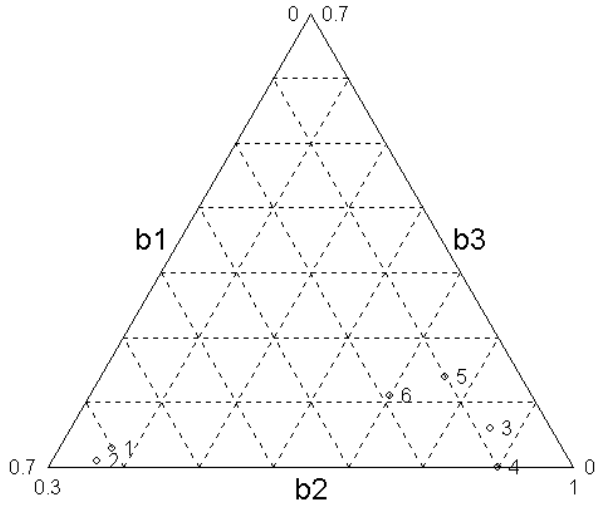
Question 7 Tracer le triangle ABC.



Question 8 Démontrer que le triangle ABC est équilatéral.

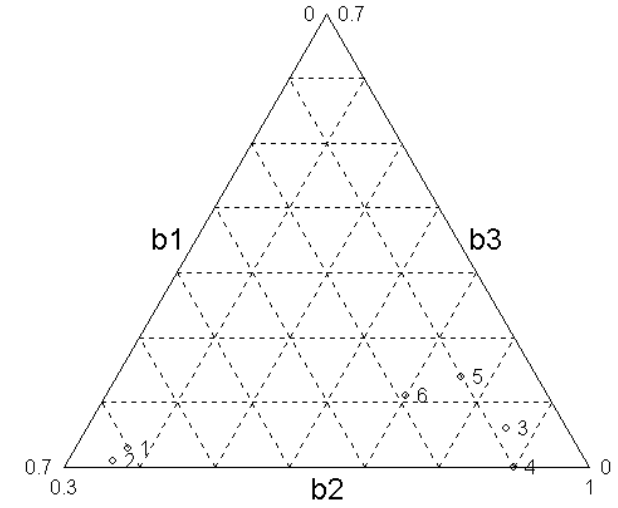
Question 9 O est l'origine. Démontrer que OA est perpendiculaire à BC.

Question 10 Placer sur le graphique suivant le centre de gravité …

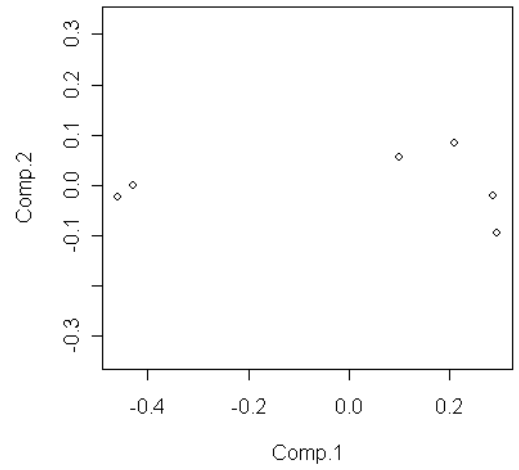


Question 11 Quelle est l'inertie de ce nuage de 6 points autour de son centre de gravité ?

Question 12 Placer sur le graphique suivant les axes principaux du nuage **centré** …



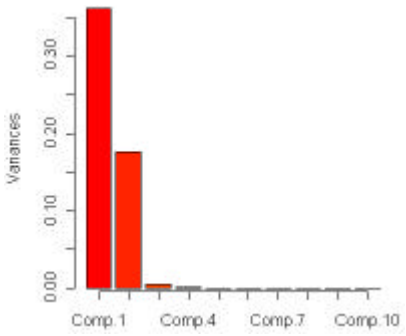
Question 13 Compléter et légender la figure :



Question 14 Quelles sont les dimensions de la matrice **C** de variances-covariances ? Donner 4 vecteurs indépendants qui vérifie $\mathbf{Cu} = 0$.

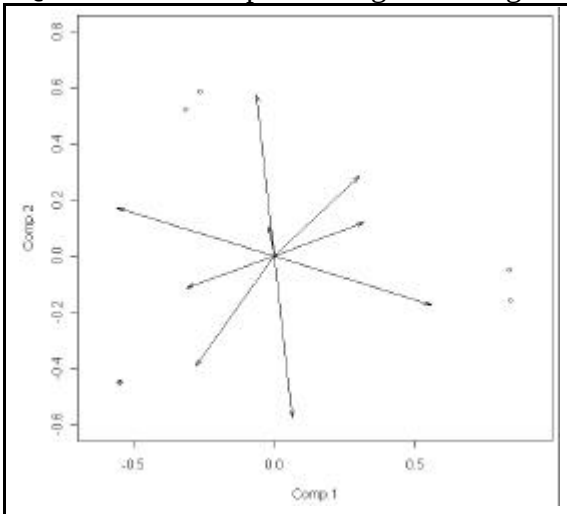
Question 15 Quel est le rang de la matrice **C** ? Donner un argument numérique **et** un argument mathématique.

Question 16 Donner une légende à la figure.



Question 17 Quelle est la propriété particulière des vecteurs propres …

Question 18 Compléter et légender la figure .



Question 19 Démontrer que pour tout matrice **A** de nombre réels $\mathbf{Au} = \mathbf{0} \Leftrightarrow \mathbf{A}^t\mathbf{Au} = \mathbf{0}$

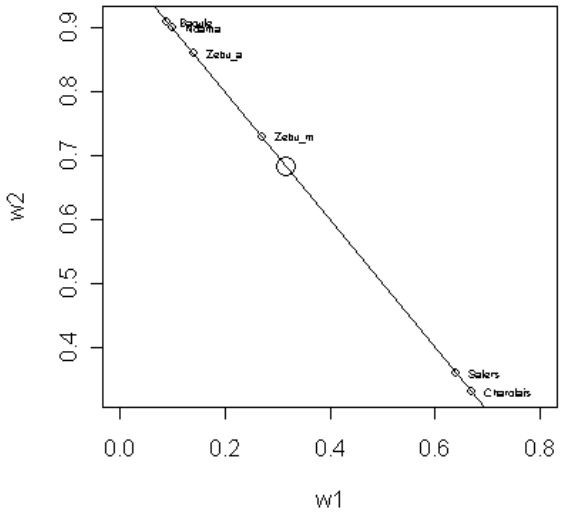
Question 20 Expliciter ...

3. Solution

Question 1

```
> plot(bovins[,9:10],xlim=c(0,0.8))
```

```
> plot(bovins[,9:10],xlim=c(0,0.8))
> text(bovins$w1,bovins$w2,row.names(bovins),pos=4,cex=0.5)
> abline(1,-1)
> points(mean(bovins$w1),mean(bovins$w2),cex=2)
```



Nuage bivarié des deux dernières colonnes du tableau bovins (système beta-lacto à deux alléles). En abscisse la fréquence du premier allèle et en ordonnée la fréquence du second allèle. Les points sont sur la droite $x + y = 1$. *Le centre de gravité est sur cette droite (moyenne des variables) et le premier axe principal est cette droite. Cette protéine isole les deux bovins d'Europe.*

Question 2

```
> apply(bovins,1,sum)
    Ndama    Baoule    Zebu_a    Zebu_m Charolais    Salers
        4         4         4         4         4         4
```

Question 3

Pour chacun des quatre sous-tableaux on a sur chaque ligne la somme de 1 :

$$\sum_{j=1}^{p} x_{ij} = 1 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij} = 1 \Rightarrow \sum_{j=1}^{p} m_j = 1$$

La somme des moyennes des variables pour chacun des 4 sous-tableaux vaut l'unité et leur somme totale vaut 4.

Question 4

```
> scale(bovins[,1:2],sca=F)
            a1    a2
Ndama      0.21 -0.21
Baoule     0.24 -0.24
Zebu_a    -0.46  0.46
Zebu_m    -0.51  0.51
Charolais  0.24 -0.24
Salers     0.28 -0.28
```

Question 5 *La somme des deux colonnes de* $\mathbf{X}_{01}$ *est nulle et plus généralement la somme des colonnes de* $\mathbf{X}_{0k}$ *est nulle car :*

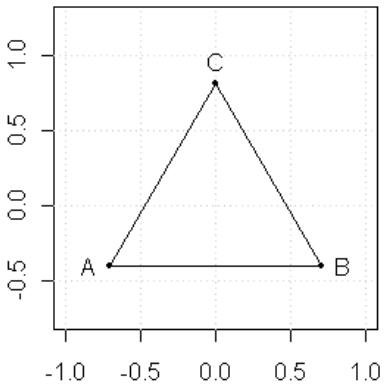
$$\sum_{j=1}^{p} x_{ij} = 1 \text{et} \sum_{j=1}^{p} m_j = 1 \Rightarrow \sum_{j=1}^{p} \left(x_{ij} - m_j \right) = 0$$

Les rangs des 4 sous-tableaux sont respectivement 1, 2, 2 et 1.

Question 6 *La remarque précédente s'écrit* $\mathbf{X}_{01}\mathbf{1}_2 = 0, \mathbf{X}_{02}\mathbf{1}_3 = 0, \mathbf{X}_{03}\mathbf{1}_3 = 0 \text{et} \mathbf{X}_{04}\mathbf{1}_2 = 0$. *On peut prendre les vecteurs colonnes de la matrice :*

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Question 7



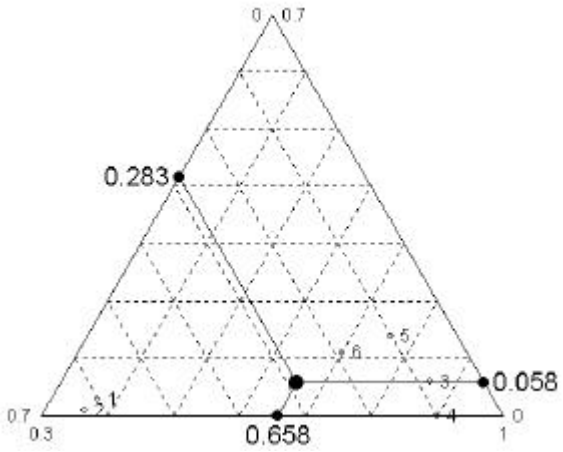
Question 8

$$AB^2=\left(\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{2}}\right)^2=2 \quad AC^2=\left(\frac{2}{\sqrt{6}}+\frac{1}{\sqrt{6}}\right)^2+\left(\frac{1}{\sqrt{2}}\right)^2=2 \quad BC^2=\left(\frac{2}{\sqrt{6}}+\frac{1}{\sqrt{6}}\right)^2+\left(\frac{1}{\sqrt{2}}\right)^2=2$$

Question 9

$$\overrightarrow{OA}=\left(-\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{6}}\right) \quad \overrightarrow{BC}=\left(-\frac{1}{\sqrt{2}},\frac{2}{\sqrt{6}}+\frac{1}{\sqrt{6}}\right)\Rightarrow\left\langle\overrightarrow{OA}\middle|\overrightarrow{BC}\right\rangle=\frac{1}{2}-\frac{1}{2}=0$$

Question 10

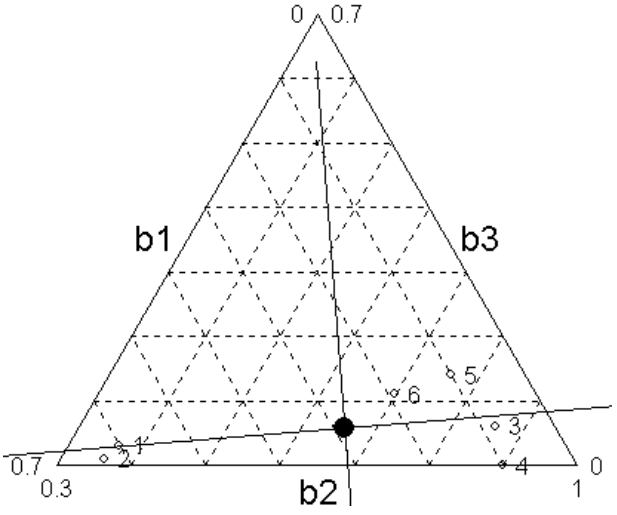



```
> plot.triangle(bovins[,3:5])
> plot.triangle(bovins[,3:5],addmean=T,labeltri=F)
```

Question 11 *Ce nuage est dans le plan* $x+y+z=1$. *L'inertie est la somme des variances soit :*

0.068+0.057+0.003=0.128

Question 12

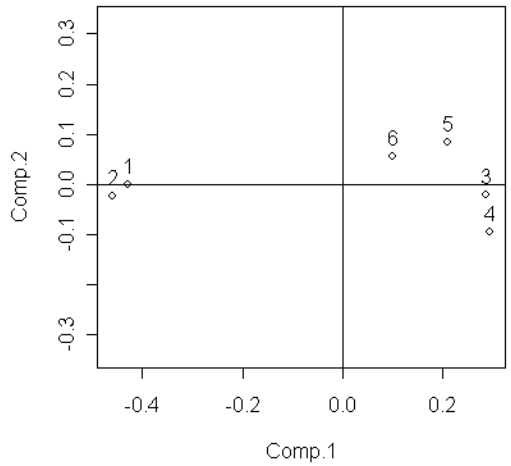


Question 13 Compléter et légender la figure obtenue par :

```
> plot(princomp(bovins[,3:5])$scores[,1:2],asp=1)

> plot(princomp(bovins[,3:5])$scores[,1:2],asp=1)
> abline(h=0,v=0)
```

```
> text(princomp(bovins[,3:5])$scores[,1],
     princomp(bovins[,3:5])$scores[,2],1:6,pos=3)
```



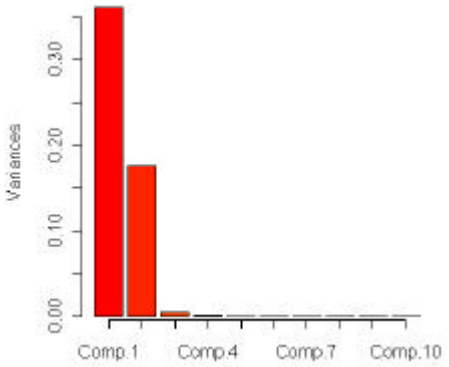
Carte factorielle de l'ACP centrée du sous-tableau $\mathbf{X}_2$ *. Le nuage de 6 individus et 3 variables est dans un plan et sur ce plan la configuration est exactement celle de la représentation triangulaire. Les axes sont portés par les axes principaux.*

Question 14 *Le tableau est 6-10 et la matrice de covariances est 10-10. Les 4 vecteurs recherchés sont ceux de la question 6 car* $\mathbf{X}_0\mathbf{u} = \mathbf{0} \Rightarrow \frac{1}{n}\mathbf{X}_0^t\mathbf{X}_0\mathbf{u} = \mathbf{0} \Rightarrow \mathbf{C}\mathbf{u} = 0$.

Question 15

Dans sdev on trouve les racines des valeurs propres. Il y a 5 valeurs propres manifestement non nulles donc le rang de ***C*** *est au moins 5. On a trouvé 4 vecteurs dans le noyau et le rang est au plus 6. En fait le tableau centré est formé de 6 vecteurs de* $\mathbb{R}^{10}$ *et de 10 vecteurs de* $\mathbb{R}^6$ *. Son rang qui est aussi celui de* ***C*** *ne peut dépasser 6. Mais la somme des 6 vecteurs de* $\mathbb{R}^{10}$ *est nulle à cause du centrage et le rang vaut 5 au plus. Il vaut donc 5 exactement*.

Question 16



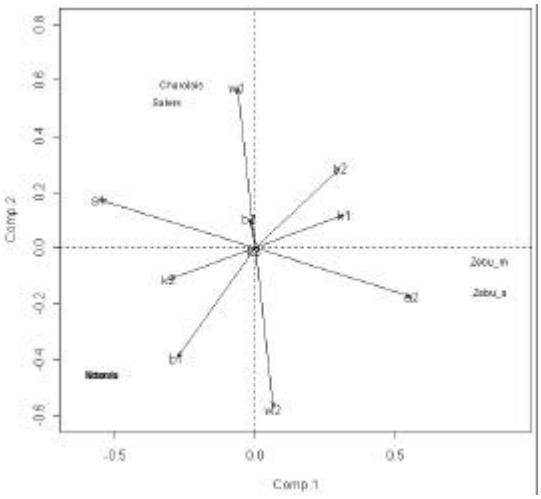
Graphe des valeurs propres du tableau ***X.*** *Les 6 points de* $\mathbb{R}^{10}$ *sont pratiquement dans un plan et la carte factorielle des individus définie par les axes 1 et 2, comme celle des variables définies par les composantes principales 1 et 2, rend compte de l'essentiel de l'information.*

Question 17

La somme des colonnes des sous-tableaux $\mathbf{X}_{0k}$ *est nulle (question 5). Les 4 vecteurs de la question 6 sont des vecteurs propres de la matrice de covariances pour la valeur propre 0. Par orthogonalité les axes principaux ont des composantes de somme nulle par blocs (par exemple*-0.28101+0.30093-0.01992=0*). On peut dire aussi : les composantes principales sont*

des combinaisons linéaires des colonnes, de même les axes principaux sont des combinaisons linéaires des lignes (la somme nulle par blocs se conserve).

Question 18



*Reporter les étiquettes. Carte factorielle des lignes du tableau **X**. Le nuage de points est très voisin de ce plan 1-2. Les 6 races forment trois groupes mais il faut plusieurs systèmes génétiques pour le mettre en évidence. Projection de la base canonique de* $\mathbb{R}^{10}$. *Les bovins d'Europe sont caractérisés par l'allèle w1, les taurins d'Afrique par b1 et les zébus par a2. Le nuage des individus est centré. Le nuage des variables est centré par tableau. On a des propriétés d'analyse factorielle multiple implicite sur ce type de tableau.*

Question 19

La condition nécessaire est évidente. La réciproque est vraie parce que :

$$\mathbf{A}^t\mathbf{A}\,\mathbf{u} = \mathbf{0} \Rightarrow \mathbf{u}^t\mathbf{A}^t\mathbf{A}\mathbf{u} = \mathbf{0} \Leftrightarrow (\mathbf{Au})^t\,\mathbf{Au} = \mathbf{0} \Leftrightarrow \langle \mathbf{Au} | \mathbf{Au} \rangle = 0 \Leftrightarrow \|\mathbf{Au}\|^2 = 0 \Leftrightarrow \mathbf{Au} = 0$$

Question 20

```
z_scale(as.matrix(bov2)%*%pr1$loadings[,1:2],scale=F)
par(mfrow=c(2,2))
par(mar=c(3,2,2,1))
ylim_range(c(-0.6,0.8))
for (i in 1:4) {
   plot(pr1$loadings[gr==i,],ylim=ylim,asp=1,type="c")
   abline(h=0,v=0,lty=1)
   text(pr1$loadings[gr==i,],names(bovins)[gr==i])
   points(pr1$scores[,1],pr1$scores[,2],type="n")
   text(pr1$scores[,1],pr1$scores[,2],row.names(bovins))
   points(z[(((i-1)*6+1) : (i*6)),],pch=20,cex=1.5)
   segments(pr1$scores[,1],pr1$scores[,2],
      z[(((i-1)*6+1) : (i*6)),1],z[(((i-1)*6+1) : (i*6)),2])
}
```

Comme chacun des nuages est pratiquement dans un sous-espace de dimension 1, soit parce qu'on a deux allèles, soit parce qu'on a en a trois dont 1 est rare, on peut lire directement la fréquence allélique sur la figure. On a trois groupes très homogènes E (pour européen), Z (pour zébus) et A (pour africains) pour l'ensemble des marqueurs. Il y a trois manières d'opposer deux des groupes au troisième. E+A contre Z (système 1); E+Z contre A (système 2), A+Z contre E (système 4). Le système 3 sépare les trois groupes. Les 4 typologies sont différentes. L'opposition A-Z est présente 3 fois sur 4 et la plus originale est la quatrième (les trois premiers loci sont sur le chromosome 6 et le quatrième sur le chromosome 11). Les ACP des tableaux de fréquences alléliques font des compromis efficaces en <u>additionnant</u> des structures indépendantes. Sur des grands ensembles de données avec des marqueurs

massivement polymorphes (microsattelites) ou des mélanges de différents types de marqueurs, ce phénomène sera caché dans une analyse simple.