TD3 : Consommation de protéines en Europe

(correction)

1 Consommation de protéines en Europe

```
> require(ade4)
> source("fonctions.R")
> protein=read.table("protein.txt")
> protein1=protein[,1:9]
> cor1=cor(protein1)
> cor1[upper.tri(cor1, diag=T)]=NA
> pca1=dudi.pca(protein1, scannf=F, nf=5)
> pca1=dudi.fixsigns(pca1, sign.co=c(-1, 1, 1, 1, 1))
> inert1=inertia.dudi(pca1, col=T, row=T)
> cest=pca1$li[protein$EST==1,]
> meanest<-sapply(cest,mean)
> nest<-nrow(cest)
> valtest<-matrix(meanest*sqrt(nest/pca1$eig[1:5]*(nrow(pca1$li)-1)/(nrow(pca1$li)-nest)), 1, 5)
> rownames(valtest)=c("EST")
> colnames(valtest)=colnames(pca1$li)
```

On considère des données qui concernent la consommation de protéines dans différents pays d'Europe en 1973.

Les pays sont : Albanie (`al`), Autriche (`at`), Belgique (`be`), Bulgarie (`bg`), Suisse (`ch`), Tchécoslovaquie (`cz`), République Démocratique d'Allemagne (`dd`), République Fédérale d'Allemagne (`de`), Danemark (`dk`), Espagne (`es`), Finlande (`fi`), France (`fr`), Grèce (`gr`), Hongrie (`hu`), Irlande (`ie`), Italie (`it`), Pays bas (`nl`), Norvège (`no`), Pologne (`pl`), Portugal (`pt`), Roumanie (`ro`), Russie (`ru`), Suède (`se`), Royaume Uni (`uk`), Yougoslavie (`yu`).

Les sources de protéines sont : viande rouge (`VROU`), viande blanche (`VBLA`), œufs (`OEUF`), lait (`LAIT`), poisson (`POISS`), céréales (`CERE`), amidon (`AMID`), légumes secs, noix et graines (`NOIX`), fruits et légumes (`FRLEG`). Ces données ont été collectées en pleine guerre froide ; on ajoute une variable `EST` qui vaut 1 pour les pays du « bloc de l'Est ».

On donne ci-dessous le tableau de données brutes, la matrice des corrélations et la représentation des 4 paires de variables (`VBLA`, `OEUF`), (`POISS`, `FRLEG`), (`OEUF`, `CERE`) et (`OEUF`, `FRLEG`).

```
> protein
```

	VROU	VBLA	OEUF	LAIT	POISS	CERE	AMID	NOIX	FRLEG	EST
al	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7	1
at	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3	0
be	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0	0
bg	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2	1
ch	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9	0
cz	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0	1
dd	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6	1
de	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8	0
dk	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4	0
es	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2	0
fi	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4	0
fr	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5	0
gr	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5	0
hu	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2	1
ie	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9	0
it	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7	0
nl	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7	0
no	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7	0
pl	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6	1
pt	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9	0
ro	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8	1
ru	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9	1
se	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0	0
uk	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3	0
yu	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2	1

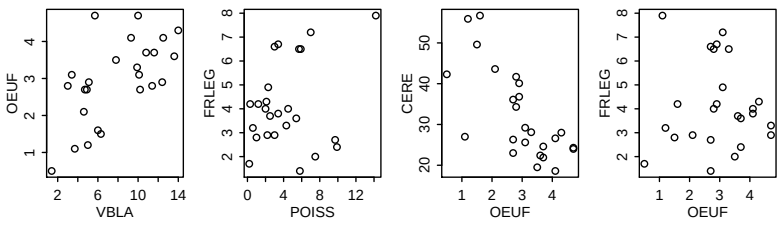
Corrélations

```
> round(cor1,2)
```

	VROU	VBLA	OEUF	LAIT	POISS	CERE	AMID	NOIX	FRLEG
VROU	NA	NA	NA	NA	NA	NA	NA	NA	NA
VBLA	0.15	NA	NA	NA	NA	NA	NA	NA	NA
OEUF	0.59	0.62	NA	NA	NA	NA	NA	NA	NA
LAIT	0.50	0.28	0.58	NA	NA	NA	NA	NA	NA
POISS	0.06	-0.23	0.07	0.14	NA	NA	NA	NA	NA
CERE	-0.50	-0.41	-0.71	-0.59	-0.52	NA	NA	NA	NA
AMID	0.14	0.31	0.45	0.22	0.40	-0.53	NA	NA	NA
NOIX	-0.35	-0.63	-0.56	-0.62	-0.15	0.65	-0.47	NA	NA
FRLEG	-0.07	-0.06	-0.05	-0.41	0.27	0.05	0.08	0.37	NA

Représentation de paires

```
> op=par(mfcol=c(1,4), mex=0.5, mar=c(5,5,4,2), oma=c(0,0,2,0))
> attach(protein)
> plot(VBLA,OEUF)
> plot(POISS,FRLEG)
> plot(OEUF,CERE)
> plot(OEUF, FRLEG)
> detach(protein)
> par(op)
```



1.1 Premier regard sur les données

Question 1 *Il y a toute un partie de données qui sont manquantes dans la matrice de corrélation (marquées NA). Expliquez comment on peut les retrouver.*

Les valeurs de la diagonale sont connues : elles valent toutes 1 (corrélation d'une variable avec elle-même). Pour les valeurs au dessus de la diagonale, on utilise le fait que la matrice est symétrique : La valeur pour (VBLA, VROU), par exemple (seconde colonne, première ligne), est égale à la corrélation de (VROU, VBLA), c'est-à-dire 0.15.

Question 2 *Quelles sont les variables qui sont particulièrement corrélées ou décorrélées ?*

— Les variables plus corrélées positivement sont NOIX et CERE $(0,65)$;
— Les variables plus corrélées négativement sont CERE et OEUF $(-0,71)$;
— Les moins corrélées sont FRLEG et OEUF $(-0,05)$ ou FRLEG et CERE $(0,05)$.

Question 3 *Pour chacun des couples (VBLA, OEUF), (POISS, FRLEG), (OEUF, CERE) et (OEUF, FRLEG), commentez la répartition des valeurs, identifiez les éventuels individus « anormaux » et expliquez le lien avec les corrélations mesurées.*

— Le couple (VBLA, OEUF) correspond à une corrélation positive mais pas très importante de $0,62$; la forme du nuage correspond à une corrélation positive de qualité moyenne. On constate en effet que les coins supérieur gauche et inférieur droit du carré sont assez vides.
— Le couple (POISS, FRLEG) correspond à une corrélation plutôt faible de $0,27$; la forme du nuage de point est difficile à interpréter à cause de la valeur anormale dans le coin en haut à droite. Après examen des données, on voit qu'il s'agit de pt.
— Le couple (CERE, OEUF) correspond à une corrélation négative de $-0,71$; on voit en effet sur le nuage de point que les coins supérieur droit et dans une moindre mesure inférieurs gauche sont vides, ce qui correspond à une direction anticorrélée.
— Enfin, le couple (FRLEG, OEUF) correspond à une très faible corrélation $(-0,05)$; le nuage est assez réparti dans le carré comme on peut s'y attendre.

1.2 Analyse en composante principale

On met pour l'instant de côté la variable EST. On obtient les données suivantes par ACP sur les données centrées-réduites sur les variables restantes : valeurs propres, corrélations avec les 5 premiers axes, valeurs test pour les 5 premiers axes, coordonnées des individus sur les 5 premiers axes, et enfin que la qualité de leur représentation par les 5 premiers axes principaux (en %).

Valeurs propres

```
> round(pca1$eig,digits=2)

[1] 4.01 1.63 1.13 0.95 0.46 0.33 0.27 0.12 0.10
```

Corrélations

```
> round(pca1$co,2)

      Comp1 Comp2 Comp3 Comp4 Comp5
VROU  -0.61  0.07  0.32  0.63  0.22
VBLA  -0.62  0.30 -0.66 -0.04 -0.20
OEUF  -0.85  0.05 -0.19  0.31  0.05
LAIT  -0.76  0.24  0.41  0.00 -0.14
POISS -0.27 -0.83  0.34 -0.21 -0.20
CERE   0.88  0.30 -0.10 -0.01  0.16
AMID  -0.59 -0.45 -0.26 -0.33  0.50
NOIX   0.84 -0.18  0.06  0.32  0.10
FRLEG  0.22 -0.69 -0.43  0.45 -0.16
```

Valeurs test pour variable EST

```
> round(valtest,2)

    Axis1 Axis2 Axis3 Axis4 Axis5
EST  2.77  1.84 -1.64 -1.79  1.54
```

```
> round(pca1$li,2)

   Axis1 Axis2 Axis3 Axis4 Axis5
al  3.56  1.66  1.80  0.23  0.02
at -1.45  1.06 -1.37  0.17 -0.95
be -1.66 -0.16 -0.22  0.53  0.77
bg  3.20  1.33 -0.15  0.22 -0.49
ch -0.93  0.77  0.16  1.19 -0.85
cz -0.38  0.62 -1.22 -0.47  0.26
dd -1.45 -0.46 -1.33 -1.16  0.43
de -2.14  0.30 -0.82  0.11 -0.07
dk -2.41 -0.29  0.77 -0.99 -0.77
es  1.34 -2.61 -0.53  0.37  0.53
fi -1.60  0.61  2.09 -1.44  0.04
fr -1.52 -0.80  0.00  2.00  0.26
gr  2.29 -1.02  0.90  1.83 -0.41
hu  1.49  0.83 -1.95 -0.22 -0.04
ie -2.72  0.78  0.02  0.44  1.04
it  1.57 -0.41 -0.13  1.25 -0.82
nl -1.68  0.93 -0.78 -0.13 -0.78
no -0.99 -0.84  1.74 -1.16 -0.42
pl -0.12 -0.54 -1.51 -0.47 -0.02
pt  1.74 -4.38 -0.04 -0.91 -0.39
ro  2.81  1.14 -0.07 -0.63  0.32
ru  0.80  0.11  0.38 -0.95  1.70
se -1.67  0.21  1.31 -0.75 -0.84
uk -1.77  0.10  1.18  1.77  1.11
yu  3.70  1.06 -0.21 -0.84  0.39
```

```
> round(abs(inert1$row.rel),1)

    Axis1 Axis2 Axis3 Axis4 Axis5
al  61.2  13.4  15.6   0.3   0.0
at  33.8  18.1  29.9   0.5  14.6
be  70.9   0.7   1.3   7.3  15.4
bg  74.1  12.8   0.2   0.3   1.8
ch  20.6  14.0   0.6  33.9  17.1
cz   4.6  12.1  47.5   7.1   2.2
dd  32.6   3.3  27.3  20.8   2.9
de  79.6   1.6  11.7   0.2   0.1
dk  66.3   1.0   6.7  11.1   6.7
es  17.2  65.3   2.7   1.3   2.7
fi  23.7   3.4  40.8  19.5   0.0
fr  25.4   7.1   0.0  43.9   0.7
gr  42.7   8.5   6.6  27.4   1.4
hu  27.6   8.7  47.6   0.6   0.0
ie  77.6   6.4   0.0   2.1  11.3
it  44.6   3.0   0.3  28.3  12.2
nl  53.0  16.4  11.6   0.3  11.4
no  15.7  11.2  48.1  21.4   2.9
pl   0.3   6.0  46.2   4.5   0.0
pt  12.7  80.5   0.0   3.5   0.6
ro  80.1  13.2   0.1   4.0   1.1
ru  12.8   0.3   2.9  18.0  58.3
se  45.4   0.7  27.9   9.2  11.4
uk  33.3   0.1  14.7  33.3  13.0
yu  86.1   7.1   0.3   4.4   0.9
```

```
> #round(inert1$row.cum[,-ncol(inert1$row.
```

Question 4 *Commentez la répartition de l'inertie : combien d'axes principaux voudrait-t-on retenir ? quelle alors est la qualité globale de représentation ?*

On sait d'après la règles de Kaiser qu'on peut conserver les axes associés aux valeurs propres supérieures à 1. On a ici a priori trois axes acceptables.

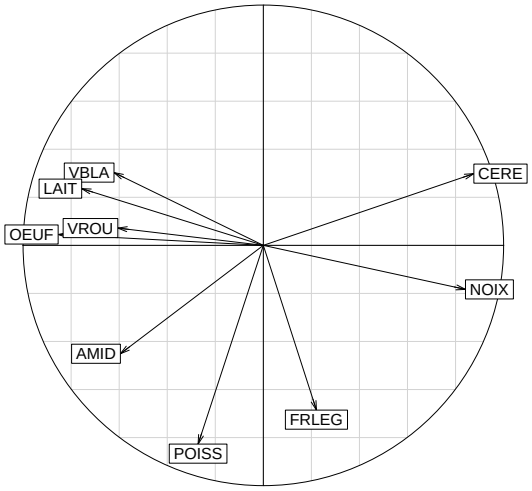
L'inertie expliquée vaut alors $6,77$, soit $75\% I_{\mathbf{g}}$ (l'inertie totale est 9). Cette valeur est correcte.

Si on avait décidé d'être un peu plus strict et de ne conserver que 2 axes, l'inertie expliquée descendrait à $63\% I_{\mathbf{g}}$. De toute façon, le nombre d'axes conservés dépend *in fine* des possibilités d'interprétation.

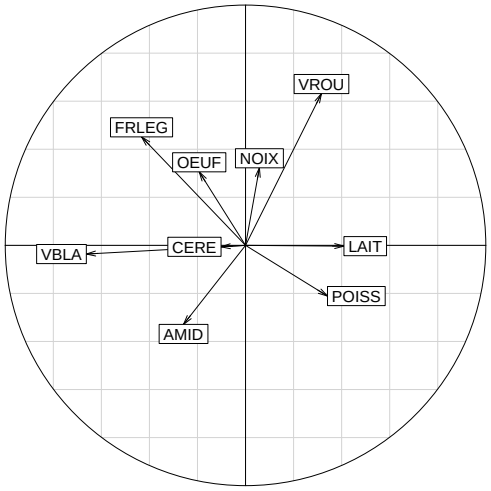
Question 5 *Quelles sont les sources de protéines qui déterminent les axes que l'on retient ? Précisez les critères utilisés. Y a-t-il un effet de taille ?*

Pour rendre la correction plus claire, on donne ici les cercles des corrélations pour les deux premiers plans principaux. Il n'est bien sûr pas demandé de faire cette représentation.

```
> s.corcircle(pca1$co)
```




```
> s.corcircle(pca1$co, 3, 4)
```



On propose de se limiter aux variables présentant une corrélation supérieure à $0,60$ pour tout les axes. Ceci nous donne les tableaux suivants

Axe 1

-	+
OEUF $(-0,85)$	CERE $(0,88)$
LAIT $(-0,76)$	NOIX $(0,84)$
VBLA $(-0,62)$	
VROU $(-0,61)$	
[AMID $(-0,59)$]	

Axe 2

-	+
POISS $(-0,83)$	
FRLEG $(-0,69)$	

Axe 3

-	+
VBLA $(-0,66)$	

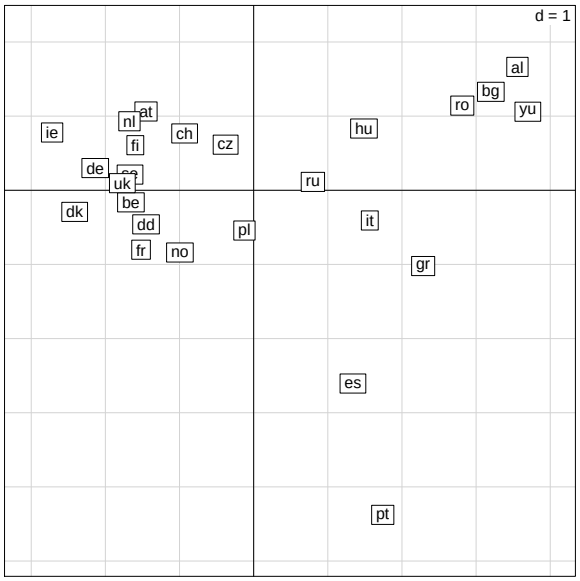
On aurait pu choisir une corrélation limite de $0,70$, qui paraît plus raisonnable. Toutefois, cela limite les variables explicatives (la différence est marquée sur l'axe 1 par une ligne de séparation) ; sur l'axe 3, on se retrouve presque sans variable. Ce n'est donc pas très satisfaisant.

On a ici clairement pas d'effet de taille, puisque les variables ont des corrélations positive et négatives avec le premier axe. Il n'y a pas de raison particulière ici de modifier les variables pour en obtenir un. Il ne peut pas y avoir de consensus entre les individus, puisque les différentes sources de protéines sont en pourcentage, et donc en compétition les unes contre les autres.

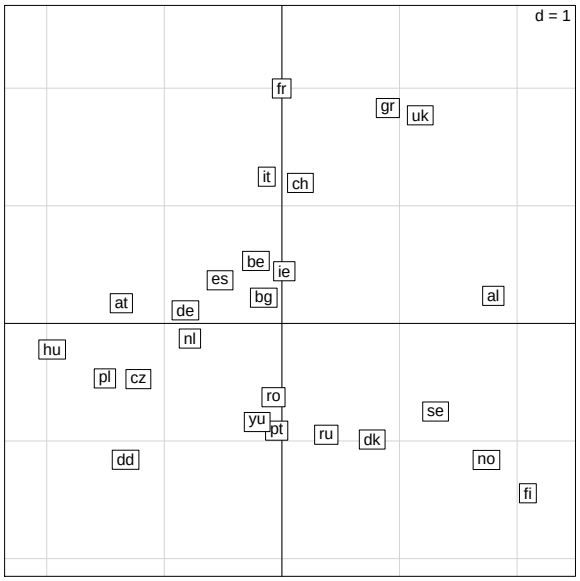
Question 6 *Quels sont les pays d'Europe qui déterminent les axes que l'on retient ? Précisez les critères utilisés.*

Là encore, on donne les projections des individus sur les deux premiers plans principaux (qui ne sont pas indispensables pour la correction) :

```
> s.label(pca1$li)
```




```
> s.label(pca1$li, 3, 4)
```



On remarque que les contributions des individus aux axes ne sont pas fournies ici. Toutefois, on sait que l'on peut raisonner sur les coordonnées de la manière suivante : sachant que l'on s'intéresse aux individus i dont la contribution à l'axe k est supérieure à α fois le poids

$$p_i \frac{(c_{ik})^2}{\lambda_k} > \alpha p_i,$$

il est équivalent de s'intéresser aux individus dont les coordonnées satisfont

$$|c_{ik}| > \sqrt{\alpha \lambda_k}.$$

Comme la qualité de la représentation globale est moyenne et qu'il y a peu d'individus, on propose ici de prendre $\alpha = 2$, ce qui donne les limites respectives de $2,83$, $1,81$, et $1,50$ pour les trois premiers axes. Avec ce critère, les pays déterminants sont

Axe 1

-	+
	`yu` $(3,70)$
	`al` $(3,56)$
	`bg` $(3,20)$
	`[ro` $(2,81)]$

Axe 2

-	+
`pt` $(-4,38)$	
`es` $(-2,61)$	

Axe 3

-	+
`hu` $(-1,95)$	`fi` $(2,09)$
`pl` $(-1,51)$	`al` $(1,80)$
	`no` $(1,74)$

En choisissant un facteur $\alpha = 3$, on aurait les limites : $3,47$, $2,21$, $1,84$. Les éléments en dessous des lignes horizontales seraient ignorés. On voit qu'il n'y a pas vraiment de gain ici, on perd plutôt en possibilités d'interprétation.

Question 7 *Comment peut-on interpréter les axes à partir des deux questions précédentes ?*

Les axes ne sont pas très simples à interpréter, surtout le troisième. Voilà toutefois ce qu'on peut dire :

- L'axe 1 met en évidence la Yougoslavie, l'Albanie, la Bulgarie et la Roumanie, qui on une faible consommation de protéines d'origine animales (œufs et lait surtout) et une forte consommation de protéines d'origine végétale (céréales, noix, légumes secs). On notera que les pays concernés sont des pays du bloc de l'Est [1].
- L'axe 2 met en évidence le Portugal et l'Espagne, qui ont une forte consommation de poisson et de fruits et légumes.
- Axe 3 : la Hongrie et la Pologne ont une consommation importante de viande blanche, contrairement à la Finlande, l'Albanie et la Norvège

Question 8 *Quels sont les cinq pays dont la qualité de représentation est mauvaise sur l'espace propre retenu ? Précisez les critères utilisés.*

On sélectionne ici les pays dont la qualité de représentation par les trois premiers axes est inférieure à 50% (la limite proposée dans le cours). Il faut ici additionner les trois premières colonnes.

On trouve alors `ru` (16%), `fr` (32,5%), `ch` (35,2%), `it` (47,9%) et `uk` (48,1%). `ru` et `ch` sont tous deux proches du centre de gravité, et on ne peut dont pas conclure sur leur qualité. Par contre `uk` et, dans une certaine mesure `fr` et `it` (pour le premier axe), sont plus éloignés ; on peut donc conclure qu'ils sont mal représentés par l'analyse.

Question 9 *Que peut-on dire du Portugal ? Calculer sa contribution à l'axe qui nous intéresse et donner des pistes sur la manière dont on aurait pu le traiter. Peut-on dire qu'il n'a pas sa place dans l'analyse ?*

Le Portugal est très éloigné sur le second axe, ce qui parait exagéré. On soupçonne que c'est un individu sur-représenté. Pour le vérifier, on calcule sa contribution à chaque axe k avec la formule

$$\frac{1}{25} \frac{(c_{20,k})^2}{\lambda_k}.$$

On obtient les valeurs suivantes en % :

```
> contr=round(pca1$li[20,1:3],2)^2/round(pca1$eig[1:3], 2)/nrow(pca1$li)
> round(contr*100, 1)

   Axis1 Axis2 Axis3
pt     3  47.1     0
```

On considère en général qu'un individu est surreprésenté quand sa contribution sur un axe est supérieure à 25%. C'est certainement le cas pour `pt` sur le second axe. La méthode habituelle dans ce cas est de refaire l'analyse sans l'individu et de réintroduire ce dernier après.

On ne peut pourtant pas dire que l'individu n'a pas sa place dans l'analyse, puisque retirer complètement le pays de l'étude nous ferait perdre de l'information.

1. on rappelle que les pays de l'est sont indiqués par la variable `EST`.

1.3 Variable supplémentaire : le bloc de l'est

Comme il a été dit en introduction, ces variables on été mesurées en pleine guerre froide. Il parait donc intéressant de regarder ce que l'on peut dire du bloc des pays de l'est (variable `EST`). Pour cela on utilise les valeurs test données plus haut.

Question 10 *Avez vous toutes les données nécessaires dans cet énoncé pour le calcul des valeurs test ? Détaillez.*

On utilise la formule

$$\text{Valeur test pour axe } k = \bar{c}_{\text{est},k}\sqrt{\frac{n_{\text{est}}}{\lambda_k}}\sqrt{\frac{n}{n-n_{\text{est}}}},$$

avec les valeurs suivantes :

- $n = 25$;
- $n_{\text{est}} = 9$;
- $\bar{c}_{\text{est},k} = \frac{1}{n_{\text{est}}}\sum_{i\in\text{est}} c_{ik}$, où c_{ik} est donné dans le tableau des coordonnées sur les axes ;
- λ_k est la k-ième valeur propre ;

Question 11 *Les conditions d'utilisation des valeurs test sont elles réunies ? En laissant de coté les problèmes éventuels, expliquez ce que les valeurs test nous apprennent.*

Pour que les conditions d'utilisation des valeurs test soient réunies, il faudrait (1) avoir au moins 30 pays et (2) avoir au moins 30 pays de l'Est. La première condition est presque remplie, la seconde ne l'est visiblement pas du tout. D'autre part, il faut qu'elle concerne une variable qualitative (c'est le cas). Les valeurs test qu'on peut interpréter sont celles qui dépassent 2 ou 3 en valeur absolue.

Comme demandé dans la question, on fait comme si les effectifs n'étaient pas un problème. On voit clairement que les pays du bloc de l'Est sont liés positivement avec l'axe 1 seulement. On peut en déduire que ces pays avaient une plus fort développement de l'activité de culture que d'élevage en ce qui concerne les sources de protéines.

2 Variables liées et valeurs propres

On cherche à montrer qu'à chaque fois que des variables sont liées par une relation linéaire, l'analyse en composante principale des données correspondantes produit une valeur propre nulle. On considère pour cela une table $\mathbf{X} = (x_i^j)$ de données avec n individus et p variables, munie d'une matrice de poids $\mathbf{D_p}$. On note $\mathbf{Y}$ la matrice des variables centrées et $\mathbf{V} = \mathbf{Y'D_pY}$ la matrice de variance-covariance. On effectue une ACP sur ces données en utilisant une métrique $\mathbf{M} = \text{diag}(m_1, \ldots, m_p)$.

Question 12 *On suppose qu'il existe des coefficients non tous nuls $w_1, \ldots, w_p$ et w_0 tels que, pour tout i,*

$$w_1 x_i^1 + w_2 x_i^2 + \ldots + w_p x_i^p = w_0.$$

Montrer que l'ACP possède un axe propre $\mathbf{a}$ *associé à une valeur propre nulle, c'est-à-dire satisfaisant* $\mathbf{VMa} = \mathbf{0}$.

Les moyennes arithmétiques des variables vérifient

$$\begin{aligned} w_1\bar{x}^1 + w_2\bar{x}^2 + \ldots + w_p\bar{x}^p &= w_1\frac{1}{n}\sum_{i=1}^n x_i^1 + \ldots + w_p\frac{1}{n}\sum_{i=1}^n x_i^p \\ &= \frac{1}{n}\sum_{i=1}^n \left(w_1 x_i^1 + \ldots + w_p x_i^p\right) \\ &= \frac{1}{n}\sum_{i=1}^n w_0 = w_0. \end{aligned}$$

Le variables centrées vérifient donc pour chaque individu i

$$w_1(x_i^1 - \bar{x}^1) + w_2(x_i^2 - \bar{x}^2) + \ldots + w_p(x_i^p - \bar{x}^p) = w_0 - w_0 = 0,$$

c'est-à-dire $\mathbf{Yw} = \mathbf{0}$, où $\mathbf{w} = (w_1, \ldots, w_p)$.

Si on pose $\mathbf{a} = \mathbf{M}^{-1}\mathbf{w}$, on a alors $\mathbf{VMa} = \mathbf{Y'D_pYMM}^{-1}\mathbf{w} = \mathbf{Y'D_pYw} = \mathbf{0}$.