TD7 : association d'antalgiques

(Correction)

```
> require(ade4)
> source("fonctions.R")
> antal1 = read.table("antal.dat")
> stats1 = stats.conting(antal1)
> #coa1 = dudi.coa(antal1,scannf=F,nf=4)
> #inert1 = inertia.dudi(coa1,r=T,c=T)
> # On enlève dou5 qui est 2*dou4+1 (pourquoi ??)
> antal2 = antal1[,1:5]
> stats2 = stats.conting(antal2)
> coa2 = dudi.coa(antal2,scannf=F,nf=3)
> inert2 = inertia.dudi(coa2,r=T,c=T)
```

On considère 160 patients ayant subi une extraction dentaire entraînant généralement une douleur vive et ne cédant qu'après de longues heures. Pour comparer l'effet de deux médicaments antalgiques (contre la douleur) et de leur association, on on les divise en 4 groupes de 40 et on les soumet à des traitements « en double insu », c'est-à-dire que ni le patient ni le docteur ne savent de quel traitement il s'agit [1]. Les quatre traitements sont notés A, B, AB et P. AB est en fait l'association de A et B, P est un placebo (médicament factice, utilisé comme situation de référence). On note pour chaque médicament l'évolution de la douleur selon une échelle de temps repérée par les identificateurs de la réaction des patients. La notation est la suivante :

- `A15m` : après 15 minutes pour le traitement A,
- `P2h` : après 2 heures pour le placebo,
- `ABp4h` après au moins 4h de traitement joint A et B (on utilise `p4h` parce que le résultat est le même après 4h, 5h ou 6h),
- etc.

Les colonnes signifient respectivement douleur nulle (`dou0`), douleur légère (`dou1`), douleur modérée (`dou2`), sévère (`dou3`), très sévère (`dou4`). Les données sont fournies ci-dessous, ainsi que les profils marginaux de colonne.

```
>   antal2[1:17,]

      dou0 dou1 dou2 dou3 dou4
A0h      0    0    0   28   12
A15m     0    2    4   28    6
A30m     0    4   16   18    2
A1h      6    4   18   10    2
A2h      6   10   12    8    4
A3h      6   10    8   12    4
A4h      4   12    4   16    4
Ap5h     0   14    6   16    4
B0h      0    0    0   22   18
B15m     0    4    6   22    8
B30m     0   12   10   16    2
B1h      0   14   16    8    2
B2h      4   16   10    8    2
B3h      4   12   12   10    2
B4h      2   12   16   10    0
B5h      2   14   10   14    0
B6h      0   18    8   14    0
```

```
>   antal2[18:33,]

       dou0 dou1 dou2 dou3 dou4
AB0h      0    0    0   18   22
AB15m     0    4    6   20   10
AB30m     0    4   20   14    2
AB1h      4   12   18    6    0
AB2h      4   20   12    4    0
AB3h     12   14    7    7    0
ABp4h    12   10   10    8    0
P0h       0    0    0   22   18
P15m      0    0    0   22   18
P30m      0    0    0   34    6
P1h       2    4   16   14    4
P2h       2    8    8   10    4
P3h       4    8    8   14    6
P4h       4    8    4   16    8
P5h       2    8    8   10   12
P6h       2    6   12   10   10
```

Profils marginaux (%)

```
>  round(stats2$mc/stats2$tot*100,1)

dou0 dou1 dou2 dou3 dou4
 6.2 20.1 21.7 37.3 14.6
```

Sous tableau douleur 2/3/4 à 15 minutes

```
> antal.sub = antal2[c(2, 10, 19, 26),
> stats.sub = stats.conting(antal.sub)
> antal.sub

      dou2 dou3 dou4
A15m     4   28    6
B15m     6   22    8
AB15m    6   20   10
P15m     0   22   18
```

1 Les données

Question 1 *Montrez que la phrase suivante est vraie : « au bout de 6 heures, la proportion des patients traités par B ressentant une douleur sévère est supérieure à la proportion des patients ayant reçu un placebo qui ressentent une douleur sévère ». Expliquez pourquoi cette phrase ne remet pas en cause l'efficacité de B.*

Le nombre de patients ayant reçu B et ressentant une douleur sévère (`dou3`) après 6 heures est 14 ; en le ramenant à l'effectif de la modalité `B6h` (40), on obtient une proportion de $14 \div 40 = 35\%$. Pour les patients ayant reçu un placebo et ressentant une douleur sévère après 6 heures, le même calcul donne $10 \div 40 = 25\%$. Cela confirme la phrase de la question.

Ce résultat ne veut bien sûr rien dire, puisque pour les niveaux de douleur supérieurs, les patients ayant reçu le placebo sont plus nombreux : si on considère les douleurs sévères ou plus (`dou3`, `dou4`), par exemple, elles sont ressenties au bout de 6 heures par $14 \div 40 = 35\%$ des patients ayant pris B, et par $20 \div 40 = 50\%$ des patients ayant pris un placebo.

1. G. D. Maïti, Étude comparative de l'efficacité de deux médicaments antalgiques et de leur association, *Les cahiers de l'analyse des données*, tome 14, n° 2 (1989), p. 157-162.

Question 2 *On s'intéresse à l'effet des différents antalgiques sur douleurs modérées à très sévères (*`dou2`*,* `dou3` *et* `dou4`*) au bout de 15 minutes (voir tableau dans les données). Deux manières alternatives de traiter la question :*

1. *La valeur du χ^2 pour ce tableau est 14.68. À l'aide du la table du χ^2, peut on dire (à 5%, puis à 1%) que l'antalgique pris est efficace sur les douleurs fortes au bout de 15 minutes ?*
2. *La p-value correspondante au tableau est 2.3%. Quelle conclusion peut-on tirer ?*

La dimension de la table étant 4×3, on s'intéresse ici à la loi du χ^2 à $(4-1)\times(3-1) = 6$ degrés de liberté. En se reportant à la table fournie, on voit que les valeurs critiques pour 5% et 1% sont respectivement $12,592$ et $16,812$.

La signification des ces valeurs est que $P(\chi^2_6 \geq 12,592) = 0.05$ et $P(\chi^2_6 \geq 16,812) = 0.01$. Ces distributions reflètent ce que l'on obtient sous l'hypothèse d'indépendance des lignes et des colonnes, c'est-à-dire en supposant que les antalgiques ne jouent pas sur les douleurs fortes au bout de 15 minutes (en particulier si le type d'antalgique ou le placebo pris ont une influence). Si la valeur du χ^2 était suffisamment grande, on pourrait déduire que les antalgiques ont un effet.

Ici on ne peut pas conclure que les antalgiques ont un effet à 1% : la valeur 14.68 est plus petite que $16,812$. Par contre contre, si on se contente de 5% d'erreur, on peut conclure que les antalgiques marchent. Ce résultat en demi-teinte est normal, car on n'a pas attendu très longtemps après l'ingestion du médicament.

2 Analyse factorielle des correspondances

On réalise l'AFC du tableau. On fournit ci-dessous les valeurs propres associées aux axes principaux et, pour les 3 premiers axes

— les coordonnées sur les axes, les contributions aux axes et la qualité de la représentation par les sous espaces factoriels pour les profils colonnes,

— les coordonnées sur les axes, et la qualité de la représentation par les sous espaces factoriels pour les profils lignes.

Valeurs propres

```
>   round(coa2$eig,2)

[1] 0.35 0.08 0.05 0.05
```

Coordonnées colonnes

```
>   round(coa2$co,2)
```

	Comp1	Comp2	Comp3
dou0	-0.77	0.94	0.18
dou1	-0.59	0.03	-0.31
dou2	-0.52	-0.28	0.32
dou3	0.38	-0.10	-0.12
dou4	0.93	0.23	0.18

Contrib. colonnes (%)

```
> colnames(inert2$col.abs)=paste0("Axis", 1:3)
> round(inert2$col.abs,1)
```

	Axis1	Axis2	Axis3
dou0	10.5	65.7	3.7
dou1	20.3	0.2	36.5
dou2	16.8	19.9	41.5
dou3	15.8	4.7	9.7
dou4	36.6	9.5	8.6

Qualité colonnes (%)

```
>  round(abs(inert2$col.cum)[,1:3],1)
```

	Axis1	Axis1:2	Axis1:3
dou0	36.6	91.7	93.7
dou1	69.9	70.1	89.6
dou2	59.9	77.0	100.0
dou3	69.5	74.5	81.1
dou4	82.7	87.9	90.9

Coord. lignes (%)

```
>   round(coa2$li,2)
```

	Axis1	Axis2	Axis3
A0h	0.93	-0.01	-0.13
A15m	0.55	-0.22	-0.17
A30m	-0.08	-0.49	0.23
A1h	-0.45	0.02	0.51
A2h	-0.42	0.23	0.17
A3h	-0.27	0.29	-0.02
A4h	-0.10	0.20	-0.32
Ap5h	-0.06	-0.17	-0.39
B0h	1.07	0.17	0.07
B15m	0.44	-0.17	-0.05
B30m	-0.18	-0.31	-0.22
B1h	-0.49	-0.38	0.02
B2h	-0.54	0.09	-0.18
B3h	-0.45	0.02	0.00
B4h	-0.56	-0.28	0.06
B5h	-0.41	-0.17	-0.27
B6h	-0.40	-0.27	-0.51
AB0h	1.16	0.28	0.19
AB15m	0.49	-0.11	0.01
AB30m	-0.23	-0.55	0.42
AB1h	-0.73	-0.13	0.22
AB2h	-0.83	0.05	-0.23
AB3h	-0.78	0.78	-0.09
ABp4h	-0.73	0.69	0.14
P0h	1.07	0.17	0.07
P15m	1.07	0.17	0.07
P30m	0.79	-0.18	-0.32
P1h	-0.13	-0.25	0.36
P2h	-0.15	-0.02	-0.01
P3h	-0.04	0.15	0.02
P4h	0.16	0.27	-0.10
P5h	0.20	0.14	0.15
P6h	0.08	0.00	0.32

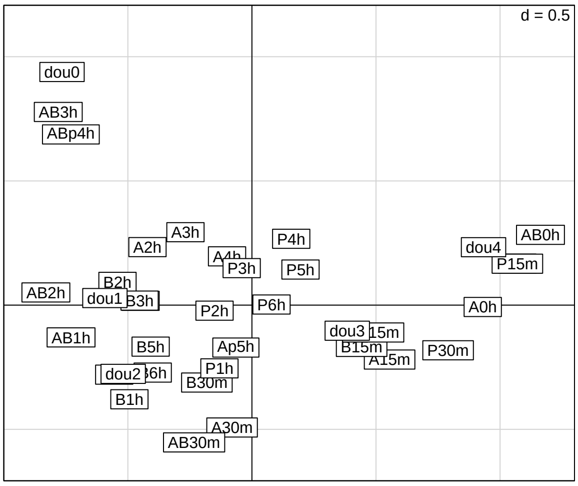
Qualité lignes (%)

```
>   round(abs(inert2$row.cum)[,1:3],1)
```

	Axis1	Axis1:2	Axis1:3
A0h	93.1	93.1	94.8
A15m	58.4	67.5	73.0
A30m	1.8	71.8	86.8
A1h	38.2	38.3	88.4
A2h	68.0	88.9	99.8
A3h	43.6	96.0	96.3
A4h	6.7	32.2	98.3
Ap5h	2.0	15.9	88.6
B0h	95.7	98.1	98.4
B15m	82.0	93.9	95.1
B30m	18.2	71.7	99.3
B1h	51.9	82.5	82.6
B2h	79.9	82.3	91.1
B3h	99.4	99.5	99.5
B4h	78.7	98.7	99.6
B5h	62.9	73.4	100.0
B6h	30.8	45.0	94.7
AB0h	84.0	88.9	91.3
AB15m	95.1	99.9	99.9
AB30m	9.9	65.8	97.6
AB1h	88.6	91.4	99.4
AB2h	81.7	82.1	88.5
AB3h	47.9	95.4	96.0
ABp4h	46.5	87.7	89.4
P0h	95.7	98.1	98.4
P15m	95.7	98.1	98.4
P30m	57.2	60.2	69.5
P1h	7.7	36.7	93.3
P2h	76.7	78.4	78.6
P3h	6.7	94.3	96.0
P4h	23.3	89.7	99.8
P5h	18.6	28.6	39.4
P6h	3.9	3.9	65.9

Premier plan principal

```
> s.label(coa2$li,xlim=c(-1,1.3),ylim=c(-0.5,1))
> s.label(coa2$co, add.p=T)
```



Question 3 *Pourquoi y a-t-il 4 valeurs propres ? Expliquez comment on détermine le nombre d'axes à conserver et calculez la proportion d'inertie totale expliquée.*

Le nombre de valeurs propres dépend du nombre de modalités de chaque variable : ici la formule est $\min(33-1, 5-1) = 4$. Quand le nombre de valeurs propres est si faible, on se contente en général de conserver 2 axes. La répartition des valeurs propres pourrait en fait conduire à ne garder que le premier. On décide d'en garder deux pour des raisons de représentation (on verra si on peut interpréter le second).

La somme des valeurs propres (l'inertie totale) est $0,53$. En conservant 2 axes, on représente $0,35 + 0,08 = 0,43 = 80\%$ de l'inertie totale.

Question 4 *Quelles sont les modalités de ligne et colonne qui déterminent les deux premiers axes principaux ? (on détaillera les critères et on cherchera à être précis dans la réponse)*

On commence par les colonnes. Pour déterminer les contributions notables, on peut comparer les contributions des profils colonne aux axes avec les poids de ces profils (les profils marginaux). Les modalités vraiment significatives sont celles dont la contribution est au moins 2 fois le poids.

Premier axe :
- en négatif : même `dou0` n'est pas significative ;
- en positif : dou4 ($36,6\% > 2 \times 14,6\%$).

Deuxième axe :
- en négatif : aucune modalité n'est significative ;
- en positif : seule `dou0` ($65,7\% \gg 2 \times 6,2\%$) est significative, et de manière très nette.

Notons qu'il aurait été possible de raisonner sur les coordonnées des modalités. C'est ce que nous ferons pour les lignes.

La contribution de chaque modalité en ligne (de coordonnée a_i) à un axe factoriel (associé à la valeur propre λ) s'écrit

$$\frac{n_{i\cdot}}{n}\frac{(a_i)^2}{\lambda}$$

et doit être comparée à son poids $n_{i\cdot}/n$. Si on cherche celles pour lesquelles le rapport est supérieur à 2, par exemple, un calcul simple montre qu'il faut s'intéresser aux modalités dont les coordonnées sur l'axe vérifient

$$|a_i| > \sqrt{2\lambda}.$$

Les résultats obtenus sont les mêmes que si on avait raisonné directement sur les contributions.

Premier axe : les modalités seront significatives quand $|a_{i1}| > \sqrt{2\lambda_1} = 0,84$.
- en négatif : à la limite, `AB2h` ($-0,83$) ;
- en positif : `AB0h` ($1,16$), `P0h`, `P15m` et `B0h` ($1,07$), `A0h` ($0,93$).

Deuxième axe : la valeur limite est maintenant $|a_{i2}| > \sqrt{2\lambda_2} = 0,4$.
- en négatif : `AB30m` ($-0,55$) et `A30m` ($-0,49$) ;
- en positif : `AB3h` ($0,78$) et `ABp4h` ($0,69$).

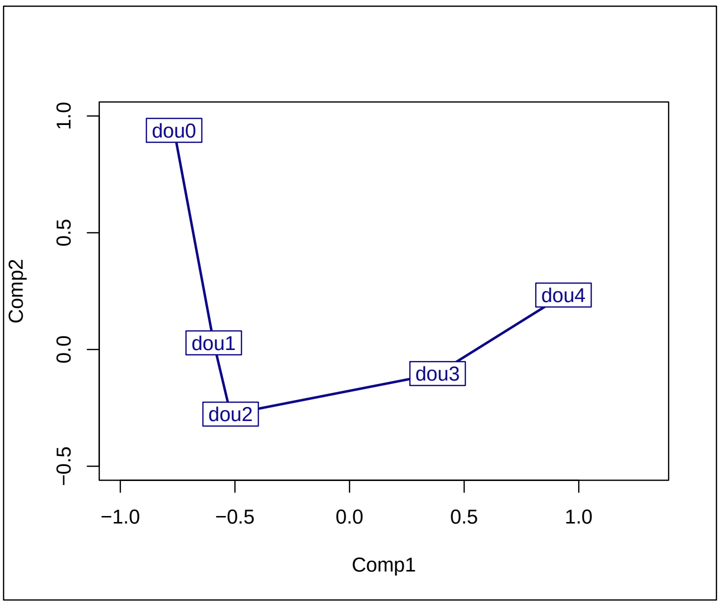
Question 5 *À partir de la question précédente, peut-on interpréter le premier plan principal ? Sur la représentation des individus, tracer une ligne reliant les douleurs dans l'ordre. Que peut on dire ? Tracer des lignes reliant les différents traitements au cours du temps. Que peut-on en déduire sur le l'efficacité de l'association de A et B ?*

La question précédente ne nous aide pas vraiment à conclure : en effet, il y a peu de variables représentées.
- le premier axe décrit surtout la douleur au moment initial.
- le second axe, tout de même, associe `dou0` avec la prise de AB depuis au moins 3 heures. On peut déduire que AB fonctionne mieux, mais c'est faible.

On est dans un cas où l'analyse par axe fonctionne mal, il est donc intéressant de se tourner vers une analyse plus graphique. C'est ce qu'on va faire ci-dessous.

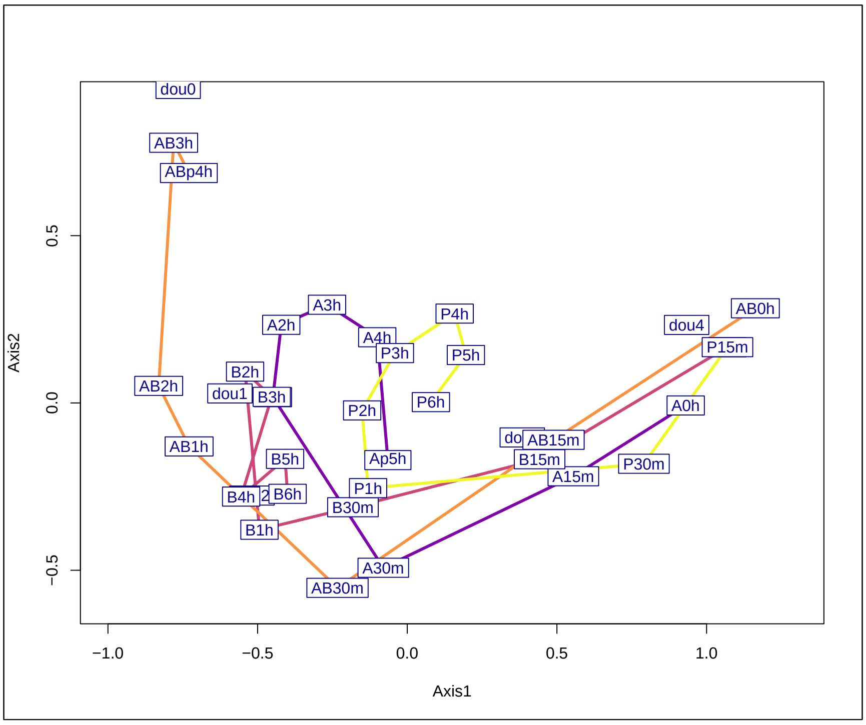
Si on relie les douleurs, on obtient une parabole : horizontalement, les douleurs sont ordonnées le long du premier axe ; cela n'était pas évident a priori, puisque ces niveaux n'étaient pas ordonnés. Verticalement, on a une opposition des extrêmes et de valeurs moyennes. C'est l'*effet Guttman*. Cet effet (courbe un peu parabolique) apparaît quand une des variables a des modalités qui s'ordonnent naturellement, comme la douleur ici. Cela implique souvent qu'on ne peut pas dire grand chose des deux premiers axes (de même que le premier axe est peu intéressant quand il y a un effet de taille).

```
> require(viridis)
> palette(plasma(5))
> plot(coa2$co[,1:2], type='l', xlim=c(-1,1.3), ylim=c(-0.5,1), lwd=2, col=1)
> s.label(coa2$co, add.p=T)
```



Si on relie les médicaments chronologiquement, on obtient maintenant des escargots qui vont vers des douleurs moyennes, sauf pour AB, qui a un comportement très différent et va dans la direction de `dou0`.

```
> plot(coa2$li[1:8,1:2],type='l',xlim=c(-1,1.3),ylim=c(-0.6,0.9), lwd=3, col=2)
> lines(coa2$li[9:17,1:2], lwd=3, col=3)
> lines(coa2$li[18:24,1:2], lwd=3, col=4)
> lines(coa2$li[25:33,1:2], lwd=3, col=5)
> s.label(coa2$co, add.p=T)
> s.label(coa2$li, add.p=T)
```



On voit donc clairement que l'effet de AB est plus stable dans le temps, là où celui des autres antalgiques (et du placebo) semblent s'estomper après quelques heures.

Question 6 *Citez les 6 profils les plus mal représentés par le premier plan principal. Peut-on en déduire quelque chose ?*

Pour trouver ces profils lignes, on s'intéresse aux données de qualité de représentation, qui sont ici les cosinus carrés de l'angle entre chaque profil ligne centré et chacun des axes. La formule classique pour cette valeur est, pour le profil i et l'axe $k = 1, 2$

$$\frac{(a_{ik})^2}{\sum_{\ell=1}^{4} (a_{ik})^2},$$

valeur qui est d'autant meilleure qu'elle est proche de 1.

Ce qui nous intéresse ici est le cosinus carré de l'angle avec le *premier plan principal*. Pour cela, il suffit de lire lire la deuxième colonne de données. On trouve : `P6h` $(3,9\%)$`,` `Ap5h` $(15,9\%)$, `P5h` $(28,6\%)$, `A4h` $(32,2\%)$, `P1h` $(36,7\%)$ et `A1h` $(38,3\%)$ pour les lignes, et rien pour les colonnes.

Par contre, on remarque que ces modalités on des coordonnées proches du centre du premier plan principal (sauf peut-être `A1h`). Il est donc un peu difficile de juger si elles sont vraiment mal représentées.

3 Valeur maximale des valeurs propres en AFC

On sait que dans une AFC, les valeurs propres vérifient toutes $\lambda_k \geq 0$. On veut montrer ici qu'on a aussi $\lambda_k \leq 1$. Pour cela on utilisera les formules de barycentre, qui relient les composantes principales de lignes a_{ik} et celles de colonnes, notées b_{jk} :

$$a_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{m_2} \frac{n_{ij}}{n_{i\cdot}} b_{jk},$$

$$b_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{\cdot j}} a_{ik}.$$

On utilise la notation suivante pour la plus grande coordonnée des modalités d'une variable :

$$\max_{1 \leq i \leq m_1} a_{ik} = \max(a_{1k}, a_{2k}, \ldots, a_{m_1 k})$$
$$\max_{1 \leq j \leq m_2} b_{jk} = \max(b_{1k}, b_{2k}, \ldots, b_{m_2 k})$$

Question 7 *Montrer que, pour toute modalité i et tout axe k,*

$$\sqrt{\lambda_k} a_{ik} \leq \max_{1 \leq j \leq m_2} b_{jk}.$$

Montrer de même que

$$\sqrt{\lambda_k} b_{jk} \leq \max_{1 \leq i \leq m_1} a_{ik}.$$

On part de la première relation barycentrique pour obtenir

$$\sqrt{\lambda_k} a_{ik} = \sum_{j=1}^{m_2} \frac{n_{ij}}{n_{i\cdot}} b_{jk} \leq \sum_{j=1}^{m_2} \frac{n_{ij}}{n_{i\cdot}} \max_{1 \leq j \leq m_2} b_{j'k} = \max_{1 \leq j \leq m_2} b_{j'k}.$$

De même, on peut écrire

$$\sqrt{\lambda_k} b_{jk} = \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{\cdot j}} a_{ik} \leq \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{\cdot j}} \max_{1 \leq i \leq m_1} a_{i'k} = \max_{1 \leq i \leq m_1} a_{ik}.$$

Question 8 *En déduire finalement que, pour tout axe k, $\lambda_k \leq 1$.*

Comme tous les a_{ik} vérifient $\sqrt{\lambda_k} a_{ik} \leq \max_{1 \leq j \leq m_2} b_{jk}$, cette relation est aussi vraie pour leur maximum :

$$\sqrt{\lambda_k} \max_{1 \leq i \leq m_1} a_{ik} \leq \max_{1 \leq j \leq m_2} b_{jk},$$

et donc, en multipliant des deux cotés par $\sqrt{\lambda_k}$, on obtient

$$\lambda_k \max_{1 \leq i \leq m_1} a_{ik} \leq \sqrt{\lambda_k} \max_{1 \leq j \leq m_2} b_{jk},$$

La deuxième inégalité que nous cherchons est obtenue comme la toute première :

$$\sqrt{\lambda_k} b_{jk} \leq \max_{1 \leq i \leq m_1} a_{ik} \text{ pour tout } j \implies \max_{1 \leq j \leq m_2} b_{jk} \leq \max_{1 \leq i \leq m_1} a_{ik}.$$

Finalement, on en déduite que

$$\lambda_k \max_{1 \leq i \leq m_1} a_{ik} \leq \sqrt{\lambda_k} \max_{1 \leq j \leq m_2} b_{jk} \leq \max_{1 \leq i \leq m_1} a_{ik}.$$

On sait que les a_{ik} sont centrés. S'ils ne sont pas tous nuls, alors certains sont négatifs et d'autres positifs, et donc $\max_{1 \leq i \leq m_1} a_{ik} > 0$. On peut donc diviser le premier et le dernier terme de l'inégalité ci-dessus par ce terme et on obtient

$$\lambda_k \leq 1.$$

Dans le cas où tous les a_{ik} sont nuls, alors $\lambda_k = \operatorname{var} \mathbf{a}_k = 0 \leq 1$.