

## ISFA 2° année 2002-2003

*Les questions sont en grande partie indépendantes.*

*Merci d'utiliser l'espace imparti pour vos réponses.*

On considère la matrice de données :

```
> ele
      JC.VGE  FM1  GM  JC.RB  FM2  JMLP
Paris      0.61 0.29 0.10  0.52 0.33 0.15
Lyon       0.59 0.28 0.13  0.51 0.31 0.18
Marseille  0.45 0.26 0.29  0.33 0.33 0.34
Lille      0.50 0.33 0.17  0.39 0.42 0.19
Bordeaux   0.52 0.34 0.14  0.49 0.36 0.15
Toulouse   0.46 0.37 0.17  0.40 0.44 0.16
Strasbourg 0.64 0.30 0.06  0.43 0.34 0.23
Nantes     0.53 0.34 0.13  0.46 0.41 0.13
Nice       0.57 0.24 0.19  0.44 0.27 0.29
Montpellier 0.54 0.32 0.14  0.40 0.36 0.24
Rennes     0.50 0.39 0.11  0.43 0.46 0.11
Toulon     0.55 0.25 0.20  0.41 0.28 0.31
Saint_Etienne 0.52 0.28 0.20  0.42 0.35 0.23
Le_Havre   0.42 0.27 0.31  0.38 0.44 0.18
```

Les 14 lignes (individus) sont 14 grandes villes. Les colonnes (variables) 1-2-3 concernent le premier tour des élections présidentielles de 1981. Les colonnes 4-5-6 concernent le premier tour des élections présidentielles de 1988. Les valeurs sont les pourcentages de voix obtenues calculés sur le total des voix obtenues par les quatre premiers candidats (valeurs arrondies au centième le plus proche). Le code des variables est :

JC.VGE - J. Chirac + V. Giscard d'Estaing (1981)

FM - F. Mitterrand (1 - 1981, 2 - 1988)

GM - G. Marchais (1981)

JC.RB - J. Chirac + R. Barre (1988)

JML - J.M. Le Pen (1988)

Les trois premières colonnes forment la matrice **A** à 14 lignes et 3 colonnes. Les trois dernières colonnes forment la matrice **B** à 14 lignes et 3 colonnes. Le tableau centré (par colonne avec la pondération uniforme) associé à **A** est noté **A<sub>0</sub>**. Le tableau centré (par colonne avec la pondération uniforme) associé à **B** est noté **B<sub>0</sub>**. On considère les matrices juxtaposées par blocs (**O** est la matrice nulle à 14 lignes et 3 colonnes) :

$$\mathbf{J} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{B} \end{bmatrix}$$

On donne :

```
> w = apply(ele,2,sum)/14
> apply(t(ele)-w,1, function(x) sum(x*x)/14)
  JC.VGE  FM1  GM  JC.RB  FM2  JMLP
0.003541 0.001910 0.004406 0.002535 0.003424 0.004509
> w
  JC.VGE  FM1  GM  JC.RB  FM2  JMLP
0.5286 0.3043 0.1671 0.4293 0.3643 0.2064
```

**1.** Quelles sont les moyennes et les variances de **J** ?

2. Donner les relations qui relient les moyennes et les variances de  $\mathbf{S}$  avec les moyennes et les variances de  $\mathbf{J}$ . En déduire les moyennes et les variances de  $\mathbf{S}$  ?
3. Quelles sont les moyennes et les variances de  $\mathbf{D}$  ?
4. Donner les coordonnées cartésiennes des 4 points de la figure 1 (l'échelle est définie par la projection euclidienne adéquate).

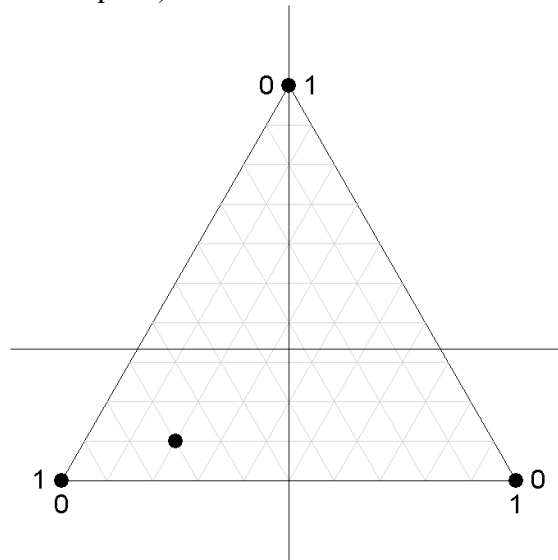


Figure 1

5. Donner une légende pour la figure 2.

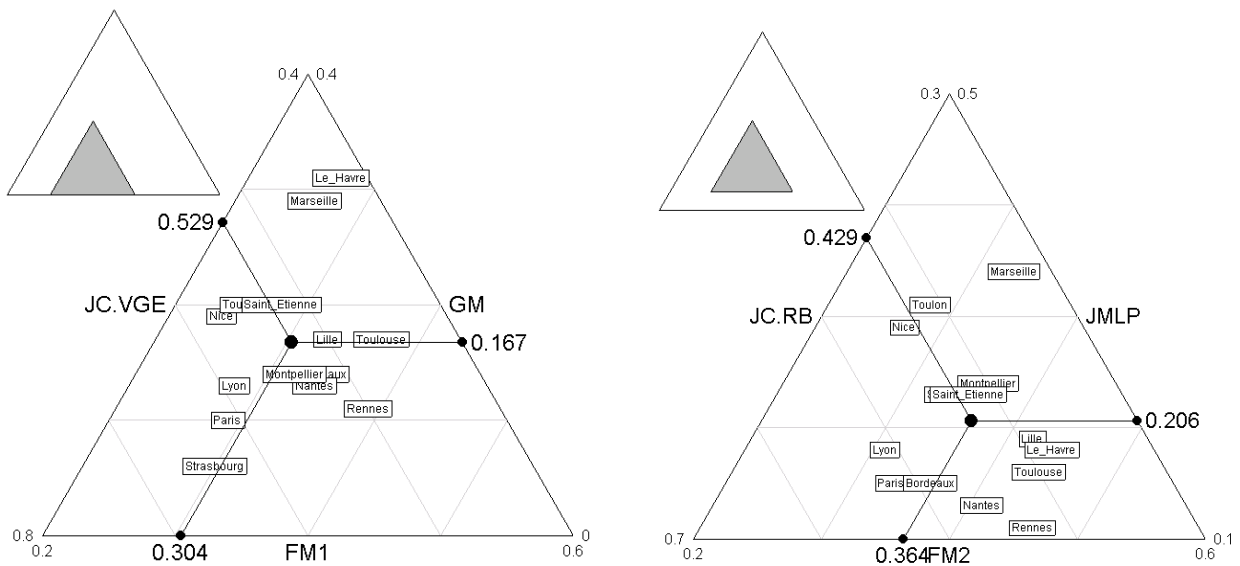


Figure 2

6. On note  $\mathbf{J}_0$ ,  $\mathbf{S}_0$  et  $\mathbf{D}_0$  les tableaux centrés (centrage par colonne, pondération uniforme) respectivement associés à  $\mathbf{J}$ ,  $\mathbf{S}$  et  $\mathbf{D}$ . Donner les rangs des trois matrices  $\mathbf{J}_0$ ,  $\mathbf{S}_0$  et  $\mathbf{D}_0$ .

```

> ana1 = prcomp(ele)
> names(ana1)
[1] "sdev"      "rotation" "x"
> ana2 = princomp(ele)
> names(ana2)
[1] "sdev"      "loadings" "center"    "scale"     "n.obs"
[6] "scores"    "call"
ana3 = dudi.pca(ele, scannf=FALSE, scal=FALSE)
> names(ana3)
[1] "tab" "cw" "lw" "eig" "rank" "nf" "c1" "l1" "co"
[10] "li" "call" "cent" "norm"
> ana1$x[1:3,1:2]
      PC1      PC2
Paris -0.12265 -0.081053
Lyon  -0.07251 -0.088406
Marseille 0.22649 0.004017
> ana2$scores[1:3,1:2]
      Comp.1  Comp.2
Paris  0.12265  0.081053
Lyon   0.07251  0.088406
Marseille -0.22649 -0.004017
> ana3$li[1:3,1:2]
      Axis1  Axis2
Paris -0.12265 -0.081053
Lyon  -0.07251 -0.088406
Marseille 0.22649 0.004017

```

7. Comment appelle-t-on les composantes `ana1$x`, `ana2$scores` ou `ana3$li` ? Les trois fonctions donnent-elle les mêmes résultats ? Donner le nom et la nature d'un autre élément directement comparable d'une procédure à l'autre.

8. Donner une légende pour la figure 3.

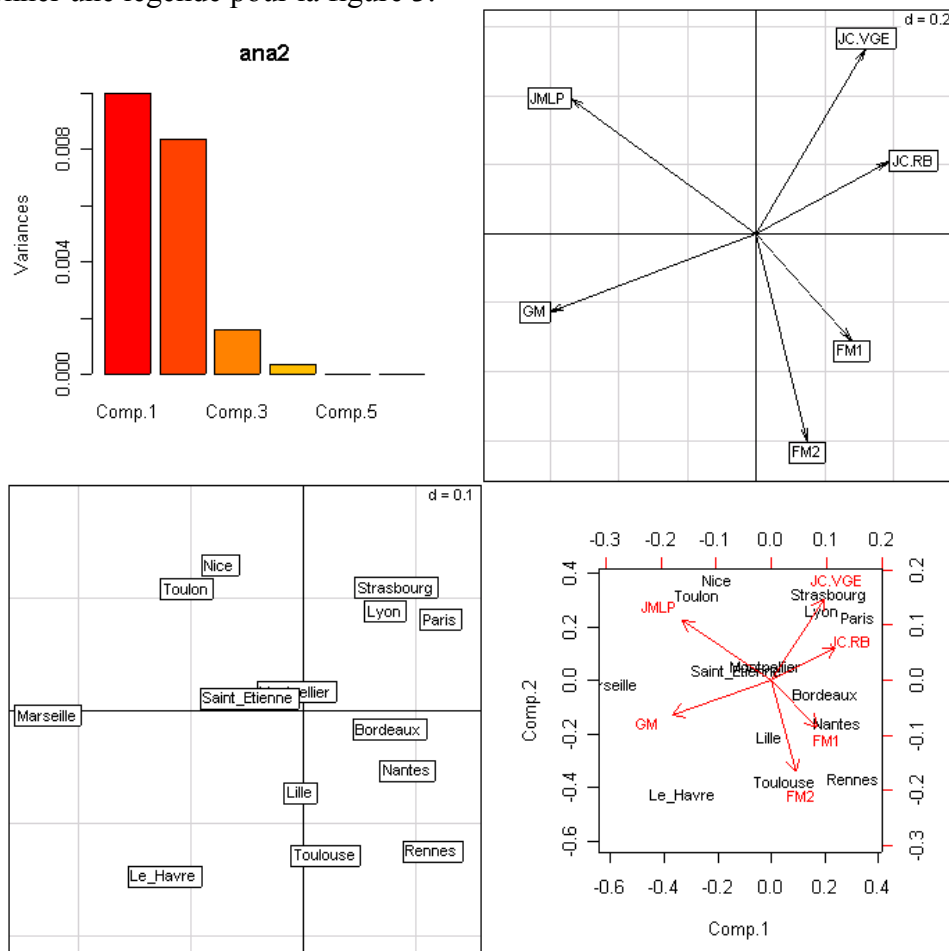


Figure 3

9. Donner les poids des lignes et des colonnes de l'analyse des correspondances du tableau e1e (question 1).

---

10. Une matrice  $\mathbf{X}$  a  $n$  lignes et 2 colonnes. On suppose que les moyennes par colonnes sont nulles (nuage centré). A la ligne  $i$ , on trouve les coordonnées  $(x_i, y_i)$  du point  $M_i$ . On suppose que le vecteur  $\mathbf{u}^t = (a, b)$  unitaire est le premier axe principal du nuage des  $n$  points  $M_i$ . Soit le nouveau nuage de  $n$  points  $P_i$  de coordonnées  $(z_i, t_i)$  définies par :

$$z_i = x_i \cos \alpha + y_i \sin \alpha$$

$$t_i = x_i \sin \alpha - y_i \cos \alpha$$

Donner le premier axe principal du nuage des  $n$  points  $P_i$ .

---

**Nom :**  
**Prénom:**

**ISFA 2 - Analyse des données - 2002/2003**

**1.** Quelles sont les moyennes et les variances de **J** ?

Empty dotted-line box for answer to question 1.

**2.** Donner les relations qui relie les moyennes et variances de **S** ...

Empty dotted-line box for answer to question 2.

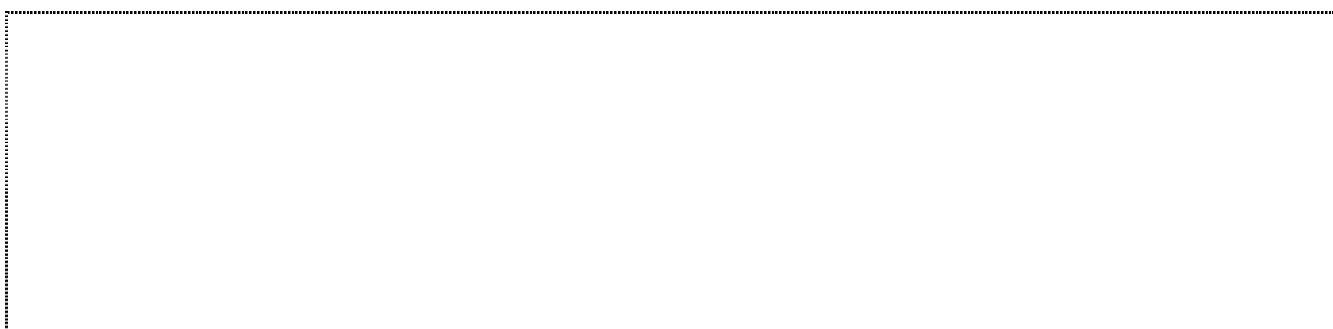
**3.** Quelles sont les moyennes et les variances de **D** ?

Empty dotted-line box for answer to question 3.

**4.** Donner les coordonnées cartésiennes des 4 points de la figure 1



**5.** Donner une légende pour la figure



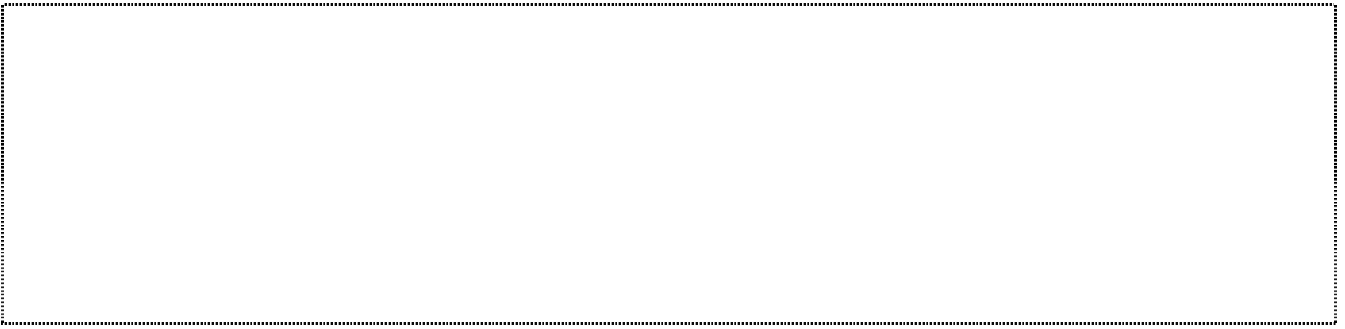
**6.** Donner les rangs des trois matrices  $\mathbf{J}_0$ ,  $\mathbf{S}_0$  et  $\mathbf{D}_0$ .



**7.** Comment appelle-t-on les composantes  $\text{ana1}\$x$ ,  $\text{ana2}\$scores$  ou  $\text{ana3}\$li$  ?



**8.** Donner une légende pour la figure 3.

A large, empty rectangular box with a dotted border, intended for the student to write a legend for Figure 3.

**9.** Donner les poids des lignes et des colonnes de l'analyse des correspondances ...

A large, empty rectangular box with a dotted border, intended for the student to provide the weights for the rows and columns of the correspondence analysis.

**10.** Donner le premier axe principal du nuage des n points

A large, empty rectangular box with a dotted border, intended for the student to provide the first principal axis of the point cloud.

## ISFA 2° année 2002-2003 - Solution

### 1. Quelles sont les moyennes et les variances de **J** ?

apply calcule une fonction par ligne ou colonne, sum calcule la somme des composantes d'un vecteur, t transpose donc les moyennes et les variances de **J** sont les valeurs affichées.

Moyennes : 0.5286 0.3043 0.1671 0.4293 0.3643 0.2064  
 Variances : 0.003541 0.001910 0.004406 0.002535 0.003424 0.004509

### 2. Donner les relations qui relie les moyennes et variances de **S** et les moyennes et variances de **J**. En déduire les moyennes et les variances de **S** ?

Les moyennes des colonnes de **S** sont les demi-sommes des moyennes de **J**. Donc :

$$\begin{aligned} m1 &= (0.5286 + 0.4293)/2 = 0.4789 \\ m2 &= (0.3043 + 0.3643)/2 = 0.3343 \\ m3 &= (0.1671 + 0.2064)/2 = 0.1868 \end{aligned}$$

Les variances sont les sommes des variances inter et intra. Noter que :

$$(x - (x+y)/2)^2 + (y - (x+y)/2)^2 = (x - y)^2/4$$

Donc par exemple :

$$\begin{aligned} v1(S) &= \text{intra} + \text{inter} = (v1(J) + v4(J))/2 + (m1(J) - m4(J))^2/4 \\ v1 &= (0.003541 + 0.002535)/2 + (0.5286 - 0.4293)^2/4 = 0.005502 \\ v2 &= (0.001910 + 0.003424)/2 + (0.3043 - 0.3643)^2/4 = 0.003567 \\ v3 &= (0.004406 + 0.004509)/2 + (0.1671 - 0.2064)^2/4 = 0.004844 \end{aligned}$$

### 3. Quelles sont les moyennes et les variances de **D** ?

Les moyennes de **D** sont simplement la moitié des moyennes de **J**, donc :

Moyennes : 0.26429 0.15214 0.08357 0.21464 0.18214 0.10321

Les variances sont les sommes des variances inter et intra, donc :

$$\begin{aligned} v1(D) &= \text{intra} + \text{inter} = v1(J)/2 + (m1(J))^2/4 \\ v1 &= 0.003541/2 + 0.5286^2/4 = 0.07162 \\ v2 &= 0.001910/2 + 0.3043^2/4 = 0.02410 \\ v3 &= 0.004406/2 + 0.1671^2/4 = 0.00918 \\ v4 &= 0.002535/2 + 0.4293^2/4 = 0.04734 \\ v5 &= 0.003424/2 + 0.3643^2/4 = 0.03489 \\ v6 &= 0.004509/2 + 0.2064^2/4 = 0.01290 \end{aligned}$$

### 4. Donner les coordonnées cartésiennes des 4 points de la figure 1 (l'échelle est définie par la projection euclidienne adéquate).

Un point de coordonnée  $(x,y,z)$  est placé sur une représentation triangulaire par les coordonnées cartésiennes  $(y-x)/\sqrt{2}$  et  $(2*z-y-x)/\sqrt{6}$ . Le point  $(1,0,0)$ , en bas et gauche a pour coordonnée  $(-0.7071, -0.4082)$ . Le point  $(0,1,0)$ , en bas à droite, a pour coordonnée  $(0.7071, -0.4082)$ . Le point  $(0,0,1)$ , en haut, a pour coordonnée  $(0, 0.8165)$ . Le point  $(0.7,0.2,0.1)$ , dans le triangle, a pour coordonnée  $(-0.3536, -0.2858)$ .

### 5. Donner une légende pour la figure 2.

Représentations triangulaires des résultats électoraux dans 14 grandes villes. A gauche, élection de 1981 et utilisation des trois catégories droite / socialiste / communiste, à droite élection de 1988 et utilisation des trois catégories extrême droite / droite / socialiste. Représentation des points moyens des nuages. En haut des figures, position de la partie utilisée du triangle dans le triangle total défini par la base canonique de  $R^3$ . Les variabilités sont comparables. Permanence d'une position centrale pour Saint-Étienne et Montpellier, d'une position à gauche non communiste pour Rennes et Toulouse, d'une position à droite pour Lyon et Paris. La typologie est profondément modifiée par le changement du troisième candidat principal.

### 6. On note $J_0$ , $S_0$ et $D_0$ les tableaux centrés (centrage par colonne, pondération uniforme) respectivement associés à **J**, **S** et **D**. Donner les rangs des trois matrices $J_0$ , $S_0$ et $D_0$ .



La somme par ligne dans  $\mathbf{S}$  vaut 1 donc, après centrage la somme par ligne vaut 0.  $\mathbf{S}_0$  est de rang 2 (pas moins, car il faudrait que les 24 points soient sur une droite). La somme par ligne dans  $\mathbf{J}$  vaut 1 pour les 3 premières colonnes et 1 pour les trois dernières. Après centrage on a deux sommes de colonnes nulles et un rang au plus égal à 4. Les deux sous-nuages étant de rang 2, le rang de  $\mathbf{J}_0$  est 4. Même argument et même résultat pour  $\mathbf{J}_0$ .

**7.** Comment appelle-t-on les composantes `ana1$x`, `ana2$scores` ou `ana3$li` ? Les trois fonctions donnent-elle les mêmes résultats ? Donner le nom et la nature d'un autre élément directement comparable d'une procédure à l'autre.

Les composantes `ana1$x`, `ana2$scores` ou `ana3$li` contiennent les coordonnées factorielles ou scores des individus en ACP. Ce sont les coordonnées des projections sur les axes principaux. Les résultats sont identiques parce que le signe d'une coordonnée est arbitraire. Un autre élément comparable est formé des axes principaux en colonnes dans `ana1$rotation`, `ana2$loadings` ou `ana3$c1`. `ana1$sdev` ou `ana2$sdev` se retrouve au carré dans `ana3$eig`.

**8.** Donner une légende pour la figure 3.

Dépouillement graphique d'une analyse en composantes principales centrée. En haut, à gauche, représentation de la répartition de la variance entre les axes principaux. L'essentiel de la structure est de dimension 2. En haut, à droite, projection de la base canonique sur les axes principaux (plan 1-2). En bas, à gauche, carte factorielle (1-2) du nuage des 14 lignes (projection du nuage sur les axes principaux). En bas à droite, biplot ou représentation simultanée des deux précédents. Les contraintes propres aux données associe les deux équipes de droite ( $F1 > 0$ ,  $F2 > 0$ ), les deux présentations du candidat de gauche ( $F1 > 0$ ,  $F2 < 0$ ) et les propriétés numériques des données rejettent les deux troisièmes ( $F1 < 0$ ), sans les associer. Marseille est passé du vote Marchais au vote LePen et se tient sur l'axe.

**9.** Donner les poids des lignes et des colonnes de l'analyse des correspondances du tableau ele (question 1).

Chacune des lignes a une somme marginale de 2 et la somme totale vaut 2 fois le nombre de lignes. Les poids des lignes sont identiques et valent  $1/14 = 0.07143$ . La somme pour une colonne vaut 14 fois la moyenne (par définition de la moyenne) et le poids de la colonne, après division par la somme totale vaut la moitié de la moyenne.

D'où les poids :

JC.VGE	FM1	GM	JC.RB	FM2	JMLP
0.26429	0.15214	0.08357	0.21464	0.18214	0.10321

**10.** Donner le premier axe principal du nuage des n points  $P_i$ .

L'axe principal est le premier vecteur propre de  $\mathbf{X}'\mathbf{X}$  et  $\mathbf{X}'\mathbf{X}\mathbf{u} = \lambda\mathbf{u}$ . La matrice  $\mathbf{Y}$  qui contient les coordonnées des nouveaux points est  $\mathbf{Y} = \mathbf{X}\mathbf{A}$  avec :

$$\mathbf{A} = \begin{bmatrix} \cos \alpha & \sin \alpha \\ \sin \alpha & -\cos \alpha \end{bmatrix}$$

Le premier axe du nouveau nuage est propre de  $\mathbf{Y}'\mathbf{Y} = \mathbf{A}'\mathbf{X}\mathbf{X}\mathbf{A}$ .

$$\mathbf{A}'\mathbf{X}\mathbf{X}\mathbf{A}\mathbf{v} = \mu\mathbf{v} \Rightarrow \mathbf{A}\mathbf{A}'\mathbf{X}\mathbf{X}\mathbf{A}\mathbf{v} = \mu\mathbf{A}\mathbf{v} \Rightarrow \mathbf{X}\mathbf{X}\mathbf{A}\mathbf{v} = \mu\mathbf{A}\mathbf{v} \Rightarrow \mathbf{A}\mathbf{v} = \mathbf{u} \Rightarrow \mathbf{v} = \mathbf{A}'\mathbf{u} \text{ et } \mu = \lambda$$

L'axe recherché  $\mathbf{v}' = (c, d)$  s'écrit :

$$c = a \cos \alpha + b \sin \alpha$$

$$d = a \sin \alpha - b \cos \alpha$$

L'axe principal est conservé par rotation du nuage.