ISFA 2° année 2000-2001

Problème d'examen (Représentation triangulaire, ACP et élections)

D. Chessel

Les exercices (17-20) sont indépendants du problème (1-16).

1. Questions

On considère la matrice **A** à $n = 14$ lignes et $p = 6$ colonnes :

```
> A
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,] 0.61 0.29 0.10 0.52 0.33 0.15  Paris
 [2,] 0.59 0.28 0.13 0.51 0.31 0.18  Lyon
 [3,] 0.45 0.26 0.29 0.33 0.33 0.34  Marseille
 [4,] 0.50 0.33 0.17 0.39 0.42 0.19  Lille
 [5,] 0.52 0.34 0.14 0.49 0.36 0.15  Bordeaux
 [6,] 0.46 0.37 0.17 0.40 0.44 0.16  Toulouse
 [7,] 0.64 0.30 0.06 0.43 0.34 0.23  Strasbourg
 [8,] 0.53 0.34 0.13 0.46 0.41 0.13  Nantes
 [9,] 0.57 0.24 0.19 0.44 0.27 0.29  Nice
[10,] 0.54 0.32 0.14 0.40 0.36 0.24  Montpellier
[11,] 0.50 0.39 0.11 0.43 0.46 0.11  Rennes
[12,] 0.55 0.25 0.20 0.41 0.28 0.31  Toulon
[13,] 0.52 0.28 0.20 0.42 0.35 0.23  Saint-Étienne
[14,] 0.42 0.27 0.31 0.38 0.44 0.18  Le Havre
```

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 :

1 - J. Chirac + V. Giscard d'Estaing (1981) 2 - F. Mitterand (1981) 3 - G. Marchais (1981)
4 - J. Chirac + R. Barre (1988) 5 - F. Mitterand (1988) 6 - J.M. Le Pen (1988)

Les trois premières colonnes de **A** forment la matrice **B** à 14 lignes et 3 colonnes. Les trois dernières colonnes de **A** forment la matrice **C** à 14 lignes et 3 colonnes. Les moyennes des variables de **A** sont :

```
0.5286 0.3043 0.1671 0.4293 0.3643 xxxxxx
```

1.1. Donner la valeur manquante.

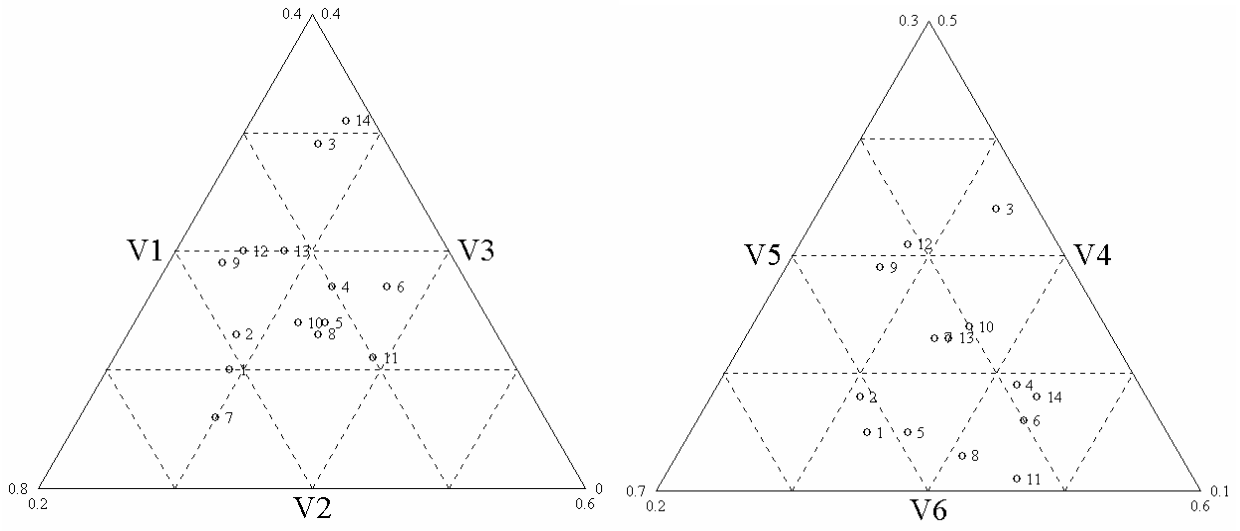


Figure 1

1.2. Placer sur la figure 1 le centre de gravité du nuage de points en indiquant votre méthode.

1.3. Donner la base orthonormale de $\mathbb{R}^3$ (pour la métrique canonique) qui permet de construire les représentations triangulaires.

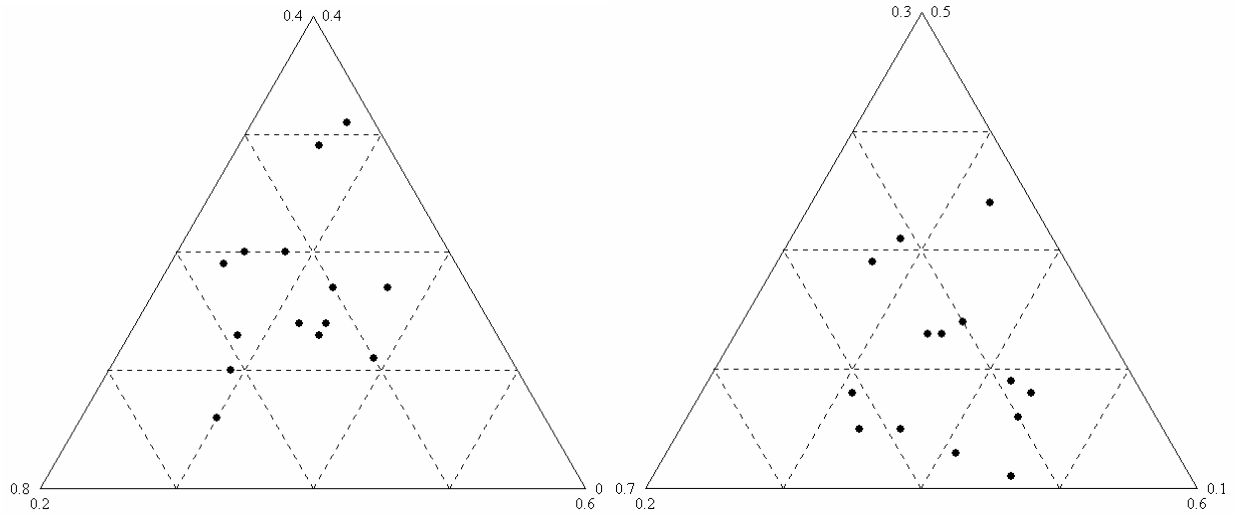


Figure 2

1.4. Compléter à votre convenance la figure 2 et rédiger un commentaire pour le résultat.

1.5. Expliquer en quoi on peut considérer que les deux parties de la figure 1 sont des cartes factorielles d'ACP centrée.

Pour coordonner ces deux représentations on fait les calculs qui suivent. Les variances des variables de **A** sont :

```
0.003541 0.001910 0.004406 0.002535 0.003424 0.004509
```

Le tableau centré $\mathbf{A}_0$ associé à **A** est :

```
[,1]    [,2]    [,3]    [,4]    [,5]    [,6]
```

```
 [1,]  0.0814 -0.0143 -0.0671  0.0907 -0.0343 -0.0564
 [2,]  0.0614 -0.0243 -0.0371  0.0807 -0.0543 -0.0264
 [3,] -0.0786 -0.0443  0.1229 -0.0993 -0.0343  0.1336
 [4,] -0.0286  0.0257  0.0029 -0.0393  0.0557 -0.0164
 [5,] -0.0086  0.0357 -0.0271  0.0607 -0.0043 -0.0564
 [6,] -0.0686  0.0657  0.0029 -0.0293  0.0757 -0.0464
 [7,]  0.1114 -0.0043 -0.1071  0.0007 -0.0243  0.0236
 [8,]  0.0014  0.0357 -0.0371  0.0307  0.0457 -0.0764
 [9,]  0.0414 -0.0643  0.0229  0.0107 -0.0943  0.0836
[10,]  0.0114  0.0157 -0.0271 -0.0293 -0.0043  0.0336
[11,] -0.0286  0.0857 -0.0571  0.0007  0.0957 -0.0964
[12,]  0.0214 -0.0543  0.0329 -0.0193 -0.0843  0.1036
[13,] -0.0086 -0.0243  0.0329 -0.0093 -0.0143  0.0236
[14,] -0.1086 -0.0343  0.1429 -0.0493  0.0757 -0.0264
```

La matrice $\mathbf{W} = \frac{1}{n}\mathbf{A}_0^t\mathbf{A}_0$ vaut :

```
        [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,]  xxxxxxx -0.00052 -0.00302  0.00200 -0.00210  0.00010
[2,] -0.00052  xxxxxxx -0.00139  0.00024  0.00199 -0.00223
[3,] -0.00302 -0.00139  xxxxxxx -0.00224  0.00011  0.00213
[4,]  0.00200  0.00024 -0.00224  xxxxxxx -0.00073 -0.00181
[5,] -0.00210  0.00199  0.00011 -0.00073  xxxxxxx -0.00270
[6,]  0.00010 -0.00223  0.00213 -0.00181 -0.00270  xxxxxxx
```

1.6. Donner les éléments de la diagonale principale.

1.7. Donner le rang et une base orthonormée du noyau de **W**.

Les deux premières valeurs propres de **W** sont :

```
1.001e-02  8.354e-03
```

Les deux vecteurs propres associés en colonnes forment la matrice **U**. **U** vaut :

```
        [,1]    [,2]
[1,] -0.3191 -0.5361
[2,] -0.2782  0.3095
[3,]  xxxxxx  0.2266
[4,] -0.3888 -0.2118
[5,] -0.1500  0.6050
[6,]  0.5389  xxxxxx
```

1.8. Donner les valeurs manquantes.

1.9. Vrai ou faux ? La matrice $\mathbf{U}\mathbf{U}^t$ est la matrice d'un projecteur orthogonal.

En utilisant les composantes des vecteurs propres, on obtient la figure 3.

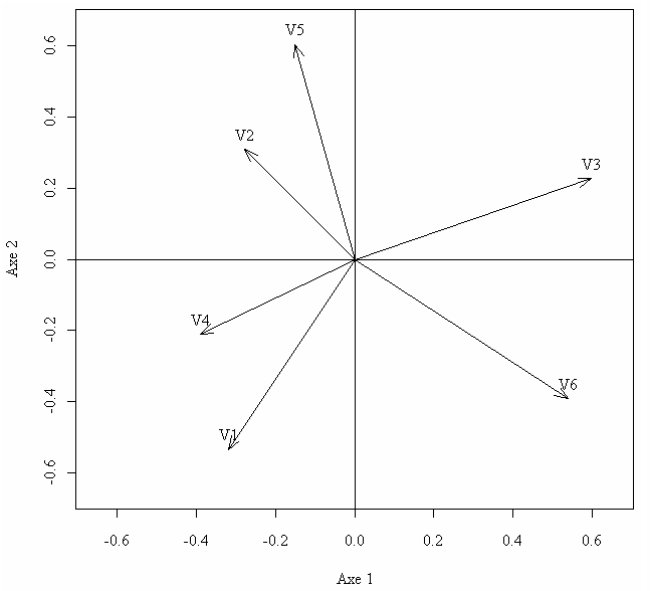


Figure 3

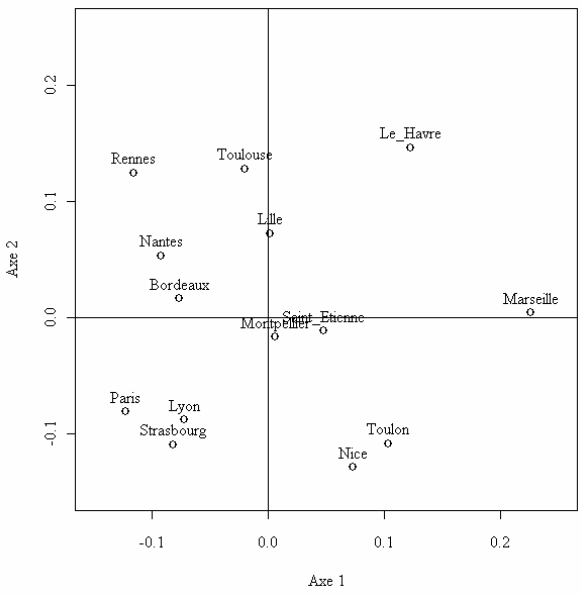


Figure 4

1.10. La figure 3 est-elle une projection euclidienne ?

Les coordonnées factorielles des lignes de l'ACP centrée de **A** sont :

```
        [,1]   [,2]
 [1,] -0.123 -0.081
 [2,] -0.073 -0.088
 [3,]  0.226  0.004
 [4,]  0.002  0.072
 [5,] -0.077  0.016
 [6,] -0.020  0.128
 [7,] -0.082 -0.109
 [8,] -0.093  0.053
 [9,]  0.073 -0.129
[10,]  0.006 -0.017
[11,] -0.115  0.125
[12,]  0.104 -0.108
[13,]  0.048 -0.011
[14,]  0.123  0.147
```

1.11. Comment ces valeurs sont-elles calculées ?

En utilisant les coordonnées factorielles on obtient la figure 4.

1.12. Est-il légitime de superposer les figures 3 et 4 ?

La matrice $\mathbf{A}_0$ est décomposée en deux blocs ($\mathbf{B}_0$ et $\mathbf{C}_0$) et on construit à l'aide de deux matrices nulles une nouvelle matrice $\mathbf{Z} = \begin{bmatrix} \mathbf{B}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_0 \end{bmatrix}$.

```
        [,1]    [,2]    [,3]    [,4]    [,5]    [,6]
 [1,]  0.0814 -0.0143 -0.0671  0.0000  0.0000  0.0000   Paris_1
 [2,]  0.0614 -0.0243 -0.0371  0.0000  0.0000  0.0000   Lyon_1
 [3,] -0.0786 -0.0443  0.1229  0.0000  0.0000  0.0000   Marseille_1
...
[12,]  0.0214 -0.0543  0.0329  0.0000  0.0000  0.0000   Toulon_1
[13,] -0.0086 -0.0243  0.0329  0.0000  0.0000  0.0000   Saint_Etienne_1
[14,] -0.1086 -0.0343  0.1429  0.0000  0.0000  0.0000   Le_Havre_1
```

```
[15,]  0.0000  0.0000  0.0000  0.0907 -0.0343 -0.0564   Paris_2
[16,]  0.0000  0.0000  0.0000  0.0807 -0.0543 -0.0264   Lyon_2
[17,]  0.0000  0.0000  0.0000 -0.0993 -0.0343  0.1336   Marseille_2

...
[26,]  0.0000  0.0000  0.0000 -0.0193 -0.0843  0.1036   Toulon_2
[27,]  0.0000  0.0000  0.0000 -0.0093 -0.0143  0.0236   Saint_Etienne_2
[28,]  0.0000  0.0000  0.0000 -0.0493  0.0757 -0.0264   Le_Havre_2
```

On calcule **ZU** (projection en individu supplémentaire) qui permet de tracer la figure 5.

1.13. Caractériser la relation qui existe entre les figures 4 et 5.

1.14. Vrai ou faux ? Les tableaux **A** et **Z** ont même matrice de covariance.

1.15. Vrai ou faux ? Le rang de la matrice **Z** vaut 4.

1.16. Oui ou Non ? Les figures obtenues représentent 90% de la variabilité contenue dans le tableau de données.

Exercices

1.17. On considère les 8 points régulièrement répartis sur le cercle unité représentés sur la figure 6. On utilise la métrique canonique et la pondération uniforme. Quelle est l'inertie de ce nuage de points (autour de l'origine) ?

1.18. Vrai ou faux ? L'inertie de ce nuage de 8 points projeté sur une droite passant par l'origine ne dépend pas de la droite.

1.19. Vrai ou faux ? L'inertie projetée sur le premier axe principal d'une ACP normée sur p variables vaut au moins 1.

1.20. Vrai ou faux ? Dans une matrice de corrélation théorique sur p variables dans laquelle tous les coefficients non diagonaux sont égaux à α l'inertie projetée sur une des composantes principales vaut $1+(p-1)\alpha$.

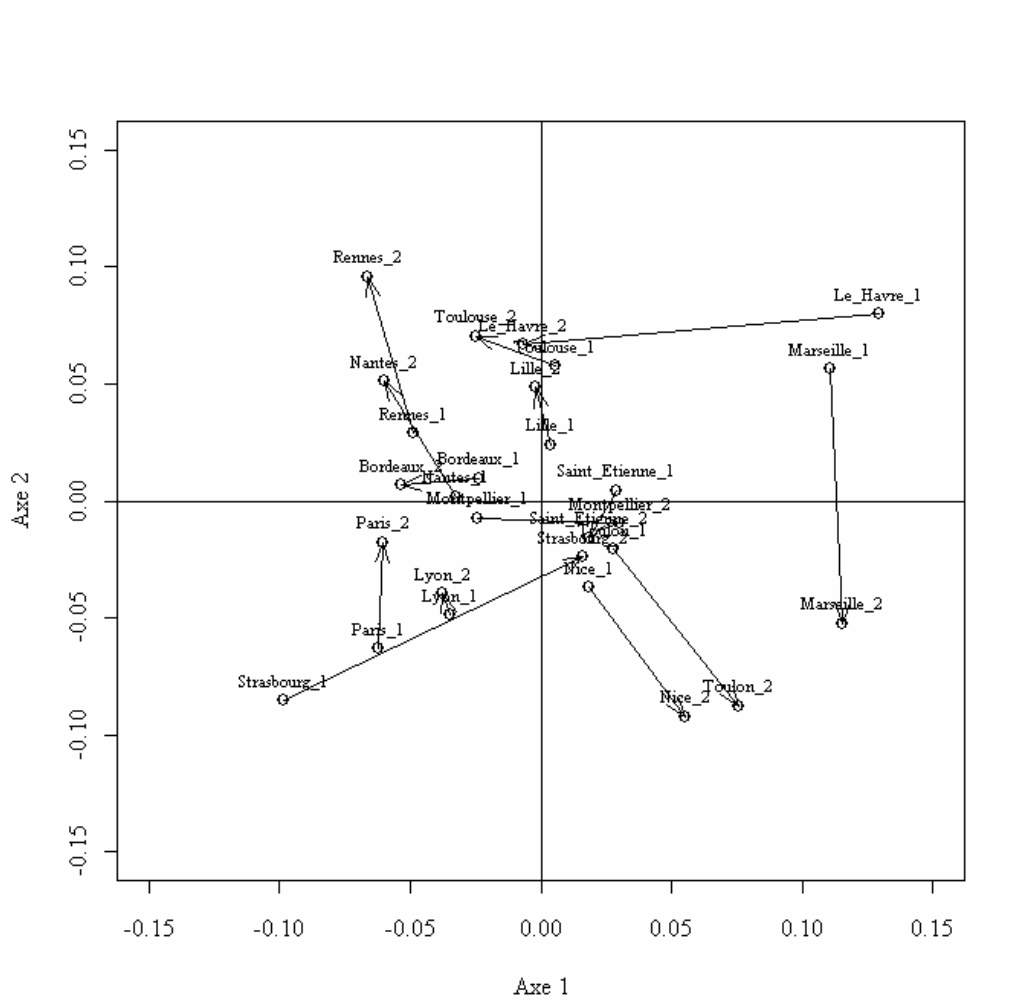


Figure 5

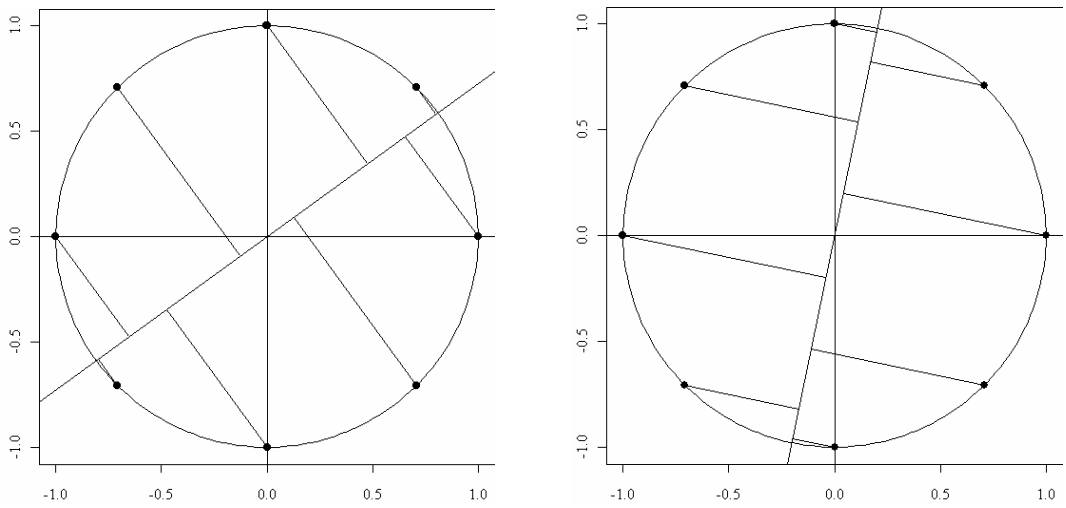
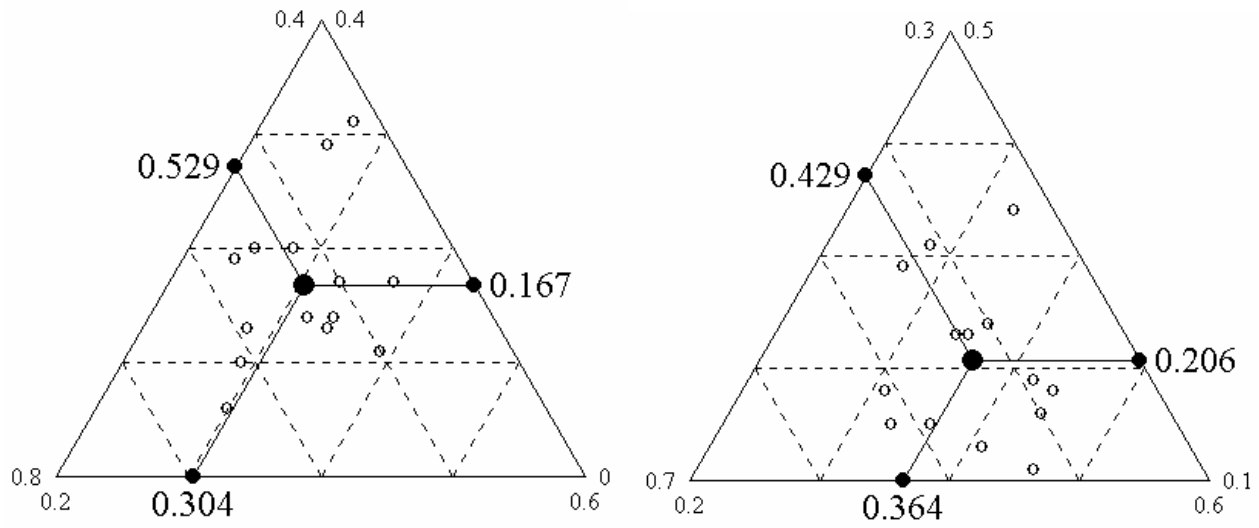


Figure 6

2. Solutions

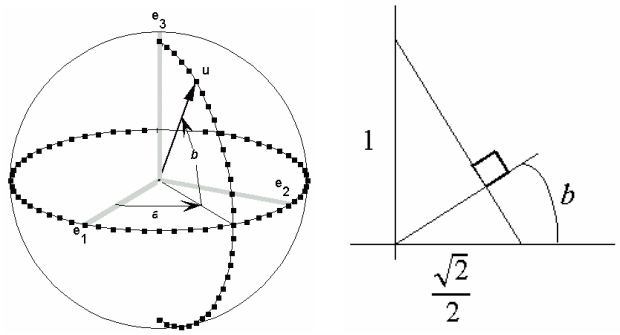
1. C'est la moyenne de la dernière variable, soit 0.2064

2. Le centre de gravité du nuage de points a pour coordonnées dans la base canonique la moyenne des variables. Ces coordonnées forment une distribution de fréquences (somme =1). Le point moyen est positionné comme une observation.



3. La base orthonormale de $\mathbb{R}^3$ (pour la métrique canonique) qui permet de tracer les représentations triangulaires est du type (Chapitre 1 du cours) :

$$\mathbf{H} = \begin{matrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{matrix} \begin{bmatrix} \begin{bmatrix} \\ \end{bmatrix} & \begin{bmatrix} \\ \end{bmatrix} & \begin{bmatrix} \\ \end{bmatrix} \end{bmatrix} \begin{matrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{matrix} = \begin{bmatrix} \cos a \cos b & -\sin a & -\cos a \sin b \\ \sin a \cos b & \cos a & -\sin a \sin b \\ \sin b & 0 & \cos b \end{bmatrix}$$



On cherche **u** pour qu'il soit perpendiculaire au plan $x+y+z=1$, soit $a=\frac{\pi}{4}$ et b donné par :

$$\cos b = \frac{\sqrt{2}}{\sqrt{3}} \quad \sin b = \frac{1}{\sqrt{3}}$$

$$\mathbf{H} = \begin{bmatrix} 1/\sqrt{3} & -1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & 2/\sqrt{6} \end{bmatrix}$$

4.

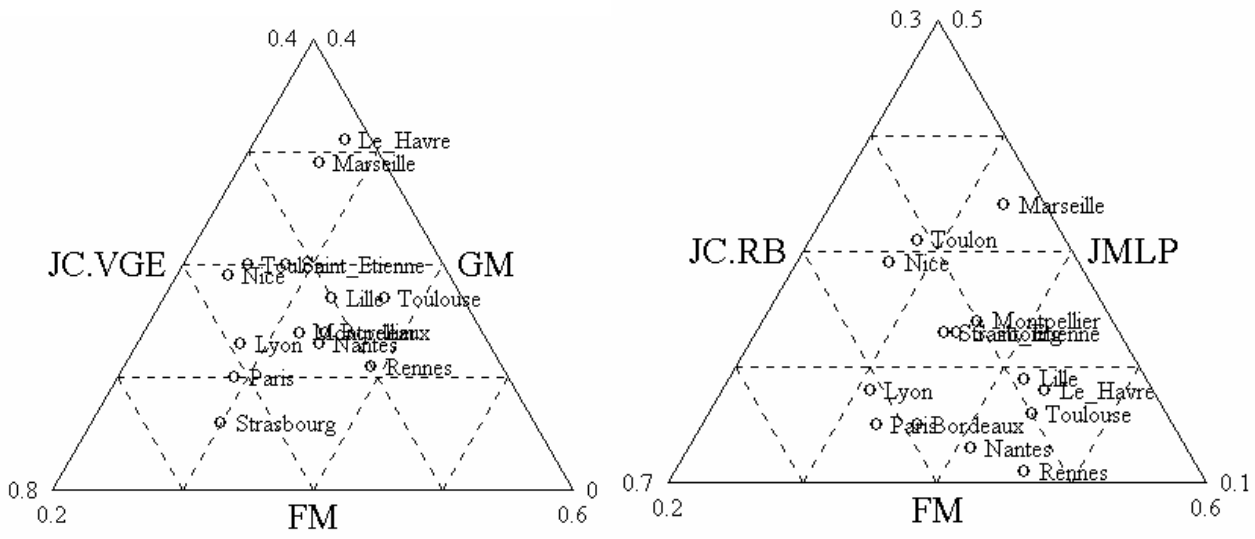
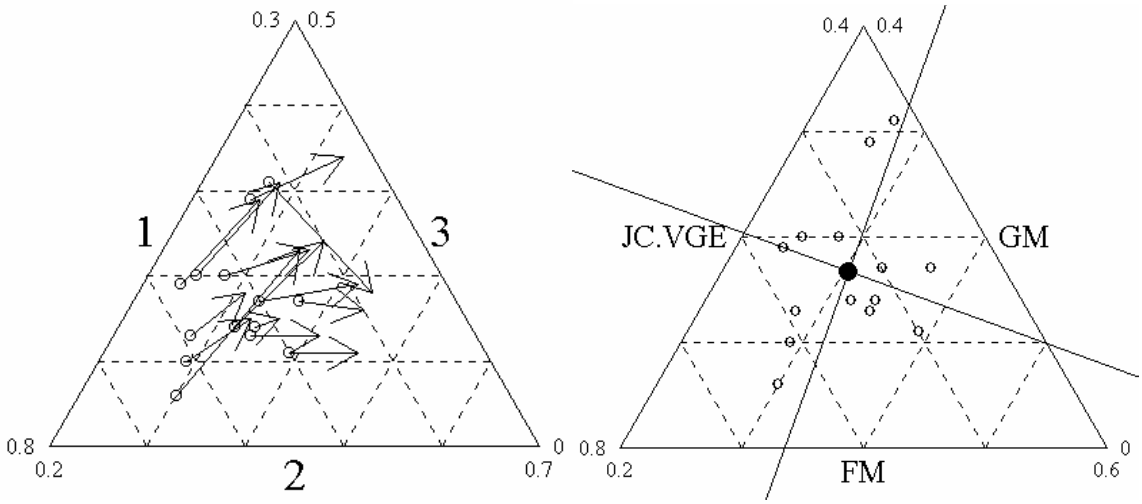


Figure 2

Figure 1 : double représentation triangulaire de résultats électoraux simplifiés. Les échelles ne sont pas directement comparables mais la figure conduit à comparer deux typologies de villes suivant leurs résultats électoraux. Lyon et Paris restent des villes de droite, Rennes, Nantes, Toulouse restent des villes de gauche, Montpellier reste proche de la moyenne. Une ville a largement modifié sa position relative dans l'ensemble (Le Havre). Cette figure peut conduire à des conclusions erronées car la troisième variable a une signification instable. Pour information : à gauche, la double représentation superposée qui donne une information très différente et, à droite, la représentation des axes dans le plan (voir 5) qui est une question difficile à laquelle vous avez échappé.



On peut aussi discuter des bornes des deux triangles, des notions de permanence de structure, de glissement du nuage, de déplacements des points, … Le tout est de donner un sens concret à la figure.

5. Les deux parties de la figure 1 sont des cartes factorielles d'ACP centrée. En effet, le nuage des 14 points de $\mathbb{R}^3$ est dans le plan affine $x+y+z=1$. Le centre de gravité est dans ce plan et le nuage centré est dans le plan $x+y+z=0$. Les axes principaux sont donc dans ce plan et le plan du triangle est le plan des axes principaux. A une rotation près, la représentation triangulaire est exactement celle de l'analyse en composantes principales centrée.

6. Les éléments de la diagonale principale sont tout simplement les variances, soient 0.0035, 0.0019, …, 0.0045.

7. La matrice s'écrit :

$$\mathbf{W} = \begin{bmatrix} \frac{\mathbf{1}}{n}\mathbf{A}_0^t\mathbf{A}_0 & \frac{\mathbf{1}}{n}\mathbf{A}_0^t\mathbf{B}_0 \\ \frac{\mathbf{1}}{n}\mathbf{B}_0^t\mathbf{A}_0 & \frac{\mathbf{1}}{n}\mathbf{B}_0^t\mathbf{B}_0 \end{bmatrix}$$

Les vecteurs :

$$\mathbf{a}^t = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{b}^t = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

sont dans le noyau et en forme une base orthogonale. Les autres vecteurs propres sont orthogonaux et ne peuvent annuler la matrice ($\mathbf{A}_0$ et $\mathbf{B}_0$ sont de rang 2) . Le rang de $\mathbf{W}$ est 4.

8. Par orthogonalité aux vecteurs du noyau les composantes sont de somme nulle par blocs de 3.

-0.3191 - 0.2782 + x = 0 donne x = 0.5972

-0.2118 + 0.6050 + x = 0 donne x = -0.3952

9. Vrai ou faux ? La matrice est la matrice d'un projecteur orthogonal. **C'est vrai** :

$$\mathbf{U}\mathbf{U}^t\mathbf{x} = \mathbf{U}\begin{bmatrix} \langle \mathbf{u}_1 | \mathbf{x} \rangle \\ \langle \mathbf{u}_2 | \mathbf{x} \rangle \end{bmatrix} = \langle \mathbf{u}_1 | \mathbf{x} \rangle \mathbf{u}_1 + \langle \mathbf{u}_2 | \mathbf{x} \rangle \mathbf{u}_2 = \mathrm{P}^{\perp}_{sev(\mathbf{u}_1,\mathbf{u}_2)}(\mathbf{x})$$

10. La figure 3 est-elle une projection euclidienne ? Oui car

$$\mathbf{U} = \begin{matrix} \mathbf{u}_1 & \mathbf{u}_2 \end{matrix} \begin{bmatrix} \begin{bmatrix} \\ \end{bmatrix} & \begin{bmatrix} \\ \end{bmatrix} \\ \vdots & \\ \end{bmatrix} \begin{matrix} \mathbf{e}_1 \\ \\ \mathbf{e}_6 \end{matrix} \Rightarrow \mathbf{U}_{ij} = \langle \mathbf{e}_i | \mathbf{u}_j \rangle$$

La base canonique et la famille $\mathbf{u}_1, \mathbf{u}_2$ sont orthonormales. En utilisant les composantes des **u** dans la base canonique , on projette la base canonique sur le plan $\mathbf{u}_1, \mathbf{u}_2$ (principe du biplot).

11. Comment ces valeurs sont-elles calculées ?

$$\mathbf{L} = \mathbf{X}_0\mathbf{U}$$

par exemple 0.0814*-0.3191 + … + -0.056*-0.3952. Illustrer par un exemple numérique.

12. Est-il légitime de superposer les figures 3 et 4 ?

Oui, car ce sont des projections sur un plan de $\mathbb{R}^6$, d'une part d'un nuage de points de cet espace, d'autre part de la base canonique du même espace.

13. Caractériser la relation qui existe entre les figures 4 et 5.

$$\mathbf{X}_0\mathbf{U} = \begin{bmatrix} \mathbf{B}_0 | \mathbf{C}_0 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} = \mathbf{B}_0\mathbf{U}_1 + \mathbf{C}_0\mathbf{U}_2 = 2\frac{\mathbf{B}_0\mathbf{U}_1 + \mathbf{C}_0\mathbf{U}_2}{2}$$

Donc la figure 4, à la constante 2 près, représente les milieux des flèches de la figure 5.

14. Vrai ou faux ? Les tableaux A et Z ont même matrice de covariance. NON

C'est manifestement faux, car :

$$\begin{bmatrix} \frac{1}{n}\mathbf{A}_0^t\mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\mathbf{B}_0^t\mathbf{B}_0 \end{bmatrix} \neq \begin{bmatrix} \frac{1}{n}\mathbf{A}_0^t\mathbf{A}_0 & \frac{1}{n}\mathbf{A}_0^t\mathbf{B}_0 \\ \frac{1}{n}\mathbf{B}_0^t\mathbf{A}_0 & \frac{1}{n}\mathbf{B}_0^t\mathbf{B}_0 \end{bmatrix}$$

15. Vrai ou faux ? Le rang de la matrice **Z** est 4.

OUI, les vecteurs $\begin{pmatrix}1 & 1 & 1 & 0 & 0 & 0\end{pmatrix}^t$ et $\begin{pmatrix}0 & 0 & 0 & 1 & 1 & 1\end{pmatrix}^t$. Le rang ne peur dépasser 4. Il est égal à 4 à cause des sous-matrices de rang 2.

16. Oui ou Non ? Les figures obtenues représentent 90% de la variabilité …

OUI car la somme des variances donne l'inertie totale (200 e-04) et la somme des deux premières valeurs propres donnent l'inertie projetée (184 e -04). Le rapport dépasse 0.9.

17. Quelle est l'inertie de ce nuage de points ? Par définition 1.

18. Vrai ou faux ? L'inertie de ce nuage de 8 points projeté …

OUI car l'inertie projetée ne peut dépasser la première valeur propre de $\mathbf{X}^t\mathbf{X}$ et doit dépasser la seconde. Comme la matrice $\mathbf{X}^t\mathbf{X}$ est proportionnelle à l'identité (le résultat est le même pour la pondération uniforme ou la pondération unitaire) l'inertie projetée est constante.

19. Vrai ou faux ? L'inertie projetée sur le premier axe principal d'une ACP ...

OUI, car :

$$\left.\begin{aligned} &\lambda_1 + \lambda_2 + ... + \lambda_p = p \\ &\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \end{aligned}\right\} \Rightarrow \lambda_1 \geq 1$$

20. Vrai ou faux ? Dans une matrice de corrélation théorique …

OUI, car :

$$\begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \alpha \\ \alpha & \alpha & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \left(1+\left(p-1\right)\alpha\right) \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Donc $\frac{1}{\sqrt{p}}\mathbf{1}_p$ est axe principal et il existe une composante principale sur laquelle l'inertie projetée vaut la valeur propre $1+\left(p-1\right)\alpha$.

Pour refaire les calculs

	JC&VGE	FM	GM	JC&RB	FM	JMLP
Paris	0.61	0.29	0.1	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	0.5	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	0.4	0.44	0.16
Strasbourg	0.64	0.3	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	0.4	0.36	0.24
Rennes	0.5	0.39	0.11	0.43	0.46	0.11
Toulon	0.55	0.25	0.2	0.41	0.28	0.31
Saint_Etienne	0.52	0.28	0.2	0.42	0.35	0.23
Le_Havre	0.42	0.27	0.31	0.38	0.44	0.18