IUT de Caen - Département STID
Responsable : Alain LUCAS

Analyse des Données

Correction Fiche TD 4

1. Tableau des fréquences exprimées en pourcentage, et distributions marginales :

	Brun	**Châtain**	**Roux**	**Blond**	$f_{i.}$
Marron	11.49	20.11	4.39	1.18	**37.17**
Noisette	**2.53**	9.12	2.36	1.69	15.70
Vert	0.84	4.90	2.36	2.70	10.80
Bleu	3.38	14.19	2.87	15.89	36.33
$f_{.j}$	**18.24**	48.32	11.98	21.46	**100**

2. Les profils-lignes sont les distributions conditionnelles de Y sachant X. On obtient

%	**Brun**	**Châtain**	**Roux**	**Blond**	
Marron	**30.91**	54.11	11.81	3.17	**100**
Noisette	16.11	**58.10**	15.03	10.76	**100**
Vert	7.78	45.37	21.85	25.00	**100**
Bleu	9.30	**39.06**	7.90	43.74	**100**
$f_{.j}$	**18.24**	**48.32**	**11.98**	**21.46**	100

3. Chaque profil-ligne ayant quatre coordonnées, les profils-lignes seront représentés dans l'espace $\mathbb{R}^4$. La somme des coordonnées étant constante et égale à 1, seules trois coordonnées sont suffisantes pour déterminer la position d'un profil. Par conséquent, les profils-lignes vont appartenir à un hyperplan, c'est-à-dire un sous-espace de dimension 3.

4. Les profils-colonnes sont les distributions conditionnelles de X sachant Y. On obtient

%	**Brun**	**Châtain**	**Roux**	**Blond**	$f_{i.}$
Marron	62.99	41.62	36.64	5.50	**37.17**
Noisette	13.87	**18.87**	19.70	7.88	**15.70**
Vert	4.60	10.14	19.70	12.58	**10.80**
Bleu	**18.54**	29.37	23.96	**74.04**	**36.33**
	100	**100**	**100**	**100**	**100**

5. Chaque profil-colonne ayant quatre coordonnées, les profils-colonnes seront représentés dans l'espace $\mathbb{R}^4$. La somme des coordonnées étant constante et égale à 1, seules trois coordonnées sont suffisantes pour déterminer la position d'un profil. Par conséquent, les profils-colonnes vont appartenir à un hyperplan, c'est-à-dire un sous-espace de dimension 3.

6. La matrice des pondérations des profils-lignes est

$$D_n = \begin{pmatrix} 0,3717 & 0 & 0 & 0 \\ 0 & 0,1570 & 0 & 0 \\ 0 & 0 & 0,1080 & 0 \\ 0 & 0 & 0 & 0,3633 \end{pmatrix}$$

La matrice des pondérations des profils-colonnes est

$$D_p = \begin{pmatrix} 0,1824 & 0 & 0 & 0 \\ 0 & 0,4832 & 0 & 0 \\ 0 & 0 & 0,1198 & 0 \\ 0 & 0 & 0 & 0,2146 \end{pmatrix}$$

A partir de ces matrices, et des relations $M_p = D_p^{-1}$ et $M_n = D_n^{-1}$, on obtient

$$M_p = D_p^{-1} = \begin{pmatrix} 5.4825 & 0 & 0 & 0 \\ 0 & 2.0695 & 0 & 0 \\ 0 & 0 & 8.3472 & 0 \\ 0 & 0 & 0 & 4.6598 \end{pmatrix} \quad \text{et} \quad M_n = D_n^{-1} = \begin{pmatrix} 2.6903 & 0 & 0 & 0 \\ 0 & 6.3694 & 0 & 0 \\ 0 & 0 & 9.2593 & 0 \\ 0 & 0 & 0 & 2.7525 \end{pmatrix}$$

7. La matrice S à diagonaliser est $S = F_\ell^t D_n F_\ell D_p^{-1}$. Sachant que l'on a $F_\ell = D_n^{-1} F$, on a aussi $S = F^t D_n^{-1} F D_p^{-1}$. Dans les deux cas, on aboutit à

$$S = \begin{pmatrix} 0.2378 & 0.1942 & 0.1825 & 0.1083 \\ 0.5146 & 0.4956 & 0.4956 & 0.4217 \\ 0.1199 & 0.1228 & 0.1348 & 0.1043 \\ 0.1274 & 0.1873 & 0.1868 & 0.3655 \end{pmatrix}$$

Remarque : La matrice T à diagonaliser pour l'ajustement des profils-colonnes est $T = F_c D_p F_c^t D_n^{-1}$. Sachant que l'on a $F_c = F D_p^{-1}$, on a aussi $T = F D_p^{-1} F^t D_n^{-1}$. Dans les deux cas, on aboutit à

$$T = \begin{pmatrix} 0.4649 & 0.4042 & 0.3316 & 0.2741 \\ 0.1707 & 0.1700 & 0.1591 & 0.1366 \\ 0.0963 & 0.1094 & 0.1240 & 0.1144 \\ 0.2679 & 0.3161 & 0.3851 & 0.4747 \end{pmatrix}$$

8. D'après le cours, on sait que la première valeur propre est toujours égale à l'unité. Cette valeur propre est associée à un vecteur propre engendrant un sous-espace vectoriel de dimension un dit trivial. En pratique, on supprime cette valeur propre qui n'apporte pas d'information. On concentre alors l'analyse sur les trois dernières valeurs propres.

9. D'après le cours, on a

$$I = \sum_{\alpha=1}^{3} I(\alpha) \qquad \text{avec } I(\alpha) = \lambda_\alpha.$$

Par conséquent, sachant que les valeurs propres sont $\lambda_1 = 0.20877$, $\lambda_2 = 0.02223$ et $\lambda_3 = 0.0026$, on obtient :

$$I(1) = 0.20877 \qquad I(2) = 0.02223 \qquad I(3) = 0.0026$$

avec $I = 0.23360$.

On déduit alors que

$$PI(1) = 89.37\% \qquad PI(2) = 9.52\% \qquad PI(3) = 1.11\%$$

10. Sous la contrainte de devoir conserver au moins 90% d'information, on est amené à garder les deux premiers axes factoriels.

11. D'après le cours, les coordonnées des profils-lignes sur les deux premiers axes factoriels (les facteurs) se calculent via le produit matriciel suivant : $\Psi_\alpha = F_\ell D_p^{-1} u_\alpha$. Ainsi, on obtient

$$\Psi_1 = \begin{pmatrix} \mathbf{-0.4925} \\ -0.2125 \\ 0.1617 \\ 0.5474 \end{pmatrix} \quad \text{et} \quad \Psi_2 = \begin{pmatrix} -0.0883 \\ 0.1673 \\ \mathbf{0.3391} \\ -0.0831 \end{pmatrix}$$

12. D'après le cours, on a

$$Cr_\alpha(i) = f_{i.} \frac{\Psi_\alpha^2(i)}{\lambda_\alpha}$$

On obtient alors pour les profils-lignes

Profils-lignes	$Cr_1(i)$	$Cr_2(i)$
1	**43.18**	13.04
2	3.40	19.79
3	1.35	**55.83**
4	52.14	11.26

13. D'après le cours, on a

$$Qual_\alpha(i) = \frac{\Psi_\alpha^2(i)}{\|G_\ell \widetilde{E}_i\|^2_{M_p}}$$

On obtient alors pour les profils-lignes

Profils-lignes	$Qual_1(i)$	$Qual_2(i)$	$Qual_{1\times 2}(i)$
1	96.70	2.78	**99.48**
2	**54.24**	33.58	87.82
3	17.59	76.77	94.36
4	97.75	2.25	**100.00**

On peut observer que l'ensemble des profils-lignes présente une très bonne qualité de représentation.

14. D'après le cours, on a la relation

$$\sqrt{\lambda_\alpha}\varphi_\alpha(j) = \sum_{i=1}^{4} f_{i/j}\Psi_\alpha(i), \qquad \forall\ j = 1, \cdots, 4.$$

On obtient alors

$$\varphi_1 = \begin{pmatrix} \mathbf{-0.5046} \\ -0.1483 \\ -0.1296 \\ 0.8349 \end{pmatrix} \quad \text{et} \quad \varphi_2 = \begin{pmatrix} -0.2148 \\ 0.0327 \\ \mathbf{0.3196} \\ -0.0696 \end{pmatrix}$$

15. D'après le cours, on a

$$Cr_\alpha(j) = f_{.j}\frac{\varphi_\alpha^2(j)}{\lambda_\alpha}$$

On obtient alors pour les profils-colonnes

Profils-colonnes	$Cr_1(j)$	$Cr_2(j)$
1	22.25	**37.88**
2	5.09	2.32
3	0.96	55.13
4	**71.7**	4.67

16. D'après le cours, on a

$$Qual_\alpha(j) = \frac{\varphi_\alpha^2(j)}{\|G_c \widetilde{V}_j\|^2_{M_n}}$$

Par conséquent, on obtient

Profils-colonnes	$Qual_1(j)$	$Qual_2(j)$	$Qual_{1\times 2}(j)$
1	**83.70**	15.12	**98.28**
2	86.59	4.21	90.80
3	13.36	**81.26**	94.62
4	99.14	0.69	**99.83**

On peut observer que l'ensemble des profils-colonnes présente une très bonne qualité de représentation.

17. Via la méthode explicitée dans le TD, on obtient le tableau

Axe Factoriel 1		**Axe Factoriel 2**	
Coord. négative	**Coord. positive**	**Coord. négative**	**Coord. positive**
Marron (43%)	Bleu (52%)		Noisette (20%) Vert (56%)
Brun (22%)	Blond (72%)	Brun (38%)	Roux (55%)

Concernant le premier axe, on note clairement une opposition entre d'un côté des couleurs sombres (Marron et Brun), et de l'autre côté des couleurs claires (Bleu et Blond). Concernant le deuxième axe, l'interprétation est plus difficile du fait d'une faible part d'inertie. Cet axe est vraisemblablement caractéristique d'une structure moins forte que la première. A travailler par la suite...

18. (a) Le graphique permettant d'obtenir les profils-lignes en fonction des profils-colonnes s'obient à l'aide de la relation quasi-barycentrique suivante :

$$\sqrt{\lambda_\alpha}\Psi_\alpha(i) = \sum_{j=1}^{4} f_{j/i}\varphi_\alpha(j), \qquad \forall\ i = 1, \cdots, 4.$$

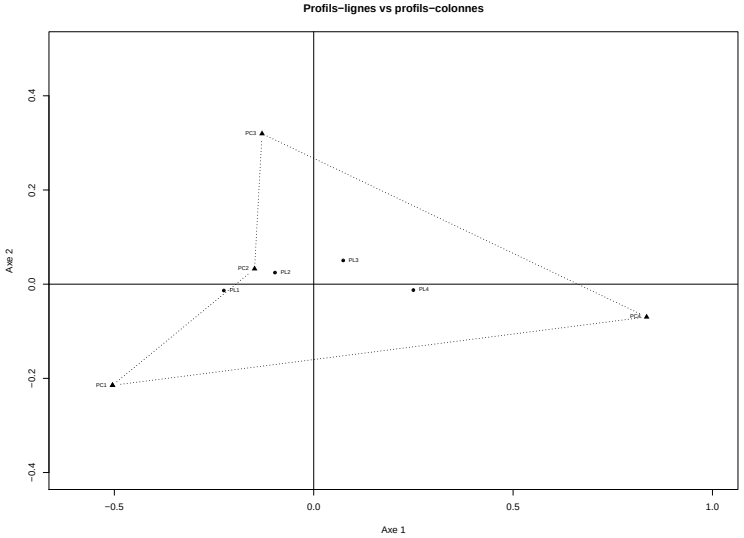
Pour réaliser le graphique, il convient de placer dans un premier temps les profils-colonnes à l'aide des coordonnées fournies par φ_α, puis de placer les profils-lignes en calculant les vecteurs $\sqrt{\lambda_\alpha}\Psi_\alpha$. Ce calcul permet d'aboutir aux vecteurs suivants :

$$\sqrt{\lambda_1}\Psi_1 = \begin{pmatrix} -0.2250 \\ -0.0971 \\ 0.0739 \\ 0.2501 \end{pmatrix} \quad \text{et} \quad \sqrt{\lambda_2}\Psi_2 = \begin{pmatrix} -0.0132 \\ 0.0249 \\ 0.0506 \\ -0.0124 \end{pmatrix}$$

Pour obtenir une représentation graphique sous R, il convient d'écrire le programme suivant :

- x=c(-0.5046,-0.1483,-0.1296,0.8349)
- y=c(-0.2148,0.0327,0.3196,-0.0696)
- x1=c(x,-0.5046)
- y1=c(y,-0.2148)
- plot(x1,y1,type="b",main="Profils-lignes par rapport aux profils-colonnes",xlab=" ",ylab=" ", xlim=c(-0.6,1),ylim=c(-0.4,0.5),lty=3)
- abline(v=0) ;abline(h=0)
- text(x,y,labels=c("PC1","PC2","PC3","PC4"),pos="2",cex=0.6)
- z1=c(-0.225,-0.0971,0.0739,0.2501)
- z2=c(-0.0132,0.0249,0.0506,-0.0124)
- points(z1,z2,cex=0.6)
- text(z1,z2,cex=0.6,pos="1",labels=c("PL1","PL2","PL3","PL4"))

Finalement, on obtient le graphique suivant :



A l'aide de cette représentation graphique, on observe que les profils-lignes PL1 et PL2 sont plutôt attirés par la modalité PC2. A l'inverse, le profil-ligne PL4 semble être attiré plutôt par la modalité PC4.

(b) Le graphique permettant d'obtenir les profils-colonnes en fonction des profils-lignes s'obient à l'aide de la relation quasi-barycentrique suivante :

$$\sqrt{\lambda_\alpha}\varphi_\alpha(j) = \sum_{i=1}^{4} f_{i/j}\Psi_\alpha(i), \qquad \forall\ j = 1, \cdots, 4.$$

Pour réaliser le graphique, il convient de placer dans un premier temps les profils-ligness à l'aide des coordonnées fournies par Ψ_α, puis de placer les profils-colonnes en calculant les vecteurs $\sqrt{\lambda_\alpha}\varphi_\alpha$. Ce calcul permet d'aboutir aux vecteurs suivants :

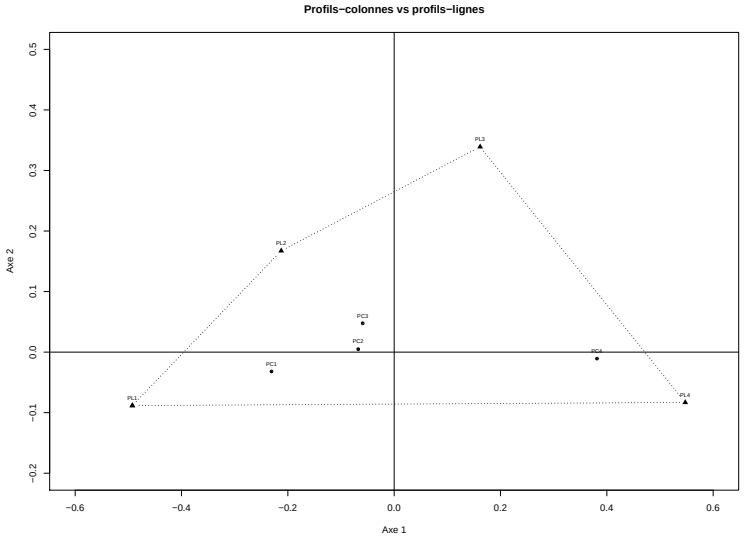
$$\sqrt{\lambda_1}\varphi_1 = \begin{pmatrix} -0.2306 \\ -0.0678 \\ -0.0592 \\ 0.3815 \end{pmatrix} \qquad \text{et} \qquad \sqrt{\lambda_2}\varphi_2 = \begin{pmatrix} -0.0320 \\ 0.0049 \\ 0.0477 \\ -0.0104 \end{pmatrix}$$

Pour obtenir une représentation graphique sous R, il convient d'écrire le programme suivant :

- x=c(-0.4925,-0.2125,0.1617,0.5474)
- y=c(-0.0883,0.1673,0.3391,-0.0831)
- x1=c(x,-0.4925)
- y1=c(y,-0.0883)
- plot(x1,y1,type="b",main="Profils-colonnes par rapport aux profils-lignes",xlab=" ",ylab=" ", xlim=c(-0.6,0.6),ylim=c(-0.2,0.5),lty=3)
- abline(v=0) ;abline(h=0)
- text(x,y,labels=c("PL1","PL2","PL3","PL4"),pos="2",cex=0.6)
- z1=c(-0.2306,-0.0678,-0.0592,0.3815)
- z2=c(-0.0320,0.0048,-0.0104,-0.032)

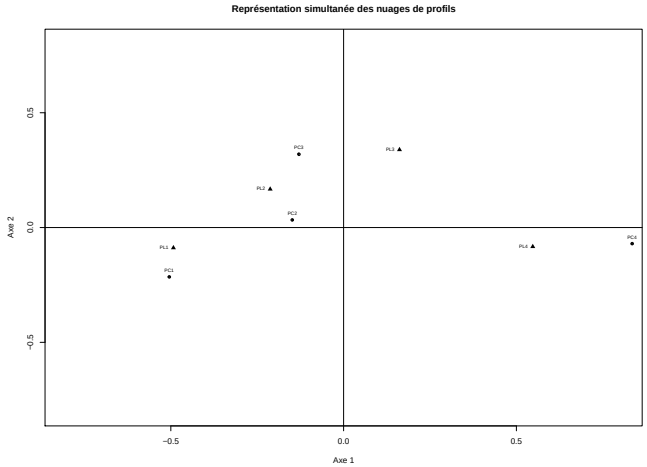
- points(z1,z2,cex=0.6)
- text(z1,z2,cex=0.6,pos="1",labels=c("PC1","PC2","PC3","PC4"))

Finalement, on obtient le graphique suivant :



A l'aide de cette représentation graphique, on note que le profil-colonne PC1 est plutôt attiré par la modalité PL1 tandis que le profil colonne PC4 est plutôt attiré par la modalité PL4.

19. En faisant usage des facteurs Ψ et φ, on obtient le graphique suivant :



Pour obtenir ce graphique sous R, le programme suivant a été écrit :

- u1=c(-0.225,-0.0971,0.0739,0.2501)
- u2=c(-0.0132,0.0249,0.0506,-0.0124)
- z1=c(-0.2306,-0.0678,-0.0592,0.3815)
- z2=c(-0.0320,0.0048,-0.0104,-0.032)
- plot(u1,u2,type="p",main="Représentation simultanée des deux nuages de profils",xlab=" ",ylab=" ", xlim=c(-0.4,0.4),ylim=c(-0.1,0.1))
- abline(h=0) ;abline(v=0)
- text(u1,u2,labels=c("PL1","PL2","PL3","PL4"),cex=0.6,pos="2")
- points(z1,z2)
- text(z1,z2,labels=c("PC1","PC2","PC3","PC4"),cex=0.6,pos="3")

20. Clairement, on retrouve les commentaires qui ont été faits précédemment. En effet, on visualise clairement l'attirance entre les modalités PC1 et PL1 d'une part, PC4 et PL4 d'autre part.

21. La lecture de ce graphique montre que les deux variables ne sont manifestement pas indépendantes. En effet, dans le cas contraire, les profils seraient semblables et concentrer autour du centre du repère, ce qui n'est pas le cas ici.

22. L'inertie totale des nuages vaut $I = 0.2336$. D'après le cours, le produit de l'inertie totale par l'effectif total est égal à la statistique du khi-deux. Autrement dit, on obtient $\chi^2 = k \times I = 138,2912$.

23. Dans le cadre d'un test d'indépendance du khi-deux, le nombre de degré de liberté est égal à $dl = (n-1)(p-1) = 9$. Le niveau du test étant ici fixé à 5%, on obtient à partir d'une table de la loi du khi-deux que le seuil c du test est $c = 16,92$. Etant donné que $\chi^2 >> c$, on rejette l'hypothèse d'indépendance.

24. Il convient d'observer dans premier temps que l'axe 1 porte l'essentiel de l'information. Par conséquent, on va être amené à observer la répartition des profils-lignes plutôt sur cet axe que sur le second. Dans ce cadre, on note que les profils-lignes PL1 et PL4 sont non seulement très bien représentés sur l'axe 1, mais sont aussi très éloignés du barycentre, et diamètralement opposés. Par conséquent, ces profils sont non seulement très différents l'un de l'autre, mais aussi très différents du profil moyen.

25. Pour l'analyse des profils-colonnes, on va procéder de la même manière que pour les profils-lignes, autrement dit en se limitant au premier axe factoriel. Sur cet axe, trois profils-colonnes sont très bien représentés : PC1, PC2, et PC4. Par ailleurs, on note que les profils PC1 et PC4 sont non seulement diamètralement opposés, mais aussi très différents du profil moyen. Enfin, le PC2 est légèrement différent du profil moyen, tout en étant plus proche de PC1 que de PC4.

26. On obtient

$$\frac{1}{\sqrt{\lambda_1}} = 2.2 \qquad \text{et} \qquad \frac{1}{\sqrt{\lambda_2}} = 6.7$$

Ces coefficients interviennent comme coefficients multiplicatifs dans le cadre des relations quasi-barycentriques. Si ces coefficients n'étaient pas présents, la coordonnée (par exemple) d'un profil-ligne i serait exactement le barycentre des profils-colonnes affectés d'un poids fonction du profil-ligne i. Or, ce coefficient multiplicatif vient modifier artificiellement la position de cette coordonnée. Ainsi, si ce coefficient est peu élevé (par rapport à 1), la représentation graphique est peu déformée tandis que si ce coefficient est élevé, la représentation graphique subit de profondes modifications pouvant amenées des erreurs d'interprétations.

27. Dans le cas présent, on va s'appuyer sur une interprétation selon le premier axe factoriel et ce pour tois raisons. La première est que le second axe factoriel apporte peu d'information (à peine 10%). La seconde est que le coefficient de distorsion est trois fois plus élevé sur cet axe que sur le premier.La troisième est que selon cet axe peu d'interactions sont visibles entre les profils des différents nuages.
Concernant le premier axe factoriel, on peut noter que les profils Marron et Brun semblent s'attirer. Autrement dit, les individus qui ont les yeux Marrons se distinguent vraisemblablement du profil moyen par une fréquence d'apparition des cheveux Bruns élevée, par rapport à la moyenne, et dans le même temps (du fait de l'opposition) d'une fréquence d'apparition des cheveux Blond plus faible que la moyenne. Inversement, les individus qui ont les cheveux bruns se distinguent vraisemblablement du profil moyen par une forte fréquence de la modalité Marron et dans le même temps d'une faible féquence de la modalité Bleu. On peut, de l'autre côté de l'axe, procéder à une analyse en tout point similaire acec les profils Bleu et Blond qui semblent s'attirer, et s'opposer aux profils Marron et Brun.

28. On peut vérifier les commentaires à l'aide des écarts relatifs entre les profils et leur barycentre (utiliser la formule $b - a/a$) :

	PC1	PC2	PC3	PC4
PL1	30.91	54.11	11.81	03.17
PL4	09.30	39.06	07.90	43.74
G_ℓ	18.24	48.32	11.98	21.46
PL1-rel	+69	+12	-1.4	-85
PL4-rel	-50	-19	-34	+104

A partir de cette table, on peut noter que le PL1 se distingue du profil moyen par une valeur significativement supérieure pour le PC1 (+69%), et une valeur significativement inférieure pour PC4 (-85%). De même, le PL4 se distingue du profil moyen par une valeur significativement supérieure pour PC4 (+104%) et significativement inférieure pour PC1 (-50%).

On peut procéder de même pour l'étude des profils-colonnes par rapport aux profils-lignes :

	PL1	PL2	PL3	PL4
PC1	62.99	13.87	04.60	18.54
PC4	05.50	07.88	12.58	74.04
G_c	37.17	15.70	10.80	36.33
PC1-rel	+69	+12	-57	-49
PC4-rel	-85	-50	+16	+104

A partir de cette table, on peut noter que le PC1 se distingue du profil moyen par une valeur significativement supérieure pour le PL1 (+69%), et une valeur significativement inférieure pour PL3 (-57%) et PL4 (-49%). De même, le PC4 se distingue du profil moyen par une valeur significativement supérieure pour PC4 (+104%) et significativement inférieure pour PL1 (-85%) et PL2 (-50%).