Analyse Factorielle des Correspondances

GHAZI BEL MUFTI

ghazi.belmufti@gmail.com

ESSAI-1 / ANALYSE DES DONNÉES

Plan

Introduction

1. Principes de l'AFC

2. Résolution de l'AFC

3. Règles d'interprétation

4. Exemple d'application

Introduction

L'AFC est une méthode d'analyse de données qualitatives. Elle permet d'analyser des tableaux de contingence issus d'un tri croisé de 2 variables nominales observées sur *n* individus.

Exemple 1 : Supposons que l'on trie les tunisiens selon 2 variables : la région dans laquelle ils habitent et leur classe d'âge. Se posent alors les questions suivantes :
• quelles sont les régions dont la structure démographique est proche, et, à l'opposé, quelles-sont les régions qui ont des pyramides des âges totalement différentes ?
• quelles sont les régions où les jeunes sont relativement nombreux ?

Exemple 2 : quel lien existe-t-il entre la région d'une entreprise et son secteur d'activité

Le test du χ^2 mesure la significativité de la liaison entre les deux variables ; si cette liaison est significative, l'AFC permet de la spécifier et de la décrire.

Tableau de contingence I

On étudie le lien entre deux variables qualitatives X et Y, X à p modalités et Y à q modalités. Le tableau de contingence n_{IJ} est une matrice de format $p \times q$. On pose $I = \{1, \ldots, p\} = [\![1, p]\!]$ et $J = \{1, \ldots, q\} = [\![1, q]\!]$.

		Y			
		.	mod j	.	tot
	.	.	.	.	.
X	mod i	.	n_{ij}	.	$n_{i.}$
	.	.	.	.	.
	tot	.	$n_{.j}$	.	n

On a alors les relations : $n_{i.} = \sum_{j=1}^{q} n_{ij}$, $n_{.j} = \sum_{i=1}^{p} n_{ij}$ et $n = \sum_{i,j} n_{ij}$.

Tableau de contingence II

On transforme les effectifs en fréquences : on obteient un nouveau tableau F de terme courant :

$$\forall (i,j) \in I \times J, f_{ij} = \frac{n_{ij}}{n}.$$

On a les lois marginales :

$$f_I = (f_{i.})_{i \in I} \in \mathbb{R}^p \text{ avec } f_{i.} = \sum_{j=1}^{q} f_{ij},$$

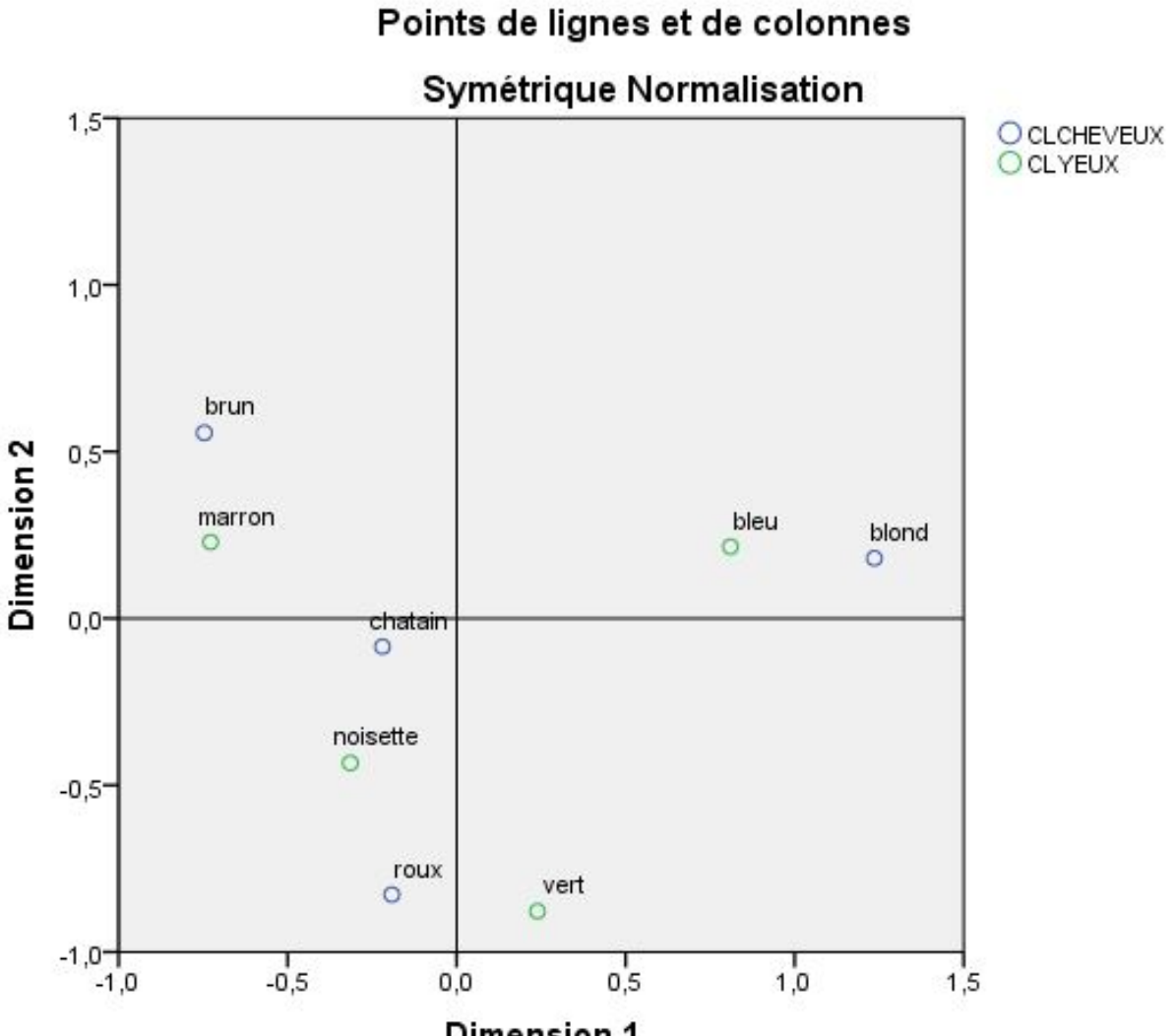
$$f_J = (f_{.j})_{j \in J} \in \mathbb{R}^q \text{ avec } f_{.j} = \sum_{i=1}^{p} f_{ij},$$

Exemple

		couleur des cheveux				
		brun	châtain	roux	blond	tot
couleur des yeux	marron	68	119	26	7	220
	noisette	15	54	14	10	93
	vert	5	29	14	16	64
	bleu	20	84	17	94	215
	tot	108	286	71	127	592

Question : y-a-t-il indépendance entre la couleur des yeux et celle des cheveux ? Sinon quels types d'associations existent entre ces couleurs ?

Principal output de l'AFC : la représentation simultanée



Transformations du tableau de contingence I

• **Tableau des Profils-colonnes :** pour tout $j \in J$, on introduit la loi conditionnnelle sur I sachant j appellé profil de la colonne j (i.e. la répartition en pourcentage des modalités $X_1, ..., X_p$ parmi les individus possédant la modalité Y_j) :

$$f_I^j = \begin{pmatrix} f_1^j \\ \vdots \\ f_p^j \end{pmatrix} = \begin{pmatrix} f_{1j}/f_{.j} \\ \vdots \\ f_{pj}/f_{.j} \end{pmatrix} \text{ et } f_I^J = (f_i^j), i \in I, j \in J$$

		couleur des cheveux				
		brun	châtain	roux	blond	profil moyen
couleur des yeux	marron	63	42	37	6	37
	noisette	14	19	20	8	16
	vert	5	10	20	13	11
	bleu	19	29	24	74	36
	tot	100	100	100	100	100

Transformations du tableau de contingence II

- **Tableau des Profils-lignes :** pour tout $i \in I$, on introduit la loi conditionnnelle sur J sachant i appellé profil de la ligne i (i.e. la répartition en pourcentage des modalités $Y_1, ..., Y_q$ parmi les individus possédant la modalité X_i) :

$$f_J^i = \begin{pmatrix} f_1^i \\ \vdots \\ f_q^i \end{pmatrix} = \begin{pmatrix} f_{i1}/f_{i.} \\ \vdots \\ f_{iq}/f_{i.} \end{pmatrix} \text{ et } f_J^I = (f_j^i), i \in I, j \in J$$

		couleur des cheveux				
		brun	châtain	roux	blond	tot
couleur des yeux	marron	31	54	12	3	100
	noisette	16	58	15	11	100
	vert	8	45	22	25	100
	bleu	9	39	8	44	100
	profil moyen	18	48	12	22	100

Choix des distances I

- La distance euclidienne usuelle entre 2 profils-lignes :

$$d^2(i,i') = \sum_{j=1}^{q} (f_{ij}/f_{i.} - f_{i'j}/f_{i'.})^2$$

Cette distance favorise les colonnes qui ont une masse $f_{.j}$ importante (i.e dans l'exemple, les couleurs de cheveux qui sont bien représentées dans la population étudiée).

L'aternative à la distance euclidienne est la disatnce du χ_2 qui est plus appropriée pour analyser les tableaux des profils lignes et colonnes :

- *Distance du* χ^2 entre deux profils-lignes :

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{q} \frac{1}{f_{.j}} (f_{ij}/f_{i.} - f_{i'j}/f_{i'.})^2$$

Choix des distances II

- *Distance du* χ^2 entre les profils-colonnes :

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{p} \frac{1}{f_{i.}}(f_{ij}/f_{.j} - f_{ij'}/f_{.j'})^2$$

Exemples :

- *Distance du* χ^2 entre les profils-colonnes brun et roux :

$$d^2_{\chi^2}(1,3) = \frac{1}{.37}(.63-.37)^2+\frac{1}{.16}(.14-.20)^2+\frac{1}{.11}(.5-.20)^2+\frac{1}{.36}(.19-.24)^2$$

- *Distance du* χ^2 entre les deux premiers profils-lignes qui sont marron et noisette :

$$d^2_{\chi^2}(1,2) = \frac{1}{.18}(.31-.16)^2+\frac{1}{.48}(.54-.58)^2+\frac{1}{.12}(.12-.15)^2+\frac{1}{.22}(.3-.11)^2$$

Nuages des lignes et des colonnes

- $\mathcal{N}_c = \{f_I^j, j \in J\}$, appelé nuage des profils-colonnes, où chaque point $f_I^j \in \mathbb{R}^p$ est muni du poids f_j.
  Le centre de gravité g_c de $\mathcal{N}_c$ = le profil moyen $(f_{1.}, ..., f_{p.})$

  En effet, on a : $g_c = f_I^J D_{f_J} 1_q = F 1_q = f_I$

- $\mathcal{N}_I = \{f_J^i, i \in I\}$, appelé nuage des profils-lignes, où chaque point $f_J^i \in \mathbb{R}^q$ est muni du poids f_i.
  Le centre de gravité g_I de $\mathcal{N}_I$ = le profil moyen $(f_{.1}, ..., f_{.q})$

Résolution de l'AFC I

L'AFC consiste à effectuer deux ACP sur deux nuages différents (i.e. $\mathcal{N}_c$ et $\mathcal{N}_l$) mais présentant une certaine symétrie.
On note :

$$D_{f_l} = Diag(f_i)_{i\in I} \in \mathcal{M}_p(\mathbb{R}) \text{ et } D_{f_J} = Diag(f_j)_{j\in J} \in \mathcal{M}_q(\mathbb{R})$$

On a

$$D_{f_l}^{-1} = Diag(1/f_i)_{i\in I} = D_{1/f_l} \text{ et } D_{f_J}^{-1} = Diag(1/f_j)_{j\in J} = D_{1/f_J}$$

- Une ACP où les individus sont les points-colonnes c_j, la métrique est $D_{f_l}^{-1}$ et la matrice des poids est D_{f_J}.
- Une ACP où les individus sont les points-lignes l_i, la métrique est $D_{f_J}^{-1}$ et la matrice des poids est D_{f_l}.

ACP sur Les profils-colonnes

On note

$$F_1 = f_I^J = (f_I^1, \ldots, f_I^q) \text{ et } F_2 = f_J^I = (f_J^1, \ldots, f_J^p)$$

F_1 est le tableau des profils colonnes et F_2 des profils lignes.
On a alors :

$$F_1 = FD_{1/f_J} \text{ et } F_2 = F'D_{1/f_I}$$

L'ACP du nuage $\mathcal{N}_c$ consiste à diagonaliser F_1F_2.

Remarque : En effet, si V est la matrice variance du nuage des profils colonnes, on montre que $VM = VD_{1/f_I}$ et la matrice $F_1D_{f_J}F_1'D_{1/f_I}$ ont les mêmes valeurs propres et que $F_1D_{f_J}F_1'D_{1/f_I} = F_1F_2$.

- Les axes factoriels sont solutions de :

$$\begin{cases} F_1F_2u_I^\alpha &= \lambda_\alpha u_I^\alpha \\ < u_I^\alpha, u_I^\beta >_{D_{1/f_I}} &= \delta_{\alpha,\beta} \end{cases}$$

- Les composantes principales sont solutions de :

$$\begin{cases} F_1'F_2'\Psi_\alpha^J &= \lambda_\alpha \Psi_\alpha^J \\ < \Psi_\alpha^J, \Psi_\alpha^J >_{D_{f_J}} &= \lambda_\alpha \delta_{\alpha,\beta} \end{cases}$$

ACP sur Les profils-lignes

L'étude du nuage des profils-lignes $\mathcal{N}_l$ se déduit de celle de $\mathcal{N}_c$ en intervertissant les rôles de I te de J. On échange F_1 et F_2.

L'ACP du nuage $\mathcal{N}_c$ consiste à diagonaliser $F_1 F_2$.

- Les axes factoriels sont solutions de :

$$\begin{cases} F_2 F_1 u_J^\alpha & = \lambda_\alpha u_J^\alpha \\ < u_J^\alpha, u_J^\beta >_{D_{1/f_J}} & = \delta_{\alpha,\beta} \end{cases}$$

- Les composantes principales sont solutions de :

$$\begin{cases} F_2' F_1' \Psi_\alpha^I & = \lambda_\alpha \Psi_\alpha^I \\ < \Psi_\alpha^I, \Psi_\beta^I >_{D_{f_I}} & = \lambda_\alpha \delta_{\alpha,\beta} \end{cases}$$

Relations quasi-barycentriques : lien entre les deux analyses

- Les 2 ACP ont les mêmes valeurs propres non nuls λ_α
- Il existe des relations entre les coordonnées des p-l et des p-c sur l'axe α dites *relations quasi-barycentriques* :

$$\begin{cases} \psi_\alpha^i = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{j=1}^{q} \frac{f_{ij}}{f_{i.}} \psi_\alpha^j \\ \\ \psi_\alpha^j = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{i=1}^{p} \frac{f_{ij}}{f_{.j}} \psi_\alpha^i \end{cases}$$

Conséquences :

- Les valeurs propres non triviales sont ≤ 1.
- Représentation simultanée des p-l et des p-c

Hypothèse d'indépendance I

Une condition nécessaire pour que l'AFC soit pertinente c'est que le lien entre les 2 variables qualitatives soit statistiquement significatif : mais comment mesurer ce lien ?

- Pour répondre à cette question nous allons nous inspirer de ce que nous savons sur les couples de v.a. Ainsi 2 v.a. $\tilde{X}$ et $\tilde{Y}$ sont indépendantes si pour tout $i \leq p$ et $j \leq q$:

$$p_{ij} = p_{i.}\, p_{.j} \quad (p_{ij} \text{ loi jointe du couple de v.a})$$

Hypothèse d'indépendance II

- Nous nous inspirons de ce principe pour l'appliquer à nos variables X et Y (même si elles ne sont pas aléatoires...) :

$$\hat{f}_{ij} = f_{i.}\, f_{.j} \text{ pour tout } i \leq p \text{ et } j \leq q$$

où $\hat{f}_{ij}$ est la fréquence jointe que l'on devrait avoir dans le tableau de fréquence dans le cas d'indépendance entre X et Y.

D'où

$$\frac{\hat{n}_{ij}}{n} = \frac{n_{i.}}{n}\,\frac{n_{.j}}{n} \text{ pour tout } i \leq p \text{ et } j \leq q$$

On en déduit que les effectifs $\hat{n}_{ij}$ que l'on devrait avoir dans le tableau de effectifs dans le cas d'indépendance entre X et Y sont donnés par :

$$\hat{n}_{ij} = \frac{n_{i.}\, n_{.j}}{n} \text{ pour tout } i \leq p \text{ et } j \leq q$$

Hypothèse d'indépendance III

- Pour répondre à la question que nous nous sommes posé et qui est "Comment mesurer ce lien ?", la réponse est par la statistique suivante dite statistique du χ_2 et dont la formule est donnée par :

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Ou encore : $\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n_{i.} n_{.j}/n)^2}{n_{i.} n_{.j}/n} = \sum_i \sum_j n \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$

Ainsi cette statistique mesure l'écart entre les effectifs observés n_{ij} et les effectifs $\hat{n}_{ij}$ que l'on devrait avoir dans le tableau des effectifs dans le cas d'indépendance entre X et Y :

Plus la valeur du χ^2 est grande, plus l'écart entre les n_{ij} et les effectifs $\hat{n}_{ij}$ est grand et plus on s'éloigne de l'indépendance entre les variables X et Y.

Hypothèse d'indépendance IV

- Hyp. d'indépendance sur les profils-lignes : ils sont tous identiques.

$$f_{ij}/f_{i.} = f_{i'j}/f_{i'.} = f_{.j} \text{ pour } j \leq q, i \text{ et } i' \leq p$$

- Exemple : si tous les profils "coul. des yeux" sont identiques entre eux, et par conséquent identiques au profil moyen correspondant, il y a indépendance entre les couleurs des yeux et celles des cheveux.
- De même, l'hyp. d'indépendance sur les p-c : $f_{ij}/f_{.j} = f_{i.} \quad \forall i$
- Examiner les proximités entre les profils revient à examiner la proximité entre chaque profil et son profil moyen.

Inertie I

- L'information contenue dans le nuage $\mathcal{N}_l$ est mesurée par la dispersion des profils-lignes autour de leur centre de gravité :

$$\begin{aligned} I(\mathcal{N}_l) &= \sum_{i=1}^{p} f_{i.} d^2_{\chi^2}(g_l, l_i) \\ &= \sum_i \sum_j \frac{f_{i.}}{f_{.j}} (f_{.j} - f_{ij}/f_{i.})^2 \\ &= \sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} \end{aligned}$$

Inertie II

- La quantité précédente étant symétrique en i et j, d'où :

$$I(\mathcal{N}_c) = I(\mathcal{N}_l) = \sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$$

- Il en découle que :

$$I_T = I(\mathcal{N}_l) = I(\mathcal{N}_c) = \chi^2 / n$$

L'inertie totale dans une AFC vaut ainsi χ^2/n.

Inertie et test d'indépendance I

- L'inertie s'exprime également par $I_T = \sum_{\alpha=1}^{min(p-1\,;\,q-1)} \lambda_\alpha$
- Comme on a :

$$n \sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} \rightsquigarrow \chi^2_{(p-1)(q-1)}$$

on en conclut que la quantité

$$nI_T = n \sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$$

permet de tester l'hypothèse H_0 d'indépendance des variables X et Y.

Inertie et test d'indépendance II

- Dans le cas de l'indépendance, nI_T aura tendance à être faible, et par conséquent, puisque I_T est la somme des valeurs propres, plus les valeurs propres sont faibles moins les facteurs sont interprétables.

Dimension	Inertie	Khi-deux	Sig	Proportion d'inertie	
				Pris en compte	Cumulé
1	0,209			0,894	0,894
2	0,022			0,095	0,989
3	0,003			0,011	1,000
Total	0,234	138,290	0,000[a]	1,000	

[a]. 9 degrés de liberté

Formes de nuages

La valeur de l'inertie est un indicateur de la dispersion du nuage :

- 2 variables sont indépendantes : l'inertie totale est faible et il n'existe pas de direction privilégiée et tous les points sont concentrés autour du centre de gravité du nuage suivant une forme sphérique.
- Une valeur propre qui tend vers 1 : pour chaque variable 2 groupes de modalités séparant le nuage de points en deux sous-nuages.
- Deux valeurs propres sont proches de 1 : on obtient trois sous-nuages et les modalités des variables se décomposent en trois groupes.
- Toutes les valeurs propres sont proches de 1 : chaque modalité est en correspondance presque exclusive avec une seule modalité de l'autre variable.

Contributions et Cosinus carrés I

Contributions

- La part de la ligne *i* dans la variance prise en compte sur l'axe α :

$$CTR(i,\alpha) = \frac{f_{i.}(\psi_\alpha^i)^2}{\lambda_\alpha}$$

- Celle de la colonne *j* à l'axe α par :

$$CTR(j,\alpha) = \frac{f_{.j}(\psi_\alpha^j)^2}{\lambda_\alpha}$$

Cosinus carrés

- La proximité entre deux points projetés sur l'axe α correspond d'autant mieux à leur distance réelle que les points sont plus proches de l'axe.

Contributions et Cosinus carrés II

- La qualité de représentation du point *i* sur l'axe α est évaluée par le cosinus de l'angle entre l'axe et le vecteur joignant le centre de gravité du nuage au point *i* :

$$cos^2_\alpha(i) = \frac{(\psi^i_\alpha)^2}{d^2_{\chi^2}(i, g_l)}$$

où $d^2_{\chi^2}(i, g_l) = \sum_{j=1}^{q} \frac{1}{f_{.j}}(f_{ij}/f_{i.} - f_{.j})^2$.

- De même :

$$cos^2_\alpha(j) = \frac{(\psi^j_\alpha)^2}{d^2_{\chi^2}(j, g_c)}$$

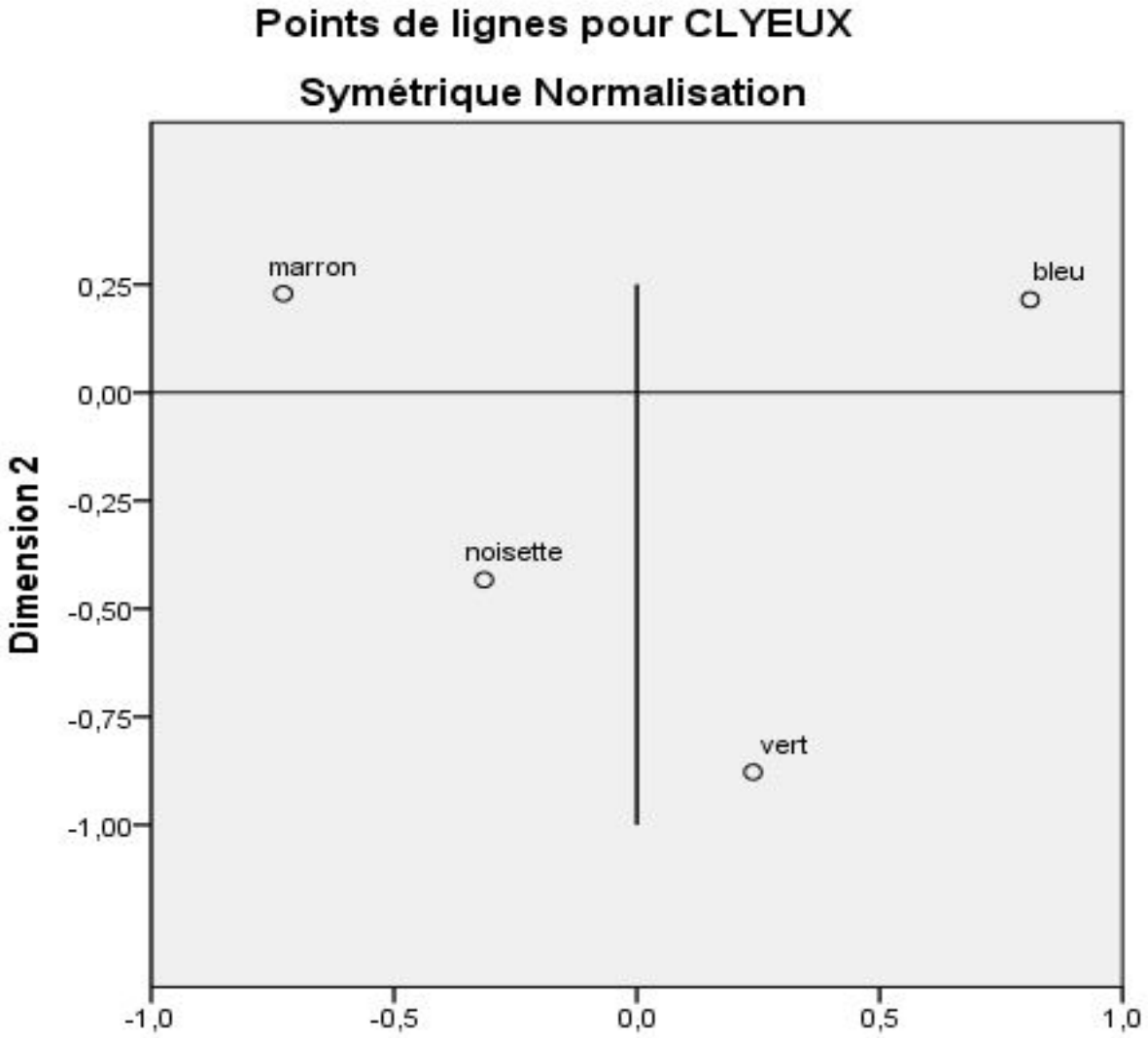
Profils-lignes I

Le premier axe est construit presque exclusivement par les yeux "marrons" et "bleus" (contributions de 43,1% et 52,1%), points situés pratiquement sur l'axe (cosinus carrés de 0,97 et 0,98).
Le second axe est surtout lié aux yeux verts.

Caractéristiques des points lignes[a]

CLYEUX	Masse	Score dans la dimension		Inertie	Contribution				
					De point à inertie de dimension		De dimension à inertie de point		
		1	2		1	2	1	2	Total
marron	,372	-,728	,229	,093	,431	,130	,967	,031	,998
noisette	,157	-,315	-,434	,013	,034	,198	,542	,336	,879
vert	,108	,239	-,878	,016	,014	,559	,176	,773	,948
bleu	,363	,810	,215	,111	,521	,112	,977	,022	1,000
Total actif	1,000			,234	1,000	1,000			

Profils-lignes II



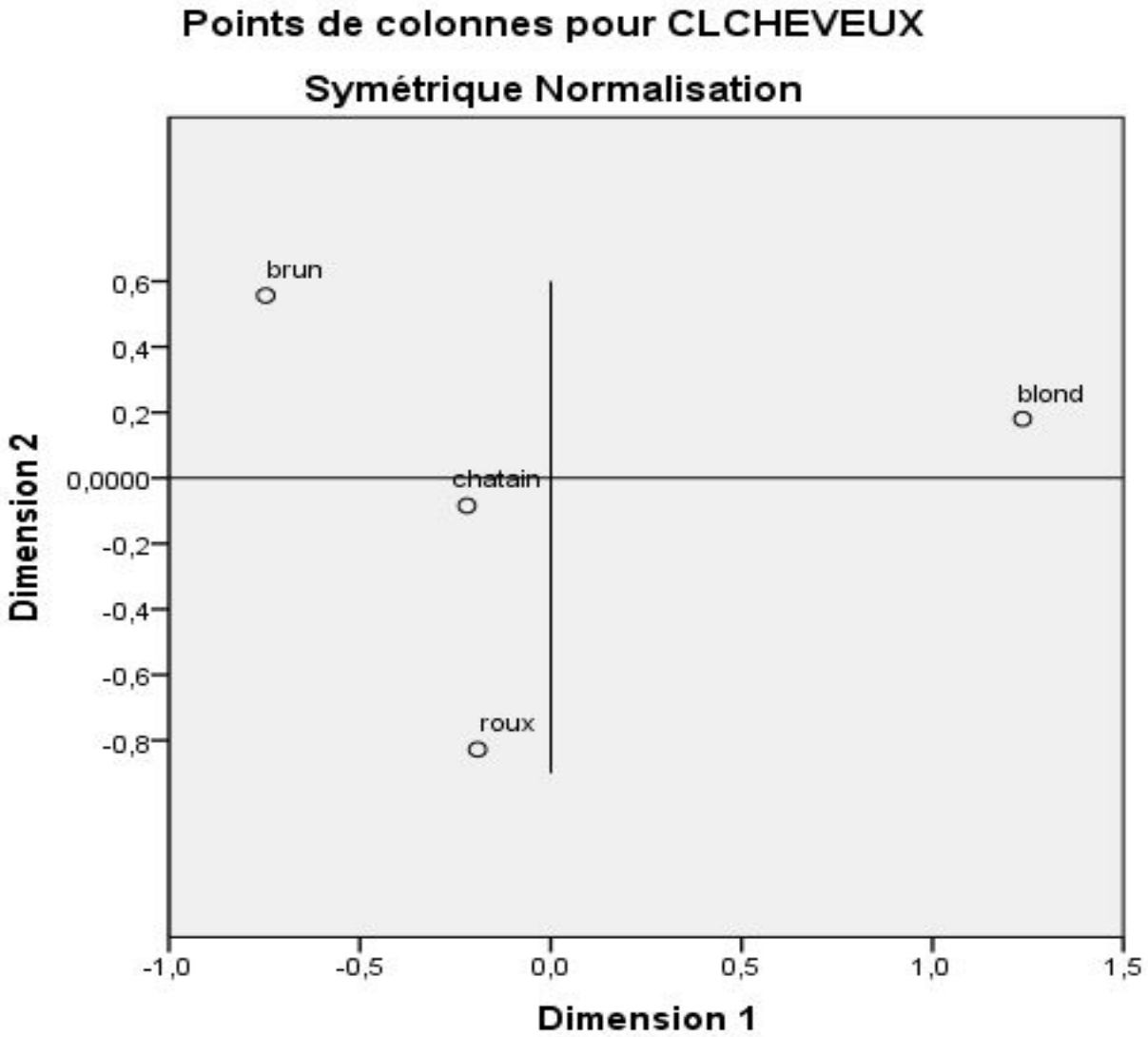
Profils-colonnes I

Sur le premier axe, la couleur des cheveux "blond" (contrib. = 71, 7%, cosinus carré = 0.99) s'oppose à toutes les autres, mais surtout à "brun". Le point "roux" a une contrib. très faible sur le premier axe (1%).
Le second axe est construit par la couleur "roux" (contributions de 55%) qui s'oppose simultanément à "brun" et "blond". La couleur "roux" est le seul p-c bien représenté sur l'axe 2 (cosinus carré de 0,81).

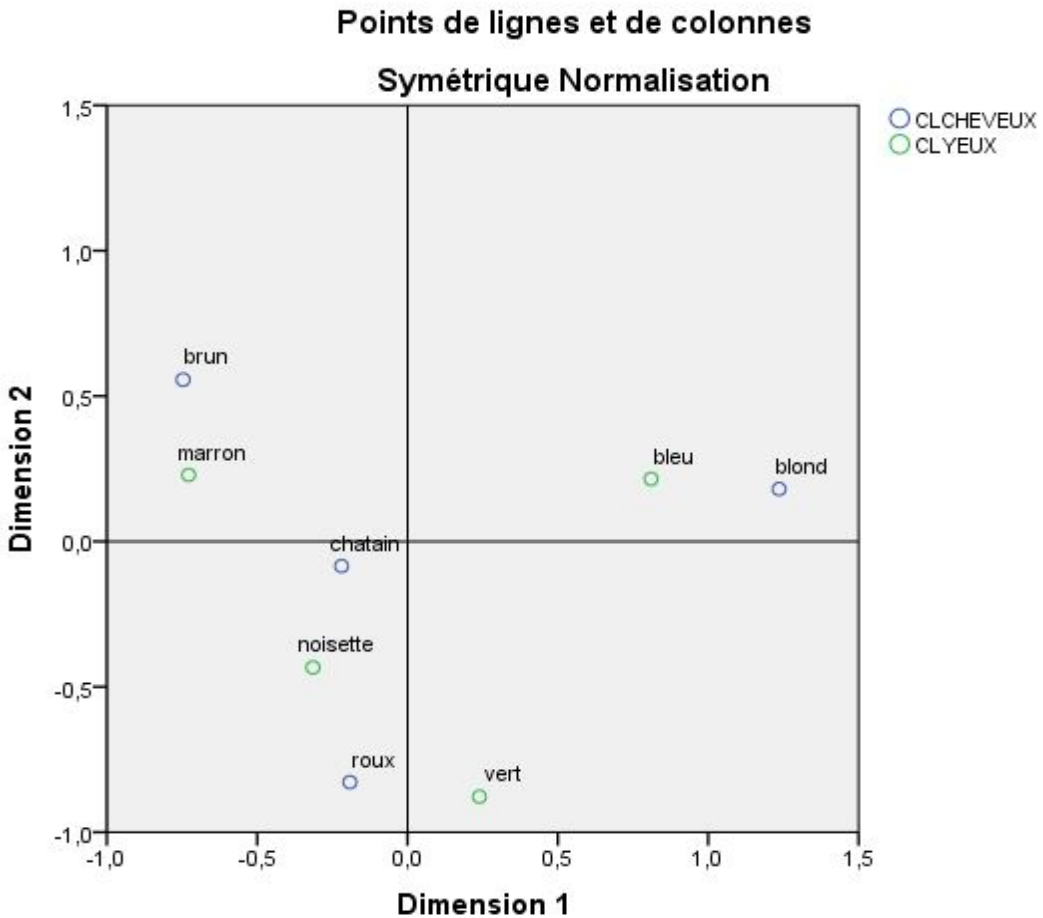
Caractéristiques des points colonnes[a]

CLCHEVEUX	Masse	Score dans la dimension		Inertie	Contribution				
					De point à inertie de dimension		De dimension à inertie de point		
		1	2		1	2	1	2	Total
brun	,182	-,746	,556	,055	,222	,379	,838	,152	,990
chatain	,483	-,219	-,085	,012	,051	,023	,864	,042	,906
roux	,120	-,192	-,828	,015	,010	,551	,133	,812	,945
blond	,215	1,236	,180	,151	,717	,047	,993	,007	1,000
Total actif	1,000			,234	1,000	1,000			

Profils-colonnes II



Représentation simultanée I



Représentation simultanée II

- Quelques associations : yeux bleus et cheveux blonds, cheveux roux et yeux verts, ainsi que cheveux bruns et yeux marrons.
- "ch. blond" est plus excentré que "y. bleu" sur le premier axe car les ch. blonds sont beaucoup mieux carctérisés par les y. bleus que l'inverse : d'après le tab. des p-c, 74% des blonds ont les yeux bleus, alors que d'après le tab. des p-l, 44% des personnes ayant les yeux bleus ont les cheveux blonds.