Chapitre 4 : Analyse Factorielle des Correspondances simples (AFC)

M1 du Master MMAS

James Ledoux

Dépt de mathématiques, Univ. Poitiers

12 juillet 2009

Escofier, B. et Pages, J. (1990).
Analyses factorielles simples et multiples.
Dunod.
Cote BU (519.23 ESC) en 5 exemplaires

Tableau

- **Données : 2 variables qualitatives X et Y**
  X : $I = \{1, \dots, p\}$ p modalités
  Y : $J = \{1, \dots, q\}$ q modalités

Table de contingence

	1	$\cdots$	j	$\cdots$	q	
1			$\vdots$			
$\vdots$			$\vdots$			
i	$\cdots$	$\cdots$	n_{ij}	$\cdots$	$\cdots$	$n_{i.}$
$\vdots$			$\vdots$			
p			$\vdots$			
			$n_{.j}$			n

$$n_{i.} = \sum_{j \in J} n_{ij} \qquad n_{.j} = \sum_{i \in I} n_{ij} \qquad n = \sum_{i \in I} \sum_{j \in J} n_{ij}$$

X et Y indépendantes ?

		Couleur des cheveux				
		brune	châtain	roux	blond	Effectifs
Couleur des yeux	marron	68	119	26	7	$n_{1.} = 220$
	noisette	15	54	14	10	$n_{2.} = 93$
	vert	5	29	14	16	$n_{3.} = 64$
	bleu	20	84	17	94	$n_{4.} = 215$
	Effectifs	$n_{.1} = 108$	$n_{.2} = 286$	$n_{.3} = 71$	$n_{.4} = 127$	$n = 592$

Table 14: Table de contingence : ventilation d'une population de 592 femmes suivant leurs couleurs des yeux et des cheveux

- **Tableau F des fréquences relatives**

$$F := \begin{array}{|c|c|c|} \hline & 1 \quad \cdots \quad q & \text{Marge col.} \\ \hline 1 \\ \vdots \\ p & f_{ij} = \frac{n_{ij}}{n} & f_{i.} \\ \hline \text{Marge lig.} & f_{.j} & 1 \\ \hline \end{array}$$

avec $f_{.j} = \sum_{i \in I} f_{ij} = \frac{n_{.j}}{n}$ $\qquad f_{i.} = \sum_{j \in J} f_{ij} = \frac{n_{i.}}{n}$

Interprétation : tableau $F := (f_{ij})_{i,j}$ d'une loi de probabilité conjointe sur l'ensemble produit $I \times J$

Marge colonne $\equiv (f_{i.})_{i \in I}$: loi marginale de la variable X

$$p \times p \quad D_I := \mathrm{dia}(f_{i.})$$

Marge ligne $\equiv (f_{.j})_{j \in J}$: loi marginale de la variable Y

$$q \times q \quad D_J := \mathrm{dia}(f_{.j})$$

- **Tableau Z_I des profils-lignes**

$Z_I := D_I^{-1} F =$

Lois conditionnelles $(Y \mid X = i)$						
	1	$\cdots$	j	$\cdots$	q	Total
i	$\frac{f_{i1}}{f_{i.}}$	$\cdots$	$\frac{f_{ij}}{f_{i.}}$	$\cdots$	$\frac{f_{iq}}{f_{i.}}$	1
Profil moyen	$f_{.1}$	$\cdots$	$f_{.j}$	$\cdots$	$f_{.q}$	1

- **Tableau Z_J des profils-colonnes**

$Z_J := F D_J^{-1} =$

Lois conditionnelles $(X \mid Y = j)$		
	j	Profil moyen
1	$\frac{f_{1j}}{f_{.j}}$	$f_{1.}$
$\vdots$	$\vdots$	$\vdots$
i	$\frac{f_{ij}}{f_{.j}}$	$f_{i.}$
$\vdots$	$\vdots$	$\vdots$
p	$\frac{f_{pj}}{f_{.j}}$	$f_{p.}$
Total	1	1

		Couleur des cheveux				
		brune	châtain	roux	blond	Total
Couleur des yeux	marron	0.31	0.54	0.12	0.03	1
	noisette	0.16	0.58	0.15	0.11	1
	vert	0.08	0.45	0.22	0.25	1
	bleu	0.09	0.39	0.08	0.44	1
Profil moyen ou marge		0.18	0.48	0.12	0.22	1

		Couleur des cheveux				Profil moyen
		brune	châtain	roux	blond	ou marge
Couleur des yeux	marron	0.63	0.42	0.37	0.06	0.37
	noisette	0.14	0.19	0.20	0.08	0.16
	vert	0.05	0.10	0.20	0.13	0.11
	bleu	0.19	0.29	0.24	0.74	0.36
	Total	1	1	1	1	1

TABLE 15: Tables des profils-ligne et profils-colonne

- **Rappel : indépendance de X et Y :**

$$\boxed{\forall (i,j) \in I \times J \qquad f_{ij} = f_{i.} f_{.j}}$$

$\iff$ **tous les profils-lignes sont égaux**
$\iff$ **tous les profils-colonnes sont égaux**

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

- **Comparer la table de contingence observée à une table de contingence $(s_{ij})_{i \in I, j \in J}$ construite sous l'hypothèse d'indépendance** avec

$$s_{ij} := \frac{n_{i.} n_{.j}}{n}$$

Une mesure de l'écart à l'indépendance

(27) $$\chi^2 = \sum_{i \subset I} \sum_{j \subset J} \frac{(n_{ij} - s_{ij})^2}{s_{ij}}$$

- **Test d'indépendance du chi-deux**

$$(H_0 : \text{ indépendance de } X \text{ et } Y) \Longrightarrow \chi^2 \approx \chi^2_{(p-1)(q-1)}$$

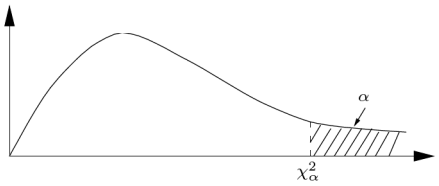


FIGURE 27: Densité du $\chi^2_{(p-1)(q-1)}$

Rejet de H_0 si $\chi^2_{obs} > \chi^2_\alpha$ où χ^2_α est la valeur seuil telle que

$$\mathbb{P}\left(\chi^2_{(p-1)(q-1)} > \chi^2_\alpha\right) = \alpha$$

- Notons que χ^2_α représente le quantile $Q(1-\alpha)$ de la loi $\chi^2_{(p-1)(q-1)}$
- α est traditionnellement appelé le risque de 1re espèce du test et sa valeur est fixée a priori.

La p-valeur

En général, on préfère calculer la p-valeur associée à un test (et c'est ce que propose les logiciels de statistique)

Définition 18 (p-valeur)

La p-valeur associée à un test est la probabilité, sous l'hypothèse H_0, d'observer une valeur de la statistique de test T plus défavorable à H_0 que la valeur observée T_{obs}.

Interprétation

La p-valeur correspond au risque « assumé » en rejetant H_0 sur la base de l'échantillon qui nous donne T_{obs}. Plus la p-valeur est faible plus le risque de se tromper en rejetant H_0 est faible.

Remarque : la p-valeur permet de déterminer la décision prise au vu de T_{obs} pour toute valeur du risque α

pour $\alpha < \mathrm{p}$ alors pas de rejet et pour $\alpha \geq \mathrm{p}$, rejet

Pour la Table 14 : $\chi^2_{obs} \approx 138.29$ (avec 3×3 degrés de liberté) et la p-valeur est $\mathbb{P}(\chi^2_9 > \chi^2_{obs}) < 10^{-3}$.

- **Objectif :**
  Analyser la liaison entre deux variables pour de grands tableaux de profils
- **Méthode :**
  Mesurer les écarts entre les profils-lignes, puis entre les profils-colonnes
- **Choix de la distance :**
  distance euclidienne usuelle

$$\forall i, i' \in I \quad d_2(\,\mathrm{PrL}(i), \mathrm{PrL}(i')\,)^2 = \sum_{j \in J} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Mais favorise les modalités de Y bien représentées dans la population ($f_{.j}$ important)
Rappel : les lignes et colonnes jouent un rôle symétrique. D'où le choix de la distance vaut pour les deux tableaux de profils

Distance dite **du chi-deux** (χ^2)

$$d_{\chi^2}(\,\mathrm{PrL}(i), \mathrm{PrL}(i')\,)^2 = \sum_{j \in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

- **Équivalence distributionnelle**
  Si $\mathrm{PrL}(i) = \mathrm{PrL}(i')$ alors on vérifie que l'on peut agréger les deux modalités i et i' en une nouvelle modalité i''

$$\boxed{(i\,; f_{i.}) + (i'\,; f_{i'.}) \to (i''\,; f_{i.} + f_{i'.})}$$

Cette agrégation ne change rien
- aux distances entre les modalités de cette variable et
- aux distances entre les modalités de l'autre variable

Remarque : même propriété si on subdivise la modalité i'' en plusieurs sous modalités de masse homogène

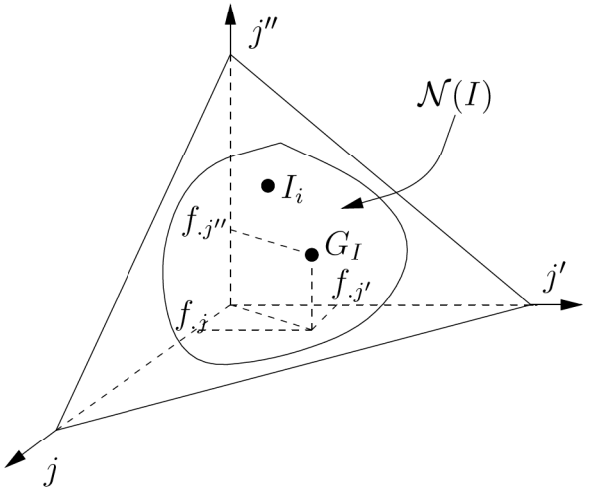
$$\mathcal{N}(I) \subseteq \mathbb{R}^{|J|}$$

- **Modalité i de la variable X :**
  $I_i \equiv \mathrm{PrL}(i) = \left(\frac{f_{ij}}{f_{i.}}\right)_{j\in J} \in \mathbb{R}^{|J|}$ avec

$$\sum_{j\in J} \frac{f_{ij}}{f_{i.}} = 1$$

- **Nuage** $\subseteq \{\ \alpha \in \mathbb{R}^{|J|}\ /\ \alpha \geq 0, \sum_{j\in J} \alpha_j = 1\ \}$ de dim. $(|J|-1)$
- **Le poids associé à I_i : $f_{i.}$**
- **Centre de gravité du nuage** est
  $OG_I = \sum_{i\in I} f_{i.}\, OI_i = (f_{.j})_{j\in J}$ ➧ marge du tableau de contingence

  G_I s'interprète comme une pseudo sous-population présentant une répartition selon les modalités de Y égale à celle de la population totale (profil moyen)



$$d_{\chi^2}(I_i, I_i') = \sum_{j\in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$$ est une distance euclidienne

$$\mathcal{N}(J) \subseteq \mathbb{R}^{|I|}$$

- Le point j est le profil-colonne $\mathrm{PrC}(j)$ de la modalité j de Y :

$$\left(\frac{f_{ij}}{f_{.j}}\right)_{i\in I}$$

- Son poids est $f_{.j}$
- Distance entre deux points = distance du chi-deux

$$d_{\chi^2}(\,\mathrm{PrC}(j), \mathrm{PrC}(j')\,)^2 = \sum_{i\in I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2$$

- $\mathcal{N}(J) \subseteq \{\ \alpha \in \mathbb{R}^{|I|}\ /\ \alpha \geq 0, \sum_{i\in I} \alpha_i = 1\ \}$ de dim. $(|I|-1)$
- G_J = centre de gravité du nuage
  $= (f_{i.})_{i\in I}$
  = marge de la variable X

- **Objectif d'une méthode factorielle :** donner une suite d'images planes approchées du nuage
- **Indépendance X et Y** $\Longleftrightarrow$ tous les profils lignes sont égaux à $G_I = (f_{.j})_{j\in J}$

  ➥ **Mesurer l'écart à l'indépendance** $\equiv$
  évaluer la dispersion du nuage autour de G_I

- **Ajustement des nuages à un sous-espace selon le principe d'inertie (projetée) maximum**

  Principe d'ajustement du nuage $\mathcal{N}(I)$ en AFC
  $\equiv$ principe d'ajustement en ACP non normée pour le « nuage des individus » avec
  - individu $\equiv$ profil-ligne
  - chaque individu a un poids $p_i := f_{i.}$
  - la distance entre les points individus est la distance du chi-deux

- **On centre le nuage des profils-lignes :** si $F := (f_{ij})_{i\in I, j\in J}$ est la table des fréquences relatives et $D_I := \mathrm{diag}(f_{i\cdot})$

$$\begin{array}{rrcl} & i \in I, \quad \mathrm{PrL}(i) & \to & \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j}\right)_{j\in J} \\ \text{(écriture matricielle)} & Z_I = D_I^{-1}F & \to & D_I^{-1}F - \mathbf{1}_{|I|}\mathbf{1}_{|J|}^\top D_J \end{array} \quad \text{➧ nouvelle origine } \equiv G_I$$

Théorème 10 (Axes factoriels du nuage centré des profils-lignes)

*Les **axes factoriels** du nuage de profils-lignes sont définis comme la base de vecteurs propres orthonormés (suivant la norme euclidienne du chi-deux) de la matrice $|J| \times |J|$*

$$F^\top D_I^{-1} F D_J^{-1} - D_J \mathbf{1}_{|J|}\mathbf{1}_{|J|}^\top = {Z_I}^\top Z_J - D_J \mathbf{1}_{|J|}\mathbf{1}_{|J|}^\top \tag{28}$$

où $D_J := \mathrm{diag}(f_{\cdot j})$.
Dans cette base, l'axe de rang $|J|$ est associé à la valeur 0 et est donc exclu de l'analyse.

➧ liste de $|J| - 1$ axes factoriels

Proposition 3 (Équivalence des analyses centrée et non-centrée)

Avec les notations du Théorème 10, on a que les axes factoriels du nuage de profils-lignes s'obtiennent également par diagonalisation de la matrice

$$F^\top D_I^{-1} F D_J^{-1} = {Z_I}^\top Z_J. \tag{29}$$

Cette matrice correspond à une analyse de l'inertie du nuage non-centré.

- **Nuage des profils-colonnes** $\mathcal{N}(J)$: Idem
  Opération de centrage : nouvelle origine $\equiv G_J$

$$\begin{array}{rcl} j \in J, \quad \mathrm{PrC}(j) & \to & \left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot}\right)_{j\in J} \\ F D_J^{-1} & \to & F D_J^{-1} - D_I \mathbf{1}_{|I|}\mathbf{1}_{|J|}^\top \end{array}$$

Théorème 11 (Axes factoriels du nuage des profils-colonnes)

Avec les notations du Théorème 10, les axes factoriels du nuage de profils-colonnes sont donnés par la base orthonormée de vecteurs propres de la matrice $|I| \times |I|$

$$Z_J {Z_I}^\top = F D_J^{-1} F^\top D_I^{-1}. \tag{30}$$

La valeur propre « triviale » 1 ainsi que l'axe correspondant sont exclus de l'analyse.

➧ liste de $|I| - 1$ axes factoriels

On peut alors vérifier les relations suivantes entre les éléments spectraux des matrices (29) et (30) :

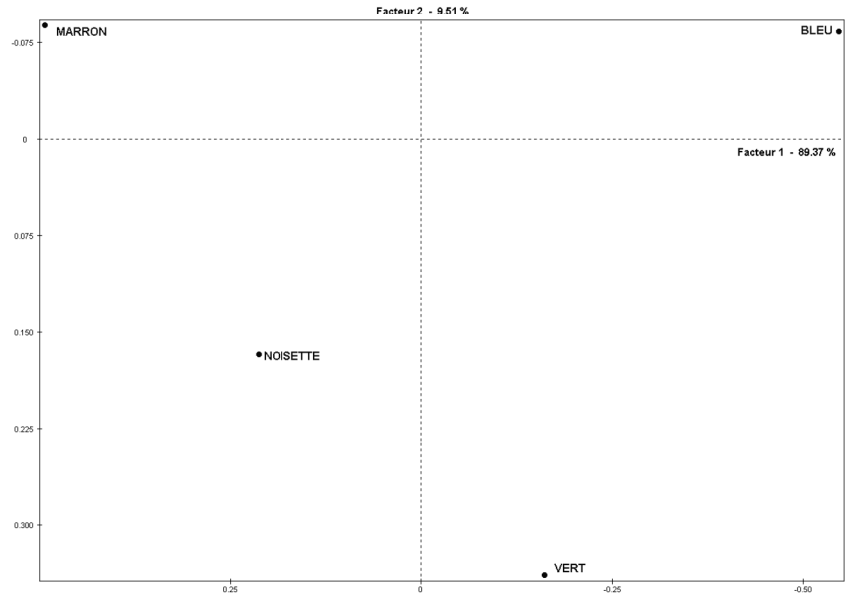
$${Z_I}^\top Z_J = F^\top D_I^{-1} F D_J^{-1} \quad \text{et } Z_J {Z_I}^\top = F D_J^{-1} F^\top D_I^{-1}.$$

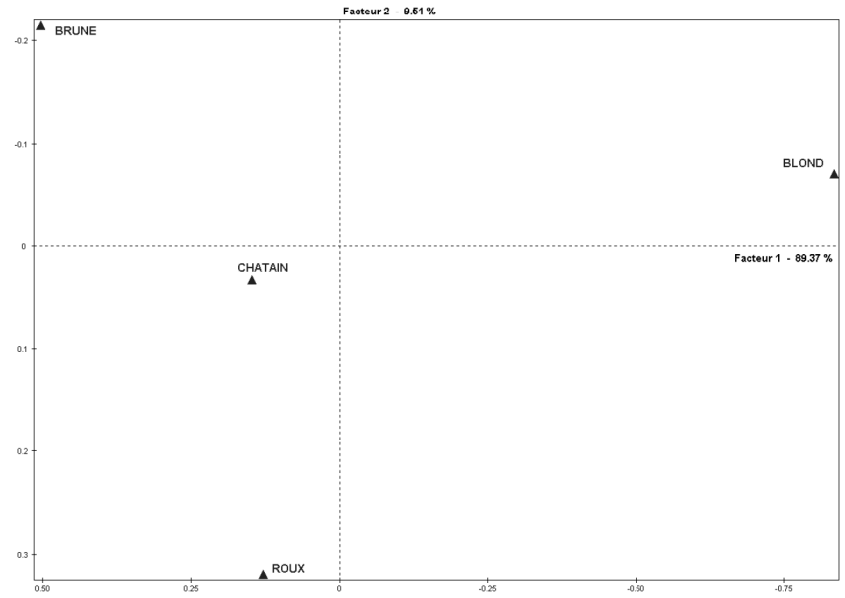
Proposition 4 (Relations spectrales)

Les matrices ${Z_I}^\top Z_J$ et $Z_J {Z_I}^\top$ ont les mêmes valeurs propres non-nulles et les axes factoriels correspondants satisfont

$$v_l = \frac{1}{\sqrt{\lambda_l}} Z_J u_l = \frac{1}{\sqrt{\lambda_l}} F D_J^{-1} u_l \tag{31a}$$

$$u_l = \frac{1}{\sqrt{\lambda_l}} {Z_I}^\top v_l = \frac{1}{\sqrt{\lambda_l}} F^\top D_I^{-1} v_l \tag{31b}$$





Lignes et colonnes représentent des objets de même nature

$$\Longrightarrow \text{ dualité plus riche qu'en ACP}$$

$$\begin{aligned}
\chi^2 &\overset{(27)}{=} \sum_i \sum_j \frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}} \\
&= n \sum_i f_{i.} \left[\sum_{j \in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 \right] \\
&= n \sum_i f_{i.}\, d_{\chi^2}(I_i, G_I)^2 \\
(32) \qquad &= n \ \text{Inertie}(\mathcal{N}(I)) \\
(33) \qquad &= n \ \text{Inertie}(\mathcal{N}(J))
\end{aligned}$$

L'inertie totale a une interprétation statistique

Définition 19 (Facteur d'une AFC)

Un facteur est défini par le vecteur des projections des points du nuage sur un axe factoriel.
Le rang d'un facteur est le rang de l'axe factoriel sur lequel la projection a lieu.

Remarque : facteur en AFC $\equiv$ composante principale en ACP

Relations de transition entre facteurs

- $G_l(i)$
  $=$ projection de la ligne i sur l'axe factoriel de rang l de $\mathcal{N}(I)$
- $F_l(j)$
  $=$ projection de la colonne j sur l'axe factoriel de rang l de $\mathcal{N}(J)$
- $\lambda_l =$ valeur propre commune à chacun des deux axes

La distance du chi-deux est associée au produit scalaire euclidien

$$\begin{aligned}
\mathcal{N}(I) : u, v \in \mathbb{R}^{|J|}, \quad \langle u, v \rangle_{\chi^2} &:= u^\top D_J^{-1} v \\
\mathcal{N}(J) : u, v \in \mathbb{R}^{|I|}, \quad \langle u, v \rangle_{\chi^2} &:= u^\top D_I^{-1} v
\end{aligned}$$

Les facteurs s'obtiennent donc par :

$$\begin{aligned}
G_l &= Z_I D_J^{-1} u_l = D_I^{-1} F D_J^{-1} u_l \\
F_l &= {Z_J}^\top D_I^{-1} v_l = D_J^{-1} F^\top D_I^{-1} v_l
\end{aligned}$$

Les relations suivantes se déduisent de la Proposition 4

Proposition 5 (Relations facteurs-axes)

$$(34a) \qquad G_l = \sqrt{\lambda_l}\, D_I^{-1}\, v_l$$
$$(34b) \qquad F_l = \sqrt{\lambda_l}\, D_J^{-1}\, u_l$$

En combinant les Propositions 4 et 5, on obtient les relations de transition suivantes entres facteurs

Proposition 6 (Relations dites quasi-barycentriques)

$$i \in I, \quad G_l(i) = \frac{1}{\sqrt{\lambda_l}} \sum_{j \in J} \left(\frac{f_{ij}}{f_{i.}} \right) F_l(j) \Longleftrightarrow G_l = \frac{1}{\sqrt{\lambda_l}} Z_I F_l$$

$$j \in J, \quad F_l(j) = \frac{1}{\sqrt{\lambda_l}} \sum_{i \in I} \left(\frac{f_{ij}}{f_{.j}} \right) G_l(i) \Longleftrightarrow F_l = \frac{1}{\sqrt{\lambda_l}} Z_J{}^\top G_l$$

avec

$$\sum_{j \in J} \left(\frac{f_{ij}}{f_{i.}} \right) = 1 \quad \textit{et} \quad \sum_{i \in I} \left(\frac{f_{ij}}{f_{.j}} \right) = 1$$

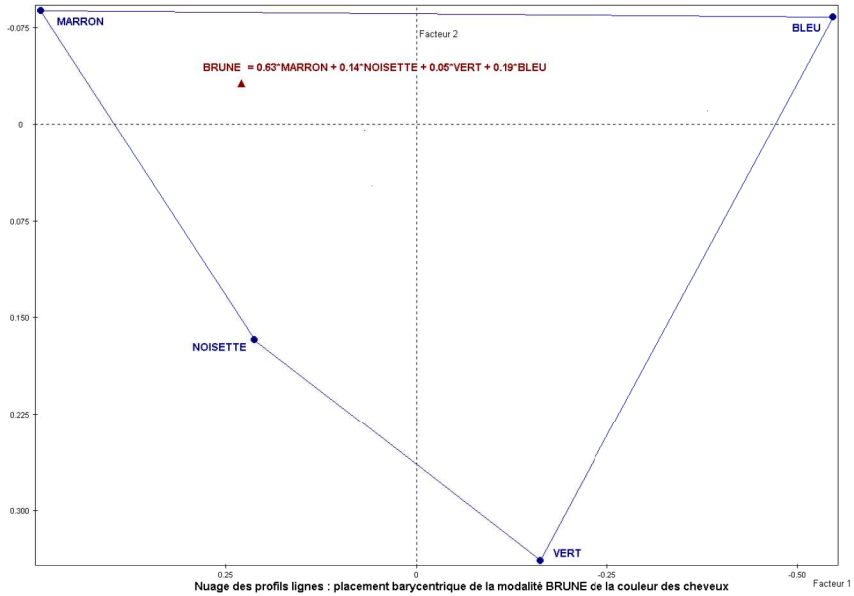
➥ **Représentations barycentriques :** « Superposition » des projections de chacun des deux nuages $\mathcal{N}(I)$ et $\mathcal{N}(J)$ sur les plans factoriel

Colonnes au barycentre des lignes

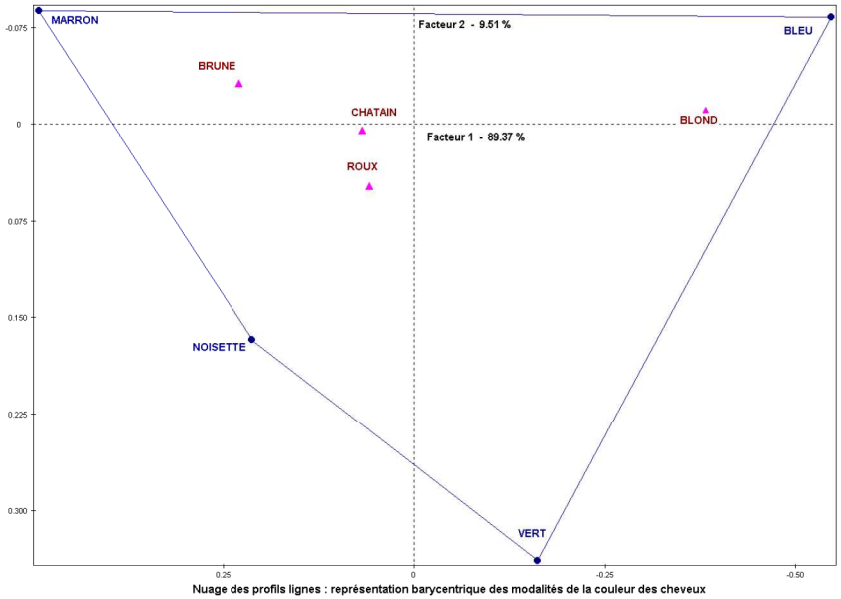
$$F_l(j) = \frac{1}{\sqrt{\lambda_l}} \sum_{i \in I} \left(\frac{f_{ij}}{f_{.j}} \right) G_l(i) \quad \text{avec} \sum_{i \in I} \left(\frac{f_{ij}}{f_{.j}} \right) = 1$$

- Au facteur $1/\sqrt{\lambda_l}$ près, la projection de la colonne j sur les axes est le barycentre des projections des points lignes, chaque ligne i étant affectée du poids $(f_{ij}/f_{.j})$, c'est à dire de la fréquence d'observation de la modalité i dans la population présentant la modalité j.
- Les modalités i « lourdes » attirent le barycentre
  ➥ une ligne i attire d'autant plus une colonne j que la valeur de f_{ij} est élevée

Lignes au barycentre des colonnes : idem en échangeant les rôles de F_l et G_l



Nuage des profils lignes : placement barycentrique de la modalité BRUNE de la couleur des cheveux

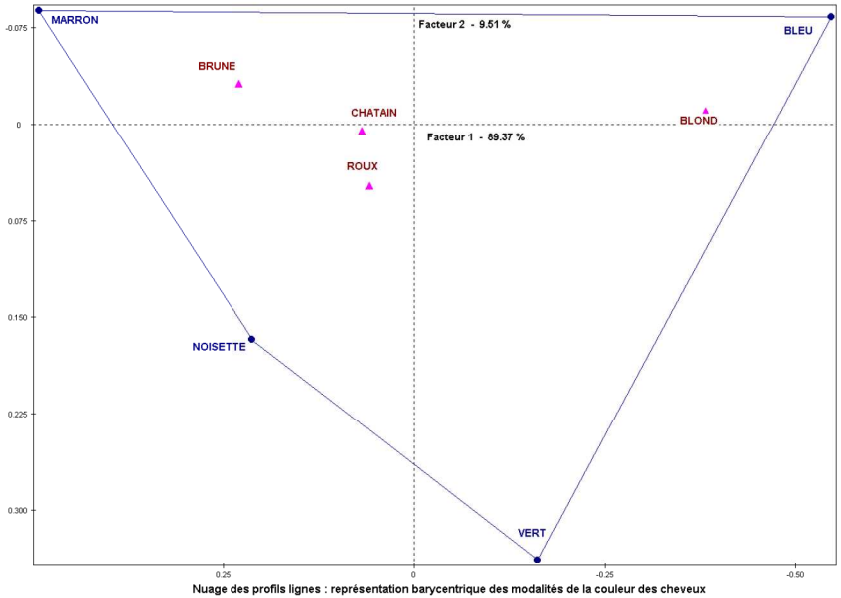


Nuage des profils lignes : représentation barycentrique des modalités de la couleur des cheveux

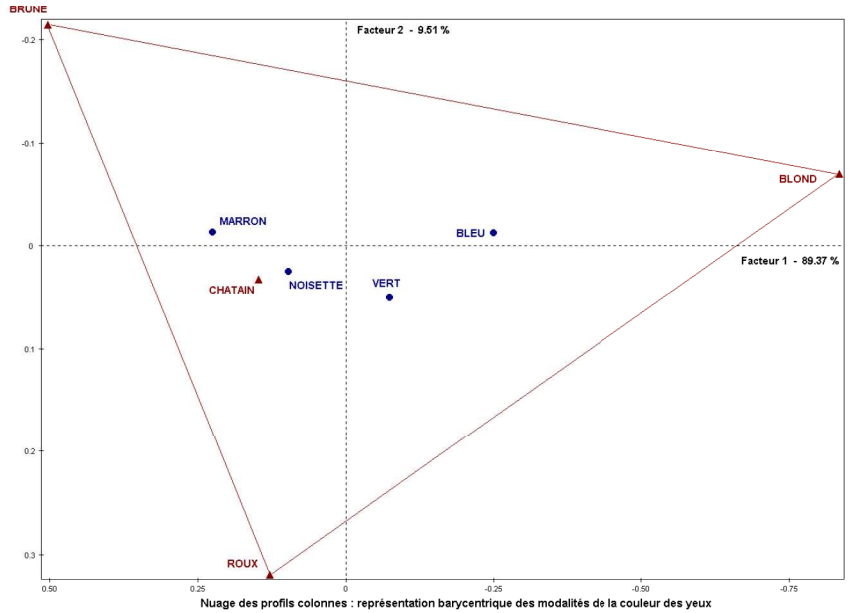
Règles de base pour la lecture des graphiques de représentation barycentrique

- La position d'un point d'un nuage s'interprète
  - comme une distance / aux autres points du **même** nuage
  - comme barycentre / à **tous** les points de l'autre nuage
- Sur chaque axe, on trouve
  du même côté d'une ligne i, les colonnes j auxquelles elle s'associe le plus
  et à l'opposé celles auxquelles elle s'associe le moins

  Toute association entre une ligne et une colonne suggérée par le graphique doit être contrôlée sur le tableau de données



Nuage des profils lignes : représentation barycentrique des modalités de la couleur des cheveux



Nuage des profils colonnes : représentation barycentrique des modalités de la couleur des yeux

Les indices d'aide à l'interprétation définis en ACP sont valides avec les différences suivantes

- **Inertie totale du nuage**

$$\begin{aligned} \text{Inertie}(\mathcal{N}(I)) &= \text{Inertie}(\mathcal{N}(J)) \\ &= \sum_{j\in J}\sum_{i\in I} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \\ &= \sum_{l=1}^{p} \lambda_l \end{aligned}$$

$$\boxed{\chi^2 = n \times \text{Inertie}(\mathcal{N}(J))} \quad \text{cf (32,33)}$$

- **Calcul de la p-valeur** pour tester l'hypothèse H_0 d'indépendance des deux variables

$$p = \mathbb{P}\left(\chi^2_{(p-1)(q-1)} > \chi^2_{obs}\right)$$

Plus α diminue plus H_0 est douteuse

- Inertie $= \sum_{l=1}^{p} \lambda_l$ est un indicateur de dispersion autour de G
Rappel : X et Y indépendantes si les profils sont identiques

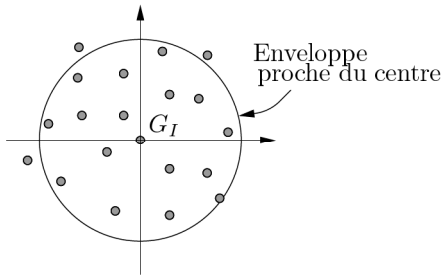
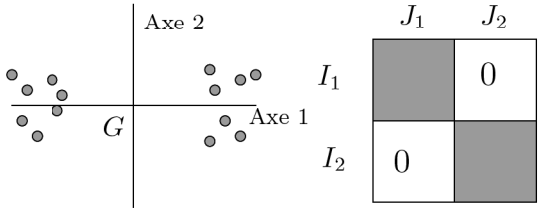


FIGURE 28: inertie totale faible et pas de direction privilégiée

- **Les valeurs propres sont ≤ 1**
  - Si une valeur propre ≈ 1 alors une réorganisation du tableau de données



  - Si deux valeur propre $\approx 1 \Longrightarrow$ 3 sous-nuages et 3 groupes de modalités
  - Si toutes les valeurs propres sont ≈ 1 chaque modalité d'une variable est en correspondance presque exclusive avec une seule modalité de l'autre

Remarque : Analyser chaque sous-nuage séparément

- **Effet Guttman**

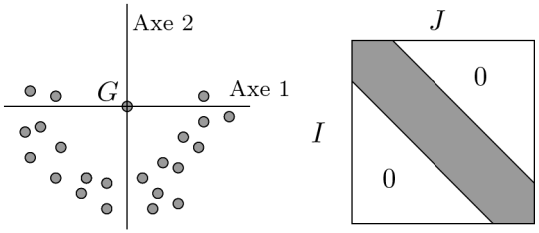


FIGURE 29: Nuage de points parabolique

Redondance des deux variables :
de la connaissance de la ligne i, on en déduit la colonne j
Toute info dans l'axe 1
Remarque : visible surtout pour les variables ordinales

- **Les point intéressants sur les plans factoriels sont toujours des points éloignés de l'origine** ($\equiv$ centre de gravité) car les plus différents du profil moyen
- **Dans le cas général, les points ont des poids différents en AFC**
  - les coordonnées d'un point sur un axe factoriel
  - sa qualité de représentation
  - sa contribution à l'inertie

sont des informations différentes

Pour l'analyse d'un axe, on s'appuie sur les points présentant

- une forte contribution et une forte qualité de représentation : points explicatifs de l'axe
- une coordonnée extrême jointe à une forte qualité de représentation : points expliqués par l'axe
  ➥ ces points sont très différents du profil moyen et cette différence est presque entièrement traduite par l'axe
- une coordonnée « extrême » jointe à une qualité moyenne de représentation : points partiellement expliqués par l'axe
  ➥ ces points représentent à un fort niveau les caractéristiques du facteur mais ces caractéristiques s'additionnent à d'autres

Remarque : cercle des corrélations avec des variables qualitatives n'a aucun sens

	CONTRIB		COS CARRES		COORD	
YEUX	Axe1	Axe2	Axe1	Axe2	Axe1	Axe2
Marron	43	13	0,97	0,03	-0,49	0,09
Noisette	3	20	0,54	0,34	-0,21	-0,17
Vert	1	56	0,18	0,77	0,16	-0,34
Bleu	52	11	0,98	0,02	0,55	0,08

	CONTRIB		COS CARRES		COORD	
CHEVEUX	Axe1	Axe2	Axe1	Axe2	Axe1	Axe2
Brun	22	38	0,84	0,15	-0,50	0,21
Chatain	5	2	0,86	0,04	-0,15	-0,03
Roux	1	55	0,13	0,81	-0,13	-0,32
Blond	72	5	0,99	0,01	0,84	0,07

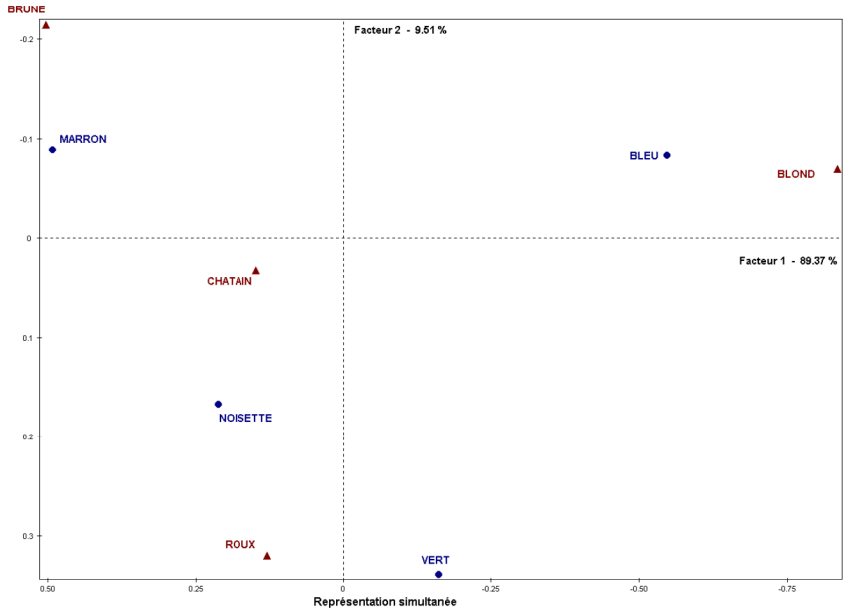
TABLE 16: Indicateurs pour les deux nuages

1. **Représentations barycentriques**
   Inutiles si les poids affectés aux profils sont du même ordre de grandeur
2. **Représentation simultanée**
   Cette représentation simultanée des deux nuages reposent sur le placement des colonnes sur le graphique des profils-lignes mais selon la formule quasi-barycentrique complète

$$F_l(j) = \boxed{\frac{1}{\sqrt{\lambda_l}}} \sum_{i \in I} \left(\frac{f_{ij}}{f_{.j}} \right) G_l(i)$$

Règles de lecture identiques / Représentations barycentriques

- **Éléments supplémentaires :** les relations quasi-barycentriques permettent de placer les modalités de variables qualitatives supplémentaires sur les graphiques



Bilan de l'AFC pour la Table 14

- **Nuage des profils-lignes**
  - Axe 1 : Opposition modalités Marron / Bleu
  - Axe 2 : Modalité Vert (Noisette)
- **Nuage des profils-colonnes**
  - Axe 1 : Opposition modalités Brun / Blond
  - Axe 2 : Modalité Roux
- **Associations**
  1. Yeux bleus $\longleftrightarrow$ Cheveux blonds (noter la non-symétrie)
  2. Yeux marrons $\longleftrightarrow$ Cheveux bruns
  3. Yeux verts (noisettes) $\longleftrightarrow$ Cheveux roux

AFC et approximation du tableau de données

- Une décomposition en valeurs singulières relativement à $\langle \cdot, \cdot \rangle_{\chi^2}$

$$\left.\begin{array}{l} u_l \text{ axe factoriel de } \mathcal{N}(I) \\ v_l \text{ axe factoriel de } \mathcal{N}(J) \end{array}\right] \rightarrow \lambda_l$$

La Prop. 4 : $Z_J u_l = F D_J^{-1} u_l = \sqrt{\lambda_l} v_l$

$$\Longrightarrow F D_J^{-1} u_l u_l^\top = \sqrt{\lambda_l} v_l u_l^\top$$

$$\Longrightarrow F \left[\sum_{l=1}^{q} D_J^{-1} u_l u_l^\top\right] = \sum_{l=1}^{q} \sqrt{\lambda_l} v_l u_l^\top$$

Comme $\{u_l, l = 1, \ldots |J|\}$ est $\langle \cdot, \cdot \rangle_{\chi^2}$-orthonormale alors

$\sum_{l=1}^{q} D_J^{-1} u_l u_l^\top = I$ et $\boxed{F = \sum_{l-1}^{q} \sqrt{\lambda_l}\ v_l u_l^\top.}$

La matrice F admet une formule de reconstruction en termes des facteurs (voir Prop. 5) :

$$\boxed{F = \sum_{l=1}^{q} \frac{1}{\sqrt{\lambda_l}}\ D_I G_l F_l^\top D_J.}$$

On en déduit la formule suivante (en exploitant également le fait que la première valeur propre est 1) :

Théorème 12 (Écart à l'indépendance)

$$\forall (i,j) \in I \times J, \quad \boxed{f_{ij} = f_{i.} f_{.j} \left(1 + \sum_{l=2}^{q} \frac{1}{\sqrt{\lambda_l}} G_l(i) F_l(j)\right)}$$

SAS

L'analyse des correspondances (simples ou multiples) est réalisée par la procédure **CORRESP** de **SAS/STAT**