Analyse Factorielle des Correspondances

AFC

AUTEUR·RICE
Benjamin Charlier

1 Tableaux de contingence bivariés

1.1 Tableau de contingence

Un tableau de contingence est une matrice $K \in \mathbb{R}^{q \times m}$ obtenue en croisant deux variables qualitatives X_1 et X_2, ayant respectivement q et m modalités, observées sur un échantillon de n individus.

▸ **Tableau**

	1	$\cdots$	s	$\cdots$	m	**Total**
1	n_{11}	$\cdots$	n_{1s}	$\cdots$	n_{1m}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	n_{i1}	$\cdots$	n_{is}	$\cdots$	n_{im}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
q	n_{q1}	$\cdots$	n_{qs}	$\cdots$	n_{qm}	$n_{q.}$
Total	$n_{.1}$	$\cdots$	$n_{.s}$	$\cdots$	$n_{.m}$	$n_{..} = n$

On note :

- n_{is} le nombre d'individus possédant les modalités i et s ;
- $n_{i.} = \sum_{s=1}^{m} n_{is}$ le nombre d'individus possédant la modalité i ;
- $n_{.s} = \sum_{i=1}^{q} n_{is}$ le nombre d'individus possédant la modalité s ;
- $n_{..} = \sum_{i=1}^{q} \sum_{s=1}^{m} n_{is} = n$ le nombre total d'individus.

1.2 Tableau des fréquences

En divisant par n le tableau K, on a le tableau des fréquences $F = (f_{is})_{1 \leq i \leq q,\, 1 \leq s \leq m}$:

▸ **Tableau**

	1	$\cdots$	s	$\cdots$	m	**Total**
1	f_{11}	$\cdots$	f_{1s}	$\cdots$	f_{1m}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	f_{i1}	$\cdots$	f_{is}	$\cdots$	f_{im}	$f_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
q	f_{q1}	$\cdots$	f_{qs}	$\cdots$	f_{qm}	$f_{q\cdot}$
Total	$f_{\cdot 1}$	$\cdots$	$f_{\cdot s}$	$\cdots$	$f_{\cdot m}$	$f_{\cdot\cdot}=1$

- Les fréquences et leurs marginales sont $f_{is}=\frac{n_{is}}{n}$, $f_{i\cdot}=\sum_{s=1}^{m} f_{is}=\frac{n_{i\cdot}}{n}$, $f_{\cdot s}=\sum_{i=1}^{q} f_{is}=\frac{n_{\cdot s}}{n}$.
- On note ensuite les vecteurs des poids des lignes et des colonnes :

$$r=(f_{1\cdot},\dots,f_{i\cdot},\dots,f_{q\cdot})^{\top},\qquad c=(f_{\cdot 1},\dots,f_{\cdot s},\dots,f_{\cdot m})^{\top}.$$

Les matrices diagonales associées sont $D_r=\mathbf{diag}(r)\in\mathbb{R}^{q\times q}$ et $D_c=\mathbf{diag}(c)\in\mathbb{R}^{m\times m}$.

1.3 Tableau des profils lignes (conditionnelle en ligne)

Le tableau des profils lignes L est obtenu en divisant chaque ligne i de F par son poids $f_{i\cdot}$.

▸ **Tableau**

	1	$\cdots$	s	$\cdots$	m
1	$f_{11}/f_{1\cdot}$	$\cdots$	$f_{1s}/f_{1\cdot}$	$\cdots$	$f_{1m}/f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	$f_{i1}/f_{i\cdot}$	$\cdots$	$f_{is}/f_{i\cdot}$	$\cdots$	$f_{im}/f_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
q	$f_{q1}/f_{q\cdot}$	$\cdots$	$f_{qs}/f_{q\cdot}$	$\cdots$	$f_{qm}/f_{q\cdot}$

On note $\ell_{is}=\frac{f_{is}}{f_{i\cdot}}=\frac{n_{is}}{n_{i\cdot}}$ le terme général de $L=D_r^{-1}F$.

- **Métrique des poids :** les profils lignes forment un nuage de q vecteurs de $\mathbb{R}^m$. Le profil ligne moyen pondérés par $D_r\in\mathbb{R}^{q\times q}$ (centre de gravité du nuage) est $c=\sum_{i=1}^{q} f_{i\cdot}\,L_{i:}\in\mathbb{R}^m$ (marges des colonnes). On peut alors centrer la matrice des profils lignes[1],

$$L^c=\left(\frac{f_{is}}{f_{i\cdot}}-c_s\right)_{i,s}=\left(\frac{f_{is}}{f_{i\cdot}}-f_{\cdot s}\right)_{i,s}=\left(\frac{f_{is}-f_{i\cdot}f_{\cdot s}}{f_{i\cdot}}\right)_{i,s}=D_r^{-1}\left(F-rc^{\top}\right).$$

- **Distance du χ^2:** métrique euclidienne avec poids $D_c^{-1}\in\mathbb{R}^{m\times m}$

$$d^2_{D_c^{-1}}(L_{i:},L_{i':})=(L_{i:}-L_{i':})^{\top}D_c^{-1}(L_{i:}-L_{i':})=\sum_{s=1}^{m}\frac{1}{f_{\cdot s}}\left(\frac{f_{is}}{f_{i\cdot}}-\frac{f_{i's}}{f_{i'\cdot}}\right)^2.$$

[1] Centrer dans cette métrique $\sum_i f_{i\cdot}\,L^c_{i:}=0_{\mathbb{R}^m}$ revient à mesurer l'écart à l'hypothèse d'indépendance stochastique ($f_{is}=f_{i\cdot}f_{\cdot s}$).

1.4 Tableau des profils colonnes (conditionnelle en colonnes)

Le tableau des profils colonnes C est obtenu en divisant chaque colonne s de F par son poids $f_{\cdot s}$.

▸ **Tableau**

	1	$\cdots$	s	$\cdots$	m
1	$f_{11}/f_{\cdot 1}$	$\cdots$	$f_{1s}/f_{\cdot s}$	$\cdots$	$f_{1m}/f_{\cdot m}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	$f_{i1}/f_{\cdot 1}$	$\cdots$	$f_{is}/f_{\cdot s}$	$\cdots$	$f_{im}/f_{\cdot m}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
q	$f_{q1}/f_{\cdot 1}$	$\cdots$	$f_{qs}/f_{\cdot s}$	$\cdots$	$f_{qm}/f_{\cdot m}$
Total	1	$\cdots$	1	$\cdots$	1

On note $c_{is} = \frac{f_{is}}{f_{\cdot s}} = \frac{n_{is}}{n_{\cdot s}}$ le terme général de $C = FD_c^{-1}$.

- **Métrique des poids :** les profils colonnes forment un nuage de m vecteurs de $\mathbb{R}^q$. Le profil colonne moyen pondérés par $D_c \in \mathbb{R}^{m\times m}$ (centre de gravité du nuage) est $r = \sum_{s=1}^{m} f_{\cdot s}\, C_{:s} \in \mathbb{R}^q$ (marges des lignes). On peut alors centrer la matrice des profils colonnes[2]:

$$C^c = \left(\frac{f_{is}}{f_{\cdot s}} - r_i\right)_{i,s} = \left(\frac{f_{is}}{f_{\cdot s}} - f_{i\cdot}\right)_{i,s} = \left(\frac{f_{is} - f_{i\cdot} f_{\cdot s}}{f_{\cdot s}}\right)_{i,s} = \left(F - rc^\top\right) D_c^{-1}.$$

- **Distance du χ^2:** métrique euclidienne avec poids $D_r^{-1} \in \mathbb{R}^{q\times q}$.

$$d^2_{D_r^{-1}}(C_{:s}, C_{:s'}) = (C_{:s} - C_{:s'})^\top D_r^{-1} (C_{:s} - C_{:s'}) = \sum_{i=1}^{q} \frac{1}{f_{i\cdot}} \left(\frac{f_{is}}{f_{\cdot s}} - \frac{f_{is'}}{f_{\cdot s'}}\right)^2.$$

[2] Centrer dans cette métrique $\sum_s f_{\cdot s}\, C^c_{:s} = 0_{\mathbb{R}^q}$ revient à mesurer l'écart à l'hypothèse d'indépendance stochastique ($f_{is} = f_{i\cdot} f_{\cdot s}$).

1.5 Liaison entre lignes et colonnes

▶ **Taux de liaison**

En cas d'indépendance entre X_1 et X_2, on a :

$$f_{is} = f_{i\cdot} f_{\cdot s} \iff n_{is} = \frac{n_{i\cdot} n_{\cdot s}}{n}.$$

avec les fréquences théoriques $f_{i\cdot} f_{\cdot s}$ et les effectifs théoriques $\frac{n_{i\cdot} n_{\cdot s}}{n}$.

1.6 Inertie et liaison entre variables

Une mesure globale de liaison entre X_1 et X_2 est

$$\chi^2 = \sum_{i=1}^{q}\sum_{s=1}^{m} \frac{\left(n_{is} - \frac{n_{i\cdot} n_{\cdot s}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot s}}{n}} = n \underbrace{\sum_{i=1}^{q}\sum_{s=1}^{m} \frac{(f_{is} - f_{i\cdot} f_{\cdot s})^2}{f_{i\cdot} f_{\cdot s}}}_{=\Phi^2}$$

- On définit une mesure locale de liaison entre une modalité i de X_1 et une modalité s de X_2 :
  - si $f_{is} \simeq f_{i\cdot} f_{\cdot s}$, il y a indépendance entre i et s ;
  - si $f_{is} > f_{i\cdot} f_{\cdot s}$, les modalités i et s s'attirent, car $\frac{f_{is}}{f_{i\cdot}} > f_{\cdot s}$ et $\frac{f_{is}}{f_{\cdot s}} > f_{i\cdot}$.
  - si $f_{is} < f_{i\cdot} f_{\cdot s}$, les modalités i et s se repoussent, car $\frac{f_{is}}{f_{i\cdot}} < f_{\cdot s}$ et $\frac{f_{is}}{f_{\cdot s}} < f_{i\cdot}$.
- En pondérant par les $f_{i\cdot} f_{\cdot s}$, on obtient $z_{is} = \frac{f_{is} - f_{i\cdot} f_{\cdot s}}{f_{i\cdot} f_{\cdot s}}$, d'où $\chi^2 = n \sum_{i,s} f_{i\cdot} f_{\cdot s} z_{is}^2$, et $\sum_{i,s} f_{i\cdot} f_{\cdot s} z_{is}^2 = \frac{\chi^2}{n} = \Phi^2$
- On a la matrice des écarts à l'indépendance (mesuré avec le taux de liaison):

$$Z = \left(\frac{f_{is} - f_{i\cdot} f_{\cdot s}}{f_{i\cdot} f_{\cdot s}}\right)_{is} = D_r^{-1}(F - rc^\top)D_c^{-1}$$

Théorème

On a

$$\frac{1}{n}\chi^2 = I(L) = I(C),$$

où $I(L)$ est l'inertie du nuage des profils lignes et $I(C)$ l'inertie du nuage des profils colonnes, calculées avec leur métrique et distance du χ^2 respective[3].

[3] On a donc $I(L) = \sum_s f_{i\cdot} d^2_{D_c^{-1}}(L_{i:}, c^\top)$ et $I(C) = \sum_i f_{\cdot s} d^2_{D_r^{-1}}(C_{:s}, c^\top)$

1.7 Exemple

```
> # Statistique du chi2 calculée directement
> chisq.test(K)$statistic
```

```
X-squared
 106.6637
```

```
> # Carré de la distance du chi2 entre un profil ligne et le profil moyen
> distsq <- function(x, mean=c) {
+   sum((x - mean)^2 / mean)
+ }
>
> # Carrés des distances de tous les profils lignes au profil moyen
> dists.row <- apply(L, 1, distsq, mean = c)
> dists.row
```

```
     Black      Brown        Red      Blond
0.38974635 0.07834092 0.18296344 0.84474722
```

```
> # Inertie du nuage des profils lignes : moyenne ponderée des distances
> iner.row <- sum(dists.row * r)
> iner.row * sum(K)
```

```
[1] 106.6637
```

```
> # Carrés des distances de tous les profils colonnes au profil moyen
> dists.col <- apply(C, 2, distsq, mean = r)
> dists.col
```

```
    Brown      Blue     Hazel     Green
0.3143532 0.4820014 0.1826037 0.1601520
```

```
> # Inertie du nuage des profils colonnes : moyenne ponderée des distances
> iner.col <- sum(dists.col * c)
> iner.col * sum(K)
```

```
[1] 106.6637
```

2 Analyse Factorielle des Correspondances

2.1 Principes

Décrire simultanément les deux nuages pondérés suivants :

1. les q profils lignes (lignes de L) pondérés par D_r et avec la métrique D_c^{-1} ;
2. les m profils colonnes, (colonnes de C) pondérés par D_c et avec la métrique D_r^{-1}.

Elle revient de manière équivalente à effectuer :

- deux ACP pondérées successives sur les q profils lignes, puis sur les m profils colonnes
- ou une ACP pondérée du tableau des écarts à l'indépendance $Z = D_r^{-1}(F - rc^\top)D_c^{-1}$ avec les pondérations D_r pour les lignes et D_c pour les colonnes.

2.2 ACP de la matrice des profils-lignes centrés

On cherche à représenter au mieux les q modalités de X_1 dans un sous-espace de $\mathbb{R}^m$ de dimension $k \leq \min(q-1, m-1)$[4].

[4] On a au plus $q-1$ dimensions pour les lignes (du fait du centrage par les poids) et $m-1$ pour les colonnes (car la somme d'un profil vaut 1).

- Les axes sont engendrés par les vecteurs propres ordonnés $v_1, \ldots, v_k$ de $L^\top D_r L D_c^{-1} \in \mathbb{R}^{m \times m}$ associés aux valeurs propres $\lambda_1 \geq \cdots \geq \lambda_k$
- À chaque étape, on ajoute une nouvelle direction qui maximise la variance (pondérée par D_r) des projections des profils-lignes centrés, sous la métrique D_c^{-1}
- Les vecteurs directeurs v_α sont D_c^{-1} orthonormés :

$$v_\alpha^\top D_c^{-1} v_\beta = \begin{cases} 1, & \text{si } \alpha = \beta, \\ 0, & \text{si } \alpha \neq \beta, \end{cases} \qquad \alpha, \beta = 1, \ldots, k$$

- La matrice $q \times k$ des coordonnées des profils-lignes projetés sur $\mathbf{Vect}(v_1, \ldots, v_k)$ est notée :

$$X = L D_c^{-1} V_k$$

- $X_{:\alpha}$ est la α-ième composante principale des profils-lignes, de moyenne nulle ($\overline{X}_{:\alpha} = 0$) et de variance $\mathbf{Var}(X_{:\alpha}) = \lambda_\alpha$.

2.3 ACP de la matrice des profils-colonnes centrés

De façon duale, on représente les m modalités de X_2 dans un sous-espace de $\mathbb{R}^q$ de dimension $k \leq \min(q-1, m-1)$[5].

[5] On a au plus $q-1$ dimensions pour les lignes (du fait du centrage par les poids) et $m-1$ pour les colonnes (car la somme d'un profil vaut 1).

- Les axes sont engendrés par les vecteurs propres ordonnés $u_1, \dots, u_k$ de $CD_cC^\top D_r^{-1} \in \mathbb{R}^{q\times q}$ associés aux valeurs propres $\lambda_1 \geq \cdots \geq \lambda_k$
- À chaque étape, on ajoute une nouvelle direction qui maximise la variance (pondérée par D_c) des projections des profils-colonnes centrés, sous la métrique D_r^{-1}
- Les vecteurs directeurs u_α sont D_r^{-1}-orthonormés :

$$u_\alpha^\top D_r^{-1} u_\beta = \begin{cases} 1, & \text{si } \alpha = \beta, \\ 0, & \text{si } \alpha \neq \beta, \end{cases} \qquad \alpha, \beta = 1, \dots, k$$

- La matrice $m \times k$ des coordonnées des profils-colonnes projetés sur $\mathbf{Vect}(u_1, \dots, u_k)$ est

$$Y = C^\top D_r^{-1} U_k$$

- $Y_{:\alpha}$ est la α-ième composante principale des profils-colonnes, de moyenne nulle ($\overline{Y}_{:\alpha} = 0$) et de variance $\mathbf{Var}(Y_{:\alpha}) = \lambda_\alpha$.

2.4 SVD généralisée des taux de liaison

L'AFC peut se calculer par une décomposition en valeurs singulières (pondérée) de la matrice $Z \in \mathbb{R}^{q\times m}$ (matrice des écarts à l'indépendance ou des profils centrés) avec les métriques

$$N = D_r^{-1} \in \mathbb{R}^{q\times q} \qquad \text{et} \qquad M = D_c^{-1} \in \mathbb{R}^{m\times m}.$$

On écrit alors

$$Z = U \Lambda V^\top,$$

où, en posant $r = \mathbf{rang}(Z)$:

- $\Lambda = \mathbf{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_r})$ est la matrice des valeurs singulières ordonnées ; les λ_α sont les valeurs propres non nulles de $ZMZ^\top N$ et de $Z^\top NZM$;
- $U \in \mathbb{R}^{q\times r}$ contient en colonnes les vecteurs propres de $ZMZ^\top N$, orthonormés pour la métrique N : $U^\top NU = I_r$;
- $V \in \mathbb{R}^{m\times r}$ contient en colonnes les vecteurs propres de $Z^\top NZM$, orthonormés pour la métrique M : $V^\top MV = I_r$.

- **Coordonnées factorielles:** profils-lignes X et profils-colonnes Y)

$$\begin{cases} X = ZMV_k = U_k\Lambda_k \\ Y = Z^\top NU_k = V_k\Lambda_k \end{cases}$$

2.5 Calcul pratique à partir de la SVD standard

- **Calcul pratique (SVD standard) :** Plutôt que d'utiliser les routines de SVD généralisée, on peut effectuer la SVD classique de $\tilde{Z} = N^{1/2}ZM^{1/2} = \tilde{U}\tilde{\Lambda}\tilde{V}^\top$ (sous métriques identités I_q, I_m). On remonte ensuite aux facteurs par :

$$U = N^{-1/2}\tilde{U}, \qquad V = M^{-1/2}\tilde{V}, \qquad \Lambda = \tilde{\Lambda}$$

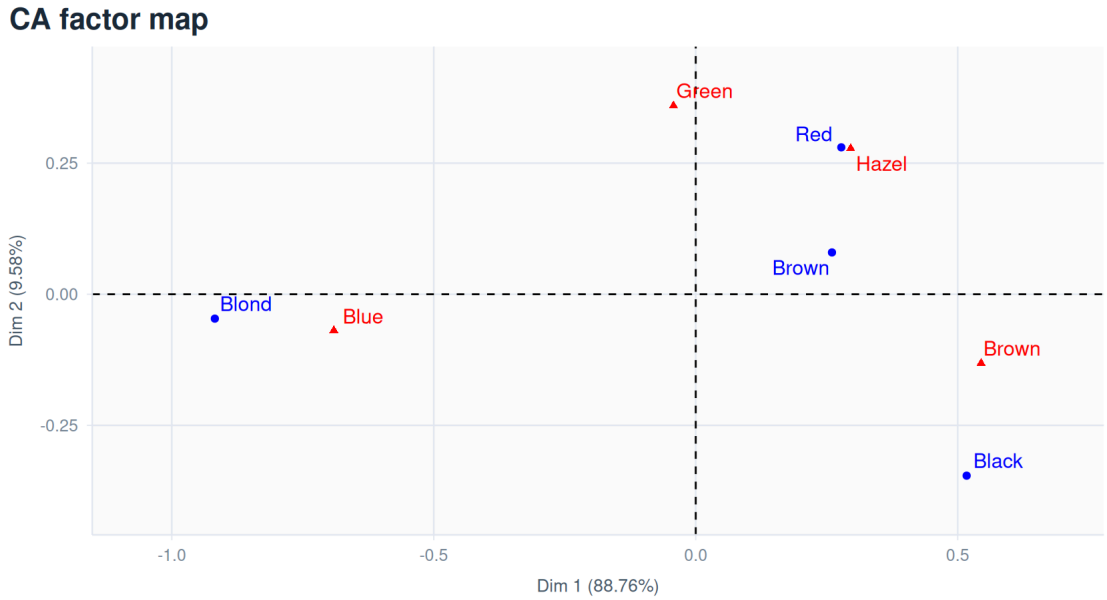
▸ **SVD Standard sur Z~\tilde ZZ~**

```
> # Métriques N = Dr^-1/2 et M = Dc^-1/2
+ N_sqrt = np.diag(1.0 / np.sqrt(r))
+ M_sqrt = np.diag(1.0 / np.sqrt(c))
+
+ # Transformation
+ Z_tilde = N_sqrt @ Z @ M_sqrt
+
+ # SVD classique : Z_tilde = U_tilde @ diag(s) @ V_tilde^T
+ U_tilde, s, Vt_tilde = svd(Z_tilde)
+
+ # Reconstitution des facteurs SVD généralisée
+ U_std = N_sqrt @ U_tilde
+ C_std = U_std @ np.diag(s)  # Coordonnées des profils-lignes
+
+ print(np.round(C_std[:, :2], 3))
```

```
[[-0.509  0.338]
 [-0.193 -0.144]
 [-0.197 -0.321]
 [ 0.76   0.085]]
```

2.6 Exemple

```
> library(FactoMineR)
>
> # Exécution de l'AFC (CA = Correspondence Analysis)
> res.ca <- CA(K, graph = TRUE)
```



- Axe 1 (inertie principale) : Oppose nettement les profils extrêmes sur la première dimension.
  - Côté lignes : opposition très marquée entre les cheveux Blond (à gauche) et Black / Brown (à droite).
  - Côté colonnes : opposition symétrique entre les yeux Blue (à gauche) et Brown (à droite).
- Axe 2 : Isole la modalité intermédiaire Red (lignes) et Hazel / Green (colonnes).

2.7 Composante principale standardisée

Une composante principale standardisée est une composante divisée par son écart-type ($\sqrt{\lambda_\alpha}$).

- La matrice Λ_k étant la matrice diagonale des valeurs singulières, les matrices X^* et Y^* des coordonnées factorielles standardisées s'écrivent :

$$\begin{cases} X^* = X\Lambda_k^{-1} \\ Y^* = Y\Lambda_k^{-1} \end{cases} \quad \Longrightarrow \quad \begin{cases} X^* = U_k \\ Y^* = V_k \end{cases}$$

- On a les relations de transition suivantes :

$$\begin{cases} X = D_r^{-1}(F - rc^\top)Y^* \\ Y = D_c^{-1}(F - rc^\top)^\top X^* \end{cases}$$

La coordonnée factorielle d'une modalité d'une variable est une combinaison linéaire des coordonnées factorielles (standardisées) des modalités de l'autre variable

$$\begin{cases} X_{i\alpha} = \sum_{s=1}^{m} \frac{f_{is}}{f_{i\cdot}} Y^*_{s\alpha} \\ Y_{s\alpha} = \sum_{i=1}^{q} \frac{f_{is}}{f_{\cdot s}} X^*_{i\alpha} \end{cases}$$

2.8 Propriétés barycentriques

On en déduit immédiatement que

$$X_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{s=1}^{m} \frac{f_{is}}{f_{i\cdot}} Y_{s\alpha} \qquad \text{et} \qquad Y_{s\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{q} \frac{f_{is}}{f_{\cdot s}} X_{i\alpha}$$

Interprétation géométrique et statistique

- **Principe du barycentre :** La coordonnée de la modalité i est la moyenne des coordonnées (standardisées) des modalités s de l'autre variable, pondérée par les fréquences conditionnelles $f_{is}/f_{i\cdot}$ (profil-ligne).
- **Dilatation :** Au coefficient de dilatation $1/\sqrt{\lambda_\alpha}$ près (puisque $\sqrt{\lambda_\alpha} < 1$), les coordonnées factorielles d'un nuage de points sont, sur chaque axe, les barycentres des coordonnées factorielles de l'autre nuage.

- **Représentation simultanée :** Ces relations justifient le tracé des profils-lignes et des profils-colonnes sur le même graphique (biplot). Mais **Attention :** la distance entre un profil-ligne et un profil-colonne ne s'interprète pas directement en termes de distance géométrique, mais en termes d'attraction/liaison (un profil-ligne se rapproche des profils-colonnes pour lesquels sa fréquence conditionnelle est élevée).

2.9 Exemple

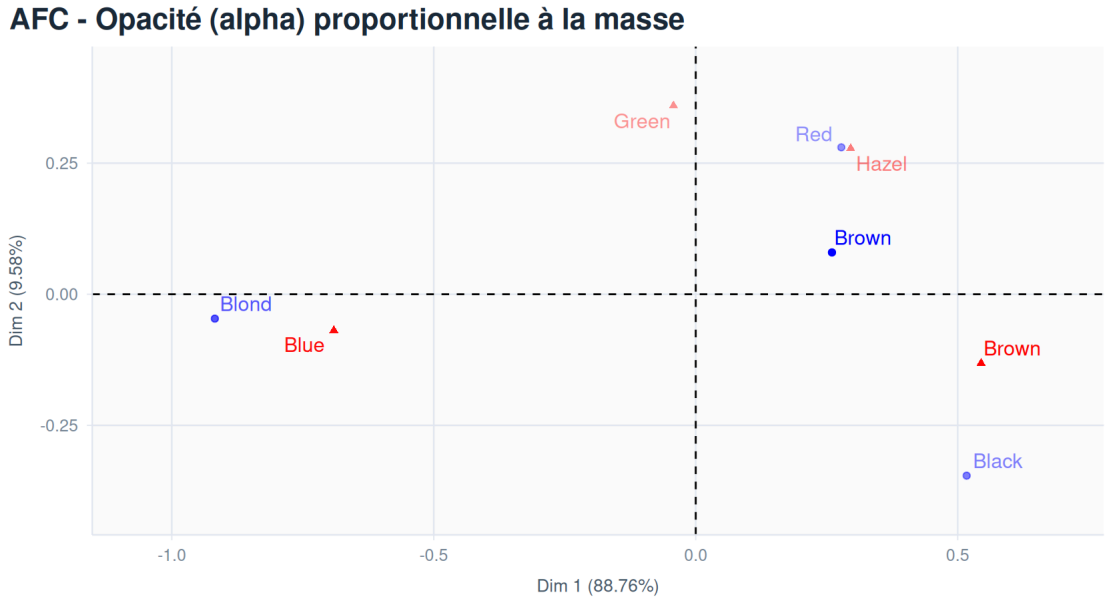
Vérification des relations barycentriques sur le 1er axe

```
> # 1. Extraction FactoMineR
> x1 <- res.ca$row$coord[, 1]                # Coordonnée x_1
> y1 <- res.ca$col$coord[, 1]                # Coordonnée y_1
> rsqrt_lambda <- 1 / sqrt(res.ca$eig[, 1])  # Valeur singulière  sqrt(lambda_1)
> rsqrt_lambda                               # poids dans les relation barycentrique
```

```
   dim 1     dim 2     dim 3
1.818304  5.535800 13.259526
```

```
> # --- Formule : x_i = (1 / sqrt(lambda_1)) * sum( (f_is / f_i.) * y_s )
> x_black_formule1 <- sum(L['Black',] * y1) * rsqrt_lambda[1]
>
>
> cat("FactoMineR (x1)                          :", x1["Black"],
+     "\nFormule avec y  (1/sqrt(lambda) * sum)   :", x_black_formule1, "\n")
```

```
FactoMineR (x1)                          : 0.5170739
Formule avec y  (1/sqrt(lambda) * sum)   : 0.5170739
```



3 Interprétation et critères de qualité

3.1 Interprétation : Inertie et test d'indépendance

- En ACP normée, l'inertie totale du nuage des points-individus est égale à p, le nombre de variables. **En AFC**, l'inertie totale des profils-lignes et des profils-colonnes est égale au ϕ^2 (le χ^2 d'indépendance divisé par l'effectif total n) :

$$I(L) = I(C) = \frac{\chi^2(X_1, X_2)}{n}$$

  La **dispersion des nuages de points est donc une mesure de liaison** entre les deux variables qualitatives, appelée aussi mesure d'écart à l'indépendance.

- **Nombre d'axes et inertie :** En AFC, il y a au plus $r = \min(q-1, m-1)$ valeurs propres non nulles. L'inertie totale s'écrit $\sum_{\alpha=1}^{r} \lambda_\alpha$. Chaque axe principal explique une part de cette inertie :

$$\frac{\lambda_\alpha}{\lambda_1 + \cdots + \lambda_r} \times 100$$

  qui s'interprète comme :
  - Le pourcentage d'inertie (ou de variance) expliqué par l'axe α,
  - La part de la liaison entre X_1 et X_2 captée par cet axe.

3.2 L'AFC en pratique

1. **Test préalable du χ^2**
   - L'inertie totale vaut $I = \frac{\chi^2}{n}$.
   - AFC pertinente **uniquement** si H_0 est rejetée ($p < 0{,}05$).
   - Évite d'interpréter le bruit d'échantillonnage.

2. **Choix du nombre d'axes**
   - Rang maximal : $r = \min(q-1,\, m-1)$.
   - Critère du coude sur l'éboulis des valeurs propres.
   - **Seuil de Benzécri :** conserver les axes où $\lambda_\alpha > \frac{1}{r}$.
3. **Interprétation des modalités**
   - $\mathrm{CTR}_\alpha(\cdot)$ **(Contributions) :** quelles modalités ***créent*** l'axe.
   - $\cos^2_\alpha(\cdot)$ **(Qualité) :** si l'axe ***explique*** bien la modalité.

3.3 Contributions absolues (CTR)

▸ **Définition**

Pour un axe factoriel α, la contribution d'une modalité mesure sa part dans l'inertie de cet axe (λ_α)

$$\begin{cases} \mathrm{CTR}_\alpha(i) = \frac{f_{i\cdot}\, x_{i\alpha}^2}{\lambda_\alpha} & \text{(Profils-lignes)} \\ \mathrm{CTR}_\alpha(s) = \frac{f_{\cdot s}\, y_{s\alpha}^2}{\lambda_\alpha} & \text{(Profils-colonnes)} \end{cases} \quad \text{avec} \quad \sum_i \mathrm{CTR}_\alpha(i) = \sum_s \mathrm{CTR}_\alpha(s) = 1$$

- **Rôle :** Identifie les modalités **responsables de la construction** de l'axe α (donne son sens à l'axe).
- **Règle de décision :** Une modalité est contributive si sa contribution excède le poids moyen : en lignes, si on a $\mathrm{CTR}_\alpha(i) > \frac{1}{q}$; en colonnes, si on a $\mathrm{CTR}_\alpha(s) > \frac{1}{m}$

> ⚠ **Attention au piège de l'excentrement**
>
> En AFC, un point très éloigné de l'origine n'est pas forcément un contributeur majeur : la contribution dépend du **produit** entre l'éloignement ($X_{i\alpha}^2$) et la masse de la modalité ($f_{i\cdot}$). Une modalité rare très excentrée peut avoir une contribution faible.

3.4 Qualité de représentation ($\cos^2$)

▸ **Définition**

Le $\cos^2$ mesure la qualité de projection d'une modalité sur l'axe factoriel α :

$$\begin{cases} \cos^2_\alpha(i) = \dfrac{x_{i\alpha}^2}{d^2_{D_r^{-1}}(L_{i:}, c^\top)} & \text{(Profils-lignes)} \\ \cos^2_\alpha(s) = \dfrac{y_{s\alpha}^2}{d^2(c_s, r)} & \text{(Profils-colonnes)} \end{cases} \quad \text{avec} \quad \sum_{\alpha=1}^{r} \cos^2_\alpha(\cdot) = 1$$

où $d^2_{D_r^{-1}}(L_{i:}, c^\top)$ est la distance du χ^2 entre le profil-ligne $L_{i:}$ et le profil moyen c.

- **Axe bien représenté :** On n'interprète la position d'un point sur un axe que si son $\cos^2_\alpha$ est élevé (proximité réelle, non un artefact de projection).
- **Proximité intra-variable :** Deux modalités d'une **même variable** proches et bien représentées ont des profils similaires.
- **Proximité inter-variables :** La proximité entre une modalité ligne et une colonne s'interprète uniquement via les **relations barycentriques** (attraction / liaison positive).

3.5 Références

- (Saporta 2006)
- François Husson, Marie Chavent, Vincent Rouanet et Pascal Lrz

Saporta, Gilbert. 2006. *Probabilités, analyse des données et statistique / Gilbert Saporta,...* 2e édition révisée et augmentée. Éditions Technip.