Analyse Factorielle des Correspondances (AFC)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique d'une table de contingence définie par deux variables qualitatives. Définition à partir de l'analyse en composantes principales des profils. Définition du modèle statistique associé, estimation. Représentation graphique simultanée des modalités des variables.

Travaux pratiques *de complexité croissante par l'études de données élémentaires.*

Retour au plan du cours.

1 Introduction

1.1 Données

On considère dans cette vignette deux variables qualitatives observées simultanément sur n individus affectés de poids identiques $1/n$. On suppose que la première variable, notée X, possède r modalités notées $x_1, \ldots, x_\ell, \ldots, x_r$, et que la seconde, notée Y, possède c modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

La table de contingence associée à ces observations, de dimension $r \times c$, est notée $\mathbf{T}$; son élément générique est $n_{\ell h}$, effectif conjoint. Elle se présente sous la forme suivante d'une table de contingence présentée dans le Tableau 1).

1.2 Notations

Les quantités $\{n_{\ell+} = \sum_{h=1}^{c} n_{\ell h}; \ell = 1, \ldots, r\}$ et $\{n_{+h} = \sum_{\ell=1}^{r} n_{\ell h}; h = 1, \ldots, c\}$ sont les *effectifs marginaux* vérifiant $\sum_{\ell=1}^{r} n_{\ell+} = \sum_{h=1}^{c} n_{+h} = n$. De façon analogue, on définit les notions de *fréquences conjointes* ($f_{\ell h} = n_{\ell h}/n$) et de *fréquences marginales* rangées dans les vecteurs :

TABLE 1 – *Table de contingence*

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell+}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

$$\begin{aligned} \mathbf{g}_r &= [f_{1+}, \ldots, f_{r+}]', \\ \text{et } \mathbf{g}_c &= [f_{+1}, \ldots, f_{+c}]'. \end{aligned}$$

Elles permettent de définir les matrices :

$$\begin{aligned} \mathbf{D}_r &= \text{diag}(f_{1+}, \ldots, f_{r+}), \\ \text{et } \mathbf{D}_c &= \text{diag}(f_{+1}, \ldots, f_{+c}). \end{aligned}$$

On sera également amené à considérer les profils–lignes et les profils–colonnes déduits de $\mathbf{T}$. Le ℓ-ième profil-ligne est

$$\{\frac{n_{\ell 1}}{n_{\ell+}}, \ldots, \frac{n_{\ell h}}{n_{\ell+}}, \ldots, \frac{n_{\ell c}}{n_{\ell+}}\}.$$

Il est considéré comme un vecteur de $\mathbb{R}^c$ et les r vecteurs ainsi définis sont disposés en colonnes dans la matrice $c \times r$

$$\mathbf{A} = \frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1}.$$

De même, le h-ième profil-colonne est

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\},$$

vecteur de $\mathbb{R}^r$, et la matrice $r \times c$ des profils-colonnes est

$$\mathbf{B} = \frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1}.$$

1.3 Liaison entre deux variables qualitatives

DÉFINITION 1. — *On dit que deux variables X et Y sont* non liées *relativement à T si et seulement si :*

$$\forall (\ell, h) \in \{1, \ldots, r\} \times \{1, \ldots, c\} : n_{\ell h} = \frac{n_{\ell +} n_{+h}}{n}.$$

Il est équivalent de dire que tous les profils-lignes sont égaux, ou encore que tous les profils-colonnes sont égaux.

Cette notion est cohérente avec celle d'indépendance en probabilités. En effet, soit $\Omega = \{1, \ldots, n\}$ l'ensemble des individus observés et $(\Omega, \mathcal{P}(\Omega), P)$ l'espace probabilisé associé où P est l'équiprobabilité ; $\mathcal{M}_X = \{x_1, \ldots, x_r\}$ et $\mathcal{M}_Y = \{y_1, \ldots, y_c\}$ désignent les ensembles de modalités, ou valeurs prises par les variables X et Y. On note $\widetilde{X}$ et $\widetilde{Y}$ les variables aléatoires associées aux 2 variables statistiques X et Y :

$$\begin{aligned} \widetilde{X} &: (\Omega, \mathcal{P}(\Omega), P) \mapsto (\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X)), \\ \widetilde{Y} &: (\Omega, \mathcal{P}(\Omega), P) \mapsto (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y)) ; \end{aligned}$$

P_X, P_Y et P_{XY} désignent respectivement les probabilités images définies par $\widetilde{X}, \widetilde{Y}$ et le couple $(\widetilde{X}, \widetilde{Y})$ sur $(\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X)), (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y))$ et $(\mathcal{M}_X \times \mathcal{M}_Y, \mathcal{P}(\mathcal{M}_X) \times \mathcal{P}(\mathcal{M}_Y))$; ce sont les probabilités empiriques. Alors, X et Y sont *non liées* si et seulement si $\widetilde{X}$ et $\widetilde{Y}$ sont *indépendantes en probabilité* (la vérification est immédiate).

On suppose maintenant qu'il existe une liaison entre X et Y que l'on souhaite étudier. La représentation graphique des profils-lignes ou des profils-colonnes, au moyen de diagrammes en barres parallèles, ainsi que le calcul de coefficients de liaison (Cramer ou Tschuprow) donnent une première idée de la variation conjointe des deux variables. Le test du χ^2 permet de plus de s'assurer du caractère significatif de cette liaison. Il est construit de la manière suivante :

l'hypothèse nulle est $H0$: $\widetilde{X}$ et $\widetilde{Y}$ sont indépendantes en probabilités ;

l'hypothèse alternative est $H1$: les variables $\widetilde{X}$ et $\widetilde{Y}$ ne sont pas indépendantes.

La statistique de test est alors

$$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{\left(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n}\right)^2}{\frac{n_{\ell +} n_{+h}}{n}};$$

elle suit asymptotiquement (pour les grandes valeurs de n), et si l'hypothèse $H0$ est vraie, une loi de χ^2 à $(r-1)(c-1)$ degrés de liberté. On rejette donc $H0$ (et l'on conclut au caractère significatif de la liaison) si χ^2 dépasse une valeur particulière (valeur ayant une probabilité faible et fixée a priori – en général 0,05 –être dépassée par une loi de χ^2 à $(r-1)(c-1)$ degrés de liberté).

1.4 Objectifs

Pour préciser la liaison existant entre les variables X et Y, on souhaite définir un modèle statistique susceptible de fournir des paramètres dont la représentation graphique (de type biplot) illustrera les "*correspondances*" entre les modalités de ces 2 variables. Cette approche sera développée au paragraphe 3.

Une autre approche, très courante dans la littérature francophone, consiste à définir l'Analyse Factorielle des Correspondances (AFC) comme étant le résultat d'une double Analyse en Composantes Principales

- l'ACP des profils–lignes,
- l'ACP des profils–colonnes,

relativement à la métrique dite du χ^2. Cette approche est présentée au paragraphe 2.

Remarque. — :

1. Toute structure d'ordre existant éventuellement sur les modalités de X ou de Y est ignorée par l'AFC
2. Tout individu présente une modalité et une seule de chaque variable.
3. Chaque modalité doit avoir été observée au moins une fois ; sinon, elle est supprimée.

2 Double ACP

2.1 Métriques du Chi2

Les correspondances entre modalités évoquées au paragraphe précédant se trouvent exprimées en termes de distances au sens d'une certaine métrique. Ainsi, chaque modalité x_ℓ de X est caractérisée par son profil–ligne représenté par le vecteur $\mathbf{a}^\ell$ de l'espace $\mathbb{R}^c$ muni de la base canonique (les coordonnées de $\mathbf{a}^\ell$ sont les éléments de la ℓ-ième colonne de $\mathbf{A}$). De même, chaque modalité y_h de Y est caractérisée par son profil–colonne représenté par le vecteur $\mathbf{b}^h$ de l'espace $\mathbb{R}^r$ muni de la base canonique.

Ces espaces sont respectivement munis des métriques, dites du χ^2, de matrices $\mathbf{D}_c^{-1}$ et $\mathbf{D}_r^{-1}$. Ainsi, la distance entre deux modalités x_ℓ et x_i de X s'écrit

$$\|\mathbf{a}^\ell - \mathbf{a}^i\|^2_{\mathbf{D}_c^{-1}} = \sum_{h=1}^{c} \frac{1}{f_{+h}}(a_h^\ell - a_h^i)^2,$$

et de même pour les modalités de Y. La métrique du χ^2 introduit les inverses des fréquences marginales des modalités de Y comme *pondérations* des écarts entre éléments de deux profils relatifs à X (et réciproquement) ; elle attribue donc plus de poids aux écarts correspondants à des modalités de *faible effectif* (rares) pour Y.

2.2 ACP des profils–colonnes

On s'intéresse ici à l'ACP du triplet $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$. Dans cette ACP, les "individus" sont les modalités de Y, caractérisées par les profils–colonnes de $\mathbf{T}$, pondérées par les fréquences marginales correspondantes et rangées en lignes dans la matrice $\mathbf{B}'$.

PROPOSITION 2. — *Les éléments de l'ACP de* $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$ *sont fournis par l'analyse spectrale de la matrice carrée,* $\mathbf{D}_r^{-1}$*–symétrique et semi–définie positive* $\mathbf{BA}$.

Preuve Elle se construit en remarquant successivement que :

1. le barycentre du nuage des profils–colonnes est le vecteur $\mathbf{g}_r$ des fréquence marginales de X,
2. la matrice $\mathbf{BD}_c\mathbf{B}' - \mathbf{g}_r\mathbf{D}_c\mathbf{g}_r'$ joue le rôle de la matrice des variances–covariances,
3. la solution de l'ACP est fournie par la D.V.S. de $(\mathbf{B}' - \mathbf{1}\mathbf{g}_r', \mathbf{D}_r^{-1}, \mathbf{D}_c)$, qui conduit à rechercher les valeurs et vecteurs propres de la matrice $(\mathbf{SM})$

$$\mathbf{BD}_c\mathbf{B}'\mathbf{D}_r^{-1} - \mathbf{g}_r\mathbf{D}_c\mathbf{g}_r' = \mathbf{BA} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1} \quad (\text{ car } \mathbf{B}'\mathbf{D}_r^{-1} = \mathbf{D}_c^{-1}\mathbf{A})$$

4. les matrices $\mathbf{BA} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1}$ et $\mathbf{BA}$ ont les mêmes vecteurs propres associées aux mêmes valeurs propres, à l'exception du vecteur $\mathbf{g}_r$ associé à la valeur propre $\lambda_0 = 0$ de $\mathbf{BA} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1}$ et à la valeur propre $\lambda_0 = 1$ de $\mathbf{BA}$.

□

On note $\mathbf{U}$ la matrice contenant les vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés de $\mathbf{BA}$. La représentation des "individus" de l'ACP réalisée fournit une représentation des modalités de la variable Y. Elle se fait au moyen des lignes de la matrice des "composantes principales" $(\mathbf{XMV})$:

$$\mathbf{C}_c = \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{U}.$$

2.3 ACP des profils–lignes

De façon symétrique (ou duale), on s'intéresse à l'ACP des "individus" modalités de X ou profils–lignes (la matrice des données est $\mathbf{A}'$), pondérés par les fréquences marginales des lignes de $\mathbf{T}$ (la matrice diagonale des poids est $\mathbf{D}_r$) et utilisant la métrique du χ^2. Il s'agit donc de l'ACP de $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$.

PROPOSITION 3. — *Les éléments de l'ACP de* $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$ *sont fournis par l'analyse spectrale de la matrice carrée,* $\mathbf{D}_c^{-1}$*–symétrique et semi–définie positive* $\mathbf{AB}$.

On obtient directement les résultats en permutant les matrices $\mathbf{A}$ et $\mathbf{B}$, ainsi que les indices c et r. Notons $\mathbf{V}$ la matrice des vecteurs propres de la matrice $\mathbf{AB}$; les coordonnées permettant la représentation les modalités de la variable X sont fournies par la matrice :

$$\mathbf{C}_r = \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V}.$$

Sachant que $\mathbf{V}$ contient les vecteurs propres de $\mathbf{AB}$ et $\mathbf{U}$ ceux de $\mathbf{BA}$, un théorème de l'annexe (st-m-explo-alglin Compléments d'algèbre linéaire)

montre qu'il suffit de réaliser une seule analyse, car les résultats de l'autre s'en déduisent simplement :

$$\begin{aligned}\mathbf{V} &= \mathbf{A}\mathbf{U}\mathbf{\Lambda}^{-1/2},\\ \mathbf{U} &= \mathbf{B}\mathbf{V}\mathbf{\Lambda}^{-1/2};\end{aligned}$$

$\mathbf{\Lambda}$ est la matrice diagonale des valeurs propres (exceptée $\lambda_0 = 0$) communes aux deux ACP

$$\begin{aligned}\mathbf{C}_c &= \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{U} = \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{B}\mathbf{V}\mathbf{\Lambda}^{-1/2} = \mathbf{D}_c^{-1}\mathbf{A}\mathbf{B}\mathbf{V}\mathbf{\Lambda}^{-1/2} = \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2},\\ \mathbf{C}_r &= \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V} = \mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2}.\end{aligned}$$

On en déduit les formules dites de *transition* :

$$\begin{aligned}\mathbf{C}_c &= \mathbf{B}'\mathbf{C}_r\mathbf{\Lambda}^{-1/2},\\ \mathbf{C}_r &= \mathbf{A}'\mathbf{C}_c\mathbf{\Lambda}^{-1/2}.\end{aligned}$$

La représentation simultanée habituellement construite à partir de ces matrices (option par défaut de SAS) n'est pas a priori justifiée. On lui donnera un sens dans les paragraphes suivants.

3 Modèles pour une table de contingence

On écrit d'abord que chaque fréquence $f_{\ell h}$ de $\mathbf{T}$ correspond à l'observation d'une probabilité théorique $p_{\ell h}$; on modélise donc la table de contingence par cette distribution de probabilités. On précise ensuite le modèle en explicitant l'écriture de $p_{\ell h}$. Différents modèles classiques peuvent être considérés.

3.1 Le modèle log–linéaire

Il consiste à écrire :

$$\ln(p_{\ell h}) = \mu + \alpha_\ell + \beta_h + \gamma_{\ell h}$$

avec des contraintes le rendant identifiable. Ce modèle, très classique, est développé par ailleurs.

3.2 Le modèle d'association

Il est encore appelé RC-modèle, ou modèle de Goodman :

$$p_{\ell h} = \gamma.\alpha_\ell.\beta_h.\exp\left(\sum_{k=1}^{q}\phi_k.\mu_{\ell k}.\nu_{hk}\right).$$

Ce modèle, muni des contraintes nécessaires, permet de structurer les interactions et de faire des représentations graphiques des lignes et des colonnes de $\mathbf{T}$ au moyen des paramètres $\mu_{\ll k}$ et ν_{hk}. Ces paramètres peuvent être estimés par maximum de vraisemblance ou par moindres carrés.

3.3 Le modèle de corrélation

On écrit ici :

$$p_{\ell h} = p_{\ell +}p_{+h} + \sum_{k=1}^{q}\sqrt{\lambda_k}u_\ell^k v_h^k, \tag{1}$$

avec $q \leq \inf(r-1, c-1)$, $\lambda_1 \geq \cdots \geq \lambda_q > 0$ et sous les contraintes d'identifiabilité suivantes :

$$\begin{aligned}\sum_{\ell=1}^{r}u_\ell^k = \sum_{h=1}^{c}v_h^k &= 0,\\ \mathbf{u}^{k'}\mathbf{D}_r^{-1}\mathbf{u}^j = \mathbf{v}^{k'}\mathbf{D}_c^{-1}\mathbf{v}^j &= \delta_{kj}.\end{aligned}$$

Remarque. — :

1. Le modèle (1) ci-dessus est équivalent au modèle considéré par Goodman :

$$p_{\ell h} = p_{\ell +}p_{+h}\left(1 + \sum_{k=1}^{q}\sqrt{\lambda_k}\xi_\ell^k\eta_h^k\right), \tag{2}$$

   moyennant une homothétie sur les paramètres.
2. La quantité $\sum_{k=1}^{q}\sqrt{\lambda_k}u_\ell^k v_h^k$ exprime l'écart à l'indépendance pour la cellule considérée.

3. Le modèle suppose que cet écart se décompose dans un sous–espace de dimension $q < \min(c-1, r-1)$.
4. Les estimations des paramètres $p_{\ell+}, p_{+h}, \lambda_k, \mathbf{u}^k, \mathbf{v}^k$ peuvent être réalisées par maximum de vraisemblance [1] ou par moindres carrés. Dans le contexte de la statistique descriptive, qui est celui de ce cours, il est naturel de retenir cette dernière solution.

3.4 Estimation Moindres Carrés dans le modèle de corrélation

3.4.1 Critère

Considérons les espaces $\mathbb{R}^c$ et $\mathbb{R}^r$ munis de leur base canonique et de leur métrique du χ^2 respectives et notons $\mathbf{P}$ le tableau des probabilités théoriques définies selon le modèle (1). Le critère des moindres carrés s'écrit alors :

$$\min_{\mathbf{P}} \left\| \frac{1}{n}\mathbf{T} - \mathbf{P} \right\|^2_{\mathbf{D_r^{-1}D_c^{-1}}}. \tag{3}$$

3.4.2 Estimation

PROPOSITION 4. — *L'estimation des paramètres de (1) en résolvant (3) est fournie par la D.V.S. de* $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$ *à l'ordre* q. *Les probabilités marginales* $p_{\ell+}$ *et* p_{+h} *sont estimées par* $f_{\ell+}$ *et* f_{+h} *tandis que les vecteurs* $\mathbf{u}^k$ *(resp.* $\mathbf{v}^k$*) sont vecteurs propres de la matrice* $\mathbf{BA}$ *(resp.* $\mathbf{AB}$*) associés aux valeurs propres* λ_k.

On obtient ainsi, d'une autre façon, l'AFC de la table de contingence $\mathbf{T}$.

Preuve Elle se construit à partir de la D.V.S. de $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$:

$$\frac{1}{n}t_\ell^h = \sum_{k=0}^{\min(r-1,c-1)} \sqrt{\lambda_k} u_\ell^k v_h^k,$$

où les vecteurs $\mathbf{u}^k$ (resp. $\mathbf{v}^k$) sont vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés (resp. $\mathbf{D}_c^{-1}$–orthonormés) de la matrice

$$\frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1}\frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1} = \mathbf{BA} \text{ (resp. } \frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1}\frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1} = \mathbf{AB}),$$

associés aux valeurs propres λ_k.

De plus, le vecteur $\mathbf{g}_r = \mathbf{u}^0$ (resp. $\mathbf{g}_c = \mathbf{v}^0$) est vecteur propre $\mathbf{D}_r^{-1}$–normé (resp. $\mathbf{D}_c^{-1}$–normé) de la matrice $\mathbf{BA}$ (resp. $\mathbf{AB}$) associé à la valeur propre $\lambda_0 = 1$. Enfin, les matrices $\mathbf{AB}$ et $\mathbf{BA}$ sont stochastiques [2] et donc les valeurs propres vérifient :

$$1 = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_q > 0.$$

En identifiant les termes, l'approximation de rang $(q+1)$ de la matrice $\mathbf{P}$ s'écrit donc :

$$\widehat{\mathbf{P}}_q = \mathbf{g}_r\mathbf{g}_c' + \sum_{k=1}^{q} \sqrt{\lambda_k}\mathbf{u}^k\mathbf{v}^{k'}$$

et les propriétés d'orthonormalité des vecteurs propres assurent que les contraintes du modèle sont vérifiées.

□

4 Représentations graphiques

4.1 Biplot

La décomposition de la matrice $\frac{1}{n}\mathbf{T}$ se transforme encore en :

$$\frac{f_{\ell h} - f_{\ell+}f_{+h}}{f_{\ell+}f_{+h}} = \sum_{k=0}^{\min(r-1,c-1)} \sqrt{\lambda_k}\frac{u_\ell^k}{f_{\ell+}}\frac{v_h^k}{f_{+h}}.$$

En se limitant au rang q, on obtient donc, pour chaque cellule (ℓ, h) de la table $\mathbf{T}$, une approximation de son écart relatif à l'indépendance comme produit scalaire des deux vecteurs

$$\frac{u_\ell^k}{f_{\ell+}}\lambda_k^{1/4} \text{ et } \frac{v_h^k}{f_{+h}}\lambda_k^{1/4},$$

termes génériques respectifs des matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/4} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/4},$$

1. On suppose alors que les n $p_{\ell h}$ sont les paramètres de lois de Poisson indépendantes conditionnellement à leur somme qui est fixée et égale à n.

2. Matrice réelle, carrée, à termes positifs, dont la somme des termes de chaque ligne (ou chaque colonne) vaut 1.

qui sont encore les estimations des vecteurs ξ_ℓ et η_h du modèle 2. Leur représentation (par exemple avec $q = 2$) illustre alors la *correspondance* entre les deux modalités x_ℓ et y_h : lorsque deux modalités, éloignées de l'origine, sont voisines (resp. opposées), leur produit scalaire est de valeur absolue importante ; leur cellule conjointe contribue alors fortement et de manière positive (resp. négative) à la dépendance entre les deux variables.

L'AFC apparaît ainsi comme la meilleure reconstitution des fréquences $f_{\ell h}$, ou encore la meilleure représentation des écarts relatifs à l'indépendance. La représentation simultanée des modalités de X et de Y se trouve ainsi pleinement justifiée.

4.2 Double ACP

Chacune des deux ACP réalisée permet une représentation des "individus" (modalités) approchant, au mieux, les distances du χ^2 entre les profils–lignes d'une part, les profils–colonnes d'autre part. Les coordonnées sont fournies cette fois par les matrices (de composantes principales)

$$\mathbf{C}_r = \mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2} \text{ et } \mathbf{C}_c = \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}.$$

Même si la représentation simultanée n'a plus alors de justification, elle reste couramment employée. En fait, les graphiques obtenus diffèrent très peu de ceux du biplot ; ce dernier sert donc de "caution" puisque les interprétations des graphiques sont identiques. On notera que cette représentation issue de la double ACP est celle réalisée par la plupart des logiciels statistiques (c'est en particulier le cas de SAS).

4.3 Représentations barycentriques

D'autres représentations simultanées, appelées barycentriques, sont proposées en utilisant les matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda},$$

ou encore les matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}.$$

Si l'on considère alors, par exemple, la formule de transition

$$\mathbf{C}_r = \mathbf{A}'\mathbf{C}_c\mathbf{\Lambda}^{-1/2} \Longleftrightarrow \mathbf{C}_r\mathbf{\Lambda}^{1/2} = \mathbf{A}'\mathbf{C}_c \Longleftrightarrow \mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda} = \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2},$$

on voit que dans la seconde des représentations ci–dessus, chaque modalité x_ℓ de X est représentée par un vecteur qui est barycentre de l'ensemble des vecteurs associés aux modalités de Y, chacun d'eux ayant pour poids l'élément correspondant du l-ième profil–ligne. Là encore, la représentation simultanée s'en trouve parfaitement justifiée. Malheureusement, dans la pratique, les représentations barycentriques sont souvent illisibles ; elles sont, de ce fait, très peu utilisées.

4.4 Autre représentation

La pratique de l'AFC montre que l'interprétation des graphiques est toujours la même, quelle que soit la représentation simultanée choisie parmi les 3 ci–dessus.

On peut ainsi envisager d'utiliser, pour une représentation simultanée des modalités de X et de Y, les coordonnées fournies respectivement par les lignes des matrices

$$\mathbf{D}_r^{-1}\mathbf{U} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}.$$

L'interprétation du graphique sera toujours la même et les matrices ci–dessus, outre leur simplicité, présentent l'avantage de conduire a une représentation graphique qui reste invariante lorsque l'on utilise la technique d'Analyse Factorielle des Correspondances Multiples sur les données considérées ici.

4.5 Aides à l'interprétation

Les qualités de représentation dans la dimension choisie et les contributions des modalités de X ou de Y se déduisent aisément de celles de l'ACP Ces quantités sont utilisées à la fois pour choisir la dimension de l'AFC et pour interpréter ses résultats dans la dimension choisie.

4.5.1 Mesure de la qualité globale

Pour une dimension donnée q ($1 \leq q \leq d = \inf(r-1, c-1)$), la qualité globale des représentations graphiques en dimension q se mesure par le rapport entre la somme des q premières valeurs propres de l'AFC et leur somme complète de 1 à d.

Compte–tenue de la propriété $\sum_{k=1}^{d} \lambda_k = \Phi^2$ (voir en 6.1), la qualité de la

représentation dans la k–ième dimension s'écrit

$$\frac{n\lambda_k}{\chi^2}.$$

On parle encore de part du khi–deux expliquée par la k–ième dimension (voir les sorties du logiciel SAS).

4.5.2 Mesure de la qualité de chaque modalité

Pour chaque modalité de X (resp. de Y), la qualité de sa représentation en dimension q se mesure par le cosinus carré de l'angle entre le vecteur représentant cette modalité dans $\mathbb{R}^c$ (resp. dans $\mathbb{R}^r$) et sa projection $\mathbf{D}_c^{-1}$–orthogonale (resp. $\mathbf{D}_r^{-1}$–orthogonale) dans le sous–espace principal de dimension q.

Ces cosinus carrés s'obtiennent en faisant le rapport des sommes appropriées des carrés des coordonnées extraites des lignes de $\mathbf{C}_r$ (resp. de $\mathbf{C}_c$).

4.5.3 Contributions à l'inertie totale

L'inertie totale (en dimension d) du nuage des profils–lignes (resp. des profils–colonnes) est égale à la somme des d valeurs propres. La part due au i–ième profil–ligne (resp. au j–ième profil–colonne) valant $f_{\ell+}\sum_{k=1}^{d}(c_{r\ell}^k)^2$ (resp. $f_{+h}\sum_{k=1}^{d}(c_{ch}^k)^2$), les contributions à l'inertie totale s'en déduisent immédiatement.

4.5.4 Contributions à l'inertie selon chaque axe

Il s'agit de quantités analogues à celles ci–dessus, dans lesquelles il n'y a pas de sommation sur l'indice k. Ces quantités sont utilisées dans la pratique pour sélectionner les modalités les plus importantes, c'est–à–dire celles qui contribuent le plus à la définition de la liaison entre les 2 variables X et Y.

4.5.5 Remarque

En général, on n'interprète pas les axes d'une AFC (en particulier parce qu'il n'y a pas de variable quantitative intervenant dans l'analyse). L'interprétation s'appuie surtout sur la position relative des différentes modalités repérées comme les plus importantes.

5 Exemple

L'exemple des données bancaires ainsi que les données d'expression génomique se prête mal à l'illustration d'une analyse des correspondances, aucun couple de variable qualitative ne conduit à des représentations intéressantes.

La table de contingence étudiée à titre d'exemple décrit la répartition des exploitations agricoles de la région Midi–Pyrénées dans les différents départements en fonction de leur taille. Elle croise la variable qualitative *département*, à 8 modalités, avec la variable *taille de l'exploitation*, quantitative découpée en 6 classes. Les données, ainsi que les résultats numériques obtenus avec la procédure `corresp` de SAS/STAT, sont fournis en annexe.

La figure 5 présente le premier plan factoriel utilisant les coordonnées obtenues par défaut, c'est–à–dire celles de la double ACP.

6 Compléments

6.1 Propriétés

- *Formule de reconstitution des données.* On appelle ainsi l'approximation d'ordre q (c'est–à–dire fournie par l'AFC en dimension q) de la table des fréquences initiales ($\frac{1}{n}\mathbf{T}$) :

$$f_{\ell h} \simeq f_{\ell+}f_{+h}\sum_{k=1}^{q}\sqrt{\lambda_k}u_\ell^k v_h^k.$$

- Les valeurs propres vérifient :

$$\sum_{k=1}^{d}\lambda_k = \Phi^2.$$

En effet, on vérifie facilement :

$$\mathrm{tr}\mathbf{AB} = \sum_{k=0}^{d}\lambda_k = 1 + \frac{\chi^2}{n} = 1 + \Phi^2;$$

d'où le résultat.

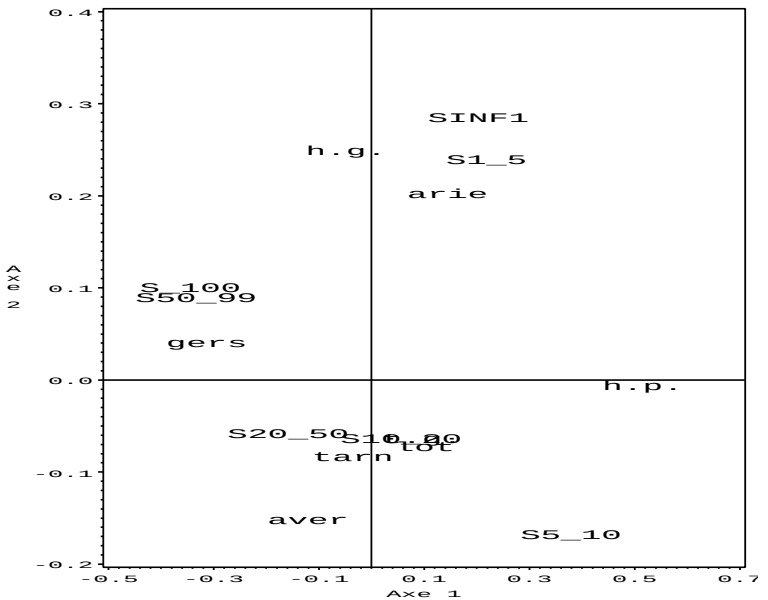

FIGURE 1 – *Répartition des exploitations agricoles par taille et par département. Premier plan de l'AFC.*

6.2 Invariance

- Les tables de contingence $\mathbf{T}$ et $\alpha\mathbf{T}, \alpha \in \mathbb{R}_+^*$, admettent la même AFC (évident).
- *Propriété d'équivalence distributionnelle* : si deux lignes de $\mathbf{T}$, ℓ et i, ont des effectifs proportionnels, alors les représentations de x_ℓ et x_i sont confondues (leurs profils sont identiques) et le regroupement de x_ℓ et x_i en une seule modalité (en additionnant les effectifs) laisse inchangées les représentations graphiques (même chose pour les colonnes de $\mathbf{T}$). Cette propriété est une conséquence de la métrique du χ^2.

6.3 Choix de la dimension

Le choix de la dimension pose les mêmes problèmes qu'en ACP De nombreuses techniques empiriques ont été proposées (essentiellement : part d'inertie expliquée, éboulis des valeurs propres). Il existe également une approche probabiliste qui peut donner des indications intéressantes. Nous la détaillons ci–dessous.

Posons

$$\widehat{n_{\ell h}^q} = n f_{\ell +} f_{+h} + n \sum_{k=1}^{q} \sqrt{\lambda_k} u_\ell^k v_h^k,$$

estimation d'ordre q de l'effectif conjoint de la cellule (ℓ, h). Alors, sous certaines conditions (échantillonnage, n grand, modèle multinomial ...), on peut montrer que

$$K_q = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - \widehat{n_{\ell h}^q})^2}{\widehat{n_{\ell h}^q}} \simeq n \sum_{k=q+1}^{d} \lambda_k$$

suit approximativement une loi de χ^2 à $(r-q-1)(c-q-1)$ degrés de liberté. On peut donc retenir pour valeur de q la plus petite dimension pour laquelle K_q est inférieure à la valeur limite de cette loi. Le choix $q = 0$ correspond à la situation où les variables sont proche de l'indépendance en probabilités ; les fréquences conjointes sont alors bien approchées par les produits des fréquences marginales.