Introduction

L'analyse des données est utilisée dès lors que les données se présentent en trop grand nombre pour être appréhendées par l'esprit humain. De nombreux domaines ont recours à l'analyse de données :

- en sciences humaines, cette technique est utilisée pour cerner les résultats des enquêtes d'opinion;
- la sociologie compte beaucoup sur l'analyse des données pour comprendre la vie et le développement de certaines populations, pour analyser les réponses à des questionnaires;
- le domaine du sport est très friand de statistiques : un médecin du sport s'interroge sur l'âge des pratiquants, leurs motivations et le sport qu'ils pratiquent;
- dans le domaine des sciences et techniques, certains chercheurs adoptent ces méthodes statistiques pour déchiffrer plusieurs caractéristiques du génome. D'autres se servent de l'analyse des données pour mettre en place un processus nécessaire à la reconnaissance des visages.

1 Analyse de données

On appelle **statistique** l'ensemble des méthodes permettant d'analyser (de traiter) des ensembles d'observations (de données). L'analyse de données est donc un domaine des statistiques qui se préoccupe de la description de données multidimensionnelles.

Il existe deux principales méthodes statistiques:

- *méthodes descriptives :* méthodes dont l'objectif est la description des données étudiées à travers leur représentation graphique des individus et/ou des variables et le calcul de résumés numériques en ayant recours à la géométrie euclidienne.
  Autres synonymes : statistiques descriptives, méthodes exploratoires.
- *méthodes inférentielles :* méthodes dont l'objectif est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population en ayant recours à des hypothèses géométriques. Il s'agit d'induire (ou d'inférer) du particulier au général.
  Autre synonyme : modélisation statistique, statistique inductive.

D'un point de vue méthodologique, on notera que la statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données : les deux aspects de la statistique se complètent bien plus qu'elles ne s'opposent.

2 Représentation des données

L'expression "Analyse de données" recouvre les techniques ayant pour objectif la description statistique des grands tableaux $n \times p$ où n lignes (ou individus / observations/ unités statistiques) variant de quelques dizaines à quelques centaines milliers voire plus, p colonnes (ou variables / facteurs/ attritbuts) variant de quelques unités à quelques dizaines.

2.1 A propos des variables

Les p variables représentant un individu peuvent être :

- quantitatives : prend des valeurs entières ou réelles. Par exemple, des mesures (valeurs numériques)

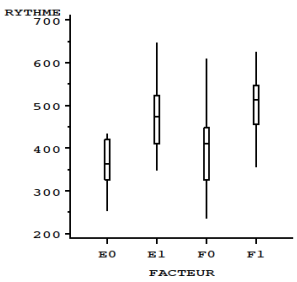
- qualitatives : ne sont pas des valeurs numériques mais des caractéristiques appelées modalités :

  - modalités naturellement ordonnées : variables ordinales.
    Exemples : mention au bac, classe d'âge.
  - sinon variable nominales
    Exemples : profession d'une population de personnes actives / situation familiale.

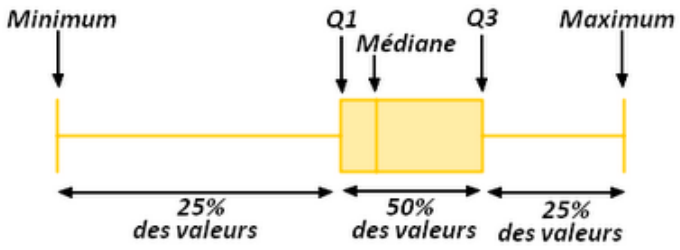
Ces variables peuvent être homogènes : représentant la même grandeur et exprimées dans la même unité ou sinon on parlera de variables hétérogènes.

2.2 Variable qualitative et variable quantitative :

Soit X la variable qualitative considérée, supposée à r modalités notées : $x_1, ...x_r$ et soit Y la variable quantitative. Une façon commode de représenter les données dans le cas d'une étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser des boîtes parallèles; il s'agit, sur un même graphique doté d'une échelle unique, de représenter pour Y une boîte-à-moustaches pour chacune des sous-populations définies par X. La comparaison de ces boîtes donne une idée assez claire de l'influence de X sur les valeurs de Y c'est-à-dire de la liaison entre les deux variables.



Boîtes parallèles



Boîte à moustache (boxplot)

2.2.1 Cas des variables qualitatives

2 variables qualitatives

Considérons un tableau de contingence ou tableau croisé de co-occurrence K, dans lequel on dispose 2 modalités respectivement en n lignes et p colonnes. A l'intersection d'une ligne et d'une colonne de K, l'élement k_{ij} représente le nombre d'individus ayant simultanément la variable nominale i et la variable nominale j. Le total marginal $k_{i.}$ est le nombre d'individu ayant la variable i alors que le total marginal $k_{.j}$ représente le nombre d'individus ayant la variable j.

		couleur des cheveux				profil moyen
		brun	châtain	roux	blond	
couleur des yeux	marron	11	20	4	1	37
	noisette	3	9	2	2	16
	vert	1	5	2	3	11
	bleu	3	14	3	16	36
profil moyen		18	48	12	21	100

Exemple de tableau de contingence

Par exemple : soit K la matrice obtenue en ventilant une population de 592 femmes suivant leurs couleurs des yeux et des cheveux. En lignes est présentée la variable "couleur des yeux" et en colonne est donnée la variable "couleur des cheveux".
k_ij = nombre de femmes ayant la couleur des yeux i et la couleur de cheveux j.
$k_{i.}$ = nombre de femmes ayant les yeux de la couleur i.
$k_{.j}$ = nombre de femmes ayant la couleur de cheveux j.

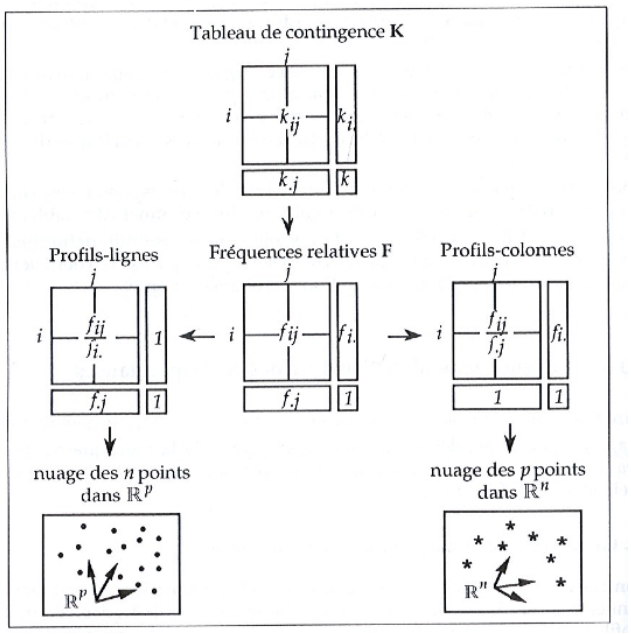
On définit ainsi des relations :

$$k_{i.} = \sum_{j=1}^{p} k_{ij}, \qquad k_{.j} = \sum_{i=1}^{n} k_{ij}, \qquad k = \sum_{i=1}^{n}\sum_{j=1}^{p} k_{ij}.$$

qui, en termes de fréquences relatives, donnent lieu aux relations :

$$f_{ij} = \frac{k_{ij}}{k} \qquad f_{i.} = \sum_{j=1}^{p} f_{ij}, \qquad f_{.j} = \sum_{i=1}^{n} f_{ij}, \qquad \sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij} = 1.$$

On construit ainsi la matrice F des fréquences relatives f_{ij}.



Transformations du tableau de contingence

Représentation des données

- Nuage des n lignes : l'ensemble des profils-lignes forme un nuage de n points dans l'espace des p colonnes. Chaque point i a pour coordonnées dans $\mathbb{R}^p$:

$$\{\frac{f_{ij}}{f_{i.}}, j = 1, .., p\}.$$

  Il est affecté d'une masse $f_{i.}$ qui est sa fréquence relative.

- Nuage des p colonnes : l'ensemble des p profils-colonnes constitue un nuage de p points dans l'espace des n lignes. Les coordonnées dans $\mathbb{R}^n$ du oint j sont données par :

$$\{\frac{f_{ij}}{f_{.j}}, i = 1, .., n\}$$

  Chaque point est affecté d'une masse $f_{.j}$.

Généralisation

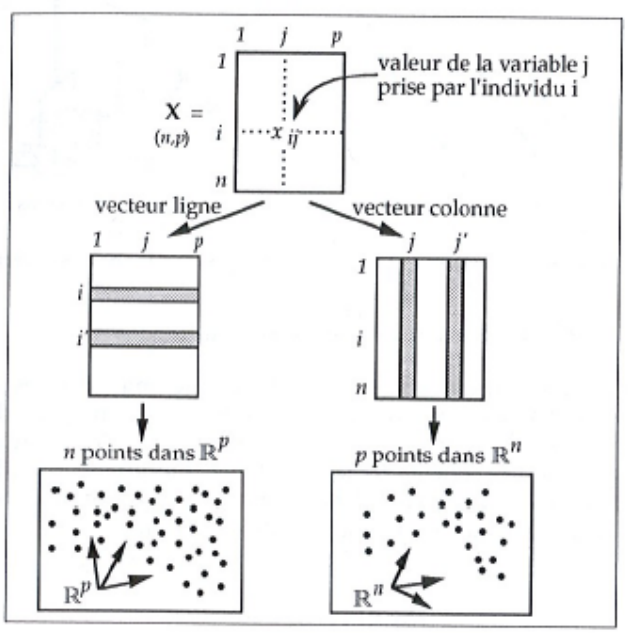
Tableau de Burt Le tableau de Burt est une généralisation particulière du tableau de contingence dans le cas où l'on étudie simultanément p variables qualitatives. Notons $X^1, .., X^p$ ces variables et appelons c_j le nombre de modalités de X^j, $j = 1, .., p$ et posons $c = \sum_{j=1}^{p} c_j$. Le tableau de Burt est une matrice carrée $c \times c$, constituée de p^2 sous-matrices.
Chacune des p sous-matrices diagonales est relative à l'une des p variables; la j^{ime} d'entre elles est carrée d'ordre c_j, diagonale, et comporte sur la diagonale les effectifs marginaux de X^j. La sous-matrice figurant dans le bloc d'indice (j, j'), $j \neq j'$, est le tableau des contingences construites en mettant X^j en lignes et $X^{j'}$ en colonnes; le tableau de Burt est donc symétrique.

	bacC	bacD	< 18	18ans	19ans	> 19	2ans	3ans	4ans
bacC	583	0	108	323	114	38	324	192	67
bacD	0	214	25	97	68	24	76	82	56
< 18	108	25	133	0	0	0	84	35	14
18ans	323	97	0	420	0	0	224	137	59
19ans	114	68	0	0	182	0	73	75	34
> 19	38	24	0	0	0	62	19	27	16
2ans	324	76	84	224	73	19	400	0	0
3ans	192	82	35	137	75	27	0	274	0
4ans	67	56	14	59	34	16	0	0	123

Tableau de Burt

2.2.2 Cas des variables quantitatives

La plupart des méthodes de statistique exploratoire multidimensionnelle repose sur des représentations géométriques des données du tableau.



Principe de représentation graphique

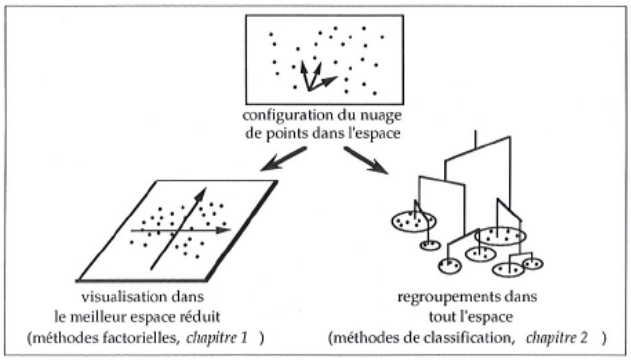
Deux nuages de points peuvent être construits :

- le nuage des n individus : nuages de points "lignes" situés dans l'espace $\mathbb{R}^p$ des variables;
- le nuage des p variables : nuages de points "colonnes" situés dans l'espace $\mathbb{R}^n$ des individus.

3 ... vers les statistiques exploratoires

Ces représentations géométriques du tableau de données nous conduisent naturellement à utiliser les notions d'espaces vectoriels. Il est donc possible de :

- définir des distances entre individus/variables,
- pondérer l'influence d'un individu/variable,
- identifier des regroupements (agrégations/clusters),
- identifier des relations/liens de dépendances (entre variables/individus).



Les deux grandes familles de méthodes

Les développements théoriques des méthodes de statistique exploratoire multidimensionnelle vont reposer sur ces notions. Il existe 2 principales approches:

- Méthodes factorielles : méthodes de réduction de dimension (ACP, AFC...)
- Méthodes de classification : produire des regroupements lignes ou colonnes, créer une partition (kmeans, CAH, SOM)

Méthodes factorielles

Les méthodes factorielles permettent de fournir des représentations synthétiques de vastes ensembles de valeurs numériques, en général sous forme de visualisations graphiques.
Pour cela, on cherche à réduire les dimensions du tableau de données en représentant les associations/dépendances entre individus et entre variables dans ces espaces de faibles dimensions.

Dans ce chapitre, les techniques suivantes seront présentées :

- *Analyse en composantes principales* s'applique aux tableaux de type "variables-individus", dont les colonnes sont des variables à valeurs numériques continues et dont les lignes sont des individus, des observations, des objets... Les proximités entre variables s'interprètent en terme de corrélation ; les proximités entre individus s'interprètent en termes de similitudes globales des valeurs observées.

- *Analyse des correspondances* s'applique aux tableaux de contingences ou tableau croisé de cooccurrence, c'est-à-dire aux tableaux de comptages obtenus par croisement de deux variables nominales. Ces tableaux ont la particularité de faire jouer un rôle identique aux lignes et aux colonnes. L'analyse fournit des représentations des associations entre lignes et colonnes de ces tableaux, fondées sur une distance entre profils (qui sont des vecteurs de fréquences conditionnelles) désignée sous le nom de distance du χ^2.

4 Analyse en composantes principales

4.1 Matrice de variance-covariance

A partir des données multidimensionnelles, on étudie généralement les liaisons entre les variables observées: c'est ce qu'on appelle l'étude des corrélations. Soit X le tableau des données de dimensions $n \times p$ d'élément x_{ij} construit à partir de n individus définis par p variables.
On définit l'individu moyen par le vecteur de $\mathbb{R}^p$ par : $g = [\overline{x_1}, ...\overline{x_p}]$.
Soit maintenant X_C le tableau centré en g de dimensions $n \times p$ défini par : $X_{C_{ij}} = x_{ij} - \overline{x_j}$.

Différentes formes de nuages de points : forme allongée, forme parabolique, forme sphérique

Matrice de variance-Covariance Σ : On appelle *la matrice de variance-covariance*, la matrice symétrique de dimension $p \times p$ dont les éléments diagonaux représentent la variance des variables et les éléments hors diagonaux la covariance entre les variables :

$$\Sigma = \frac{1}{n} X_C^T X_C$$

La covariance de la variable j et l, notée Σ_{jl}, mesure la liaison/dépendance des paramètres :

$$\Sigma_{jl} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x_j})(x_{il} - \overline{x_l})$$

Corrélation entre les variables : A partir de cette matrice Σ, on définit aussi *la corrélation entre les variables X et Y*, indépendant des unités de mesure de X et de Y. Le coefficient de corrélation est symétrique :

$$-1 \leq Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \leq 1$$

- $Corr(X, Y) = 0$, les variables sont quasiment décorrélées c'est-à-dire étant donné X, on ne peut rien dire.prédire sur la valeur de Y.
- $Corr(X, Y) = 1$, dépendance linéaire positive de X et Y.
- $Corr(X, Y) = -1$, dépendance linéaire négative de X et Y.

4.2 Méthode

Les composantes principales $C_1, .., C_q$ sont des nouvelles variables combinaison linéaire des variables d'origines $x_{.1}, ...x_{.p}$ telles que les C_k soient 2 à 2 non corrélées, de variance maximale, d'importance décroissante.

Décomposition de la variance :

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - g)^T (x_i - g)$$

Projection sur une droite : L'opérateur de projection orthogonale, noté π, sur une droite de vecteur directeur unitaire v s'écrit :

$$\Pi = vv^T$$

avec $v^T v = 1$. La variance des observations projetés s'écrit alors :

$$\begin{aligned} \sigma_V^2 &= \frac{1}{n} \sum_{i=1}^{n} (\Pi(x_i - g))^T (\Pi(x_i - g)) \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i - g)^T vv^T (x_i - g)) \\ &= v^T \Sigma v \end{aligned}$$

Recherche de la projection de variance maximale Observons que Σ est la matrice de variance-covariance. Cette matrice est symétrique définie positive. On doit donc maximiser cette variance des observations projetées:

$$\max_v v^T \Sigma v \text{ avec } v^T v = 1$$

Il s'agit d'un problème d'optimisation avec contrainte d'égalité. On introduit donc le Lagrangien :

$$\mathcal{L} = v^T \Sigma v + \lambda(1 - v^T v)$$

et on calcule les conditions nécessaires d'optimalité : $\partial_v \mathcal{L} = 0$. On obtient donc l'équation aux valeurs propres :

$$\Sigma v = \lambda v.$$

Comme la matrice Σ est symétrique définie positive, les valeurs propres sont réelles positives et les vecteurs propres peuvent être choisis orthonormés. Donc la solution est de projeter les données sur le vecteur propres associé à la plus grande valeur propres λ de Σ.

Recherche des projections de variance maximale orthogonales au premier axe : Afin de trouver le second axe de variance maximale, on résout :

$$\max_v v^T \Sigma v \text{ avec } v^T v = 1, v^T v_1 = 0.$$

Interprétation des vecteurs propres La somme des valeurs propres correspond à la variance totale:

$$Tr(\Sigma) = \sigma^2 = \sum_{i=1}^{p} \lambda_i$$

Chaque valeur propre mesure la part de variance expliquée par l'axe factoriel correspondant.

Choix de la dimension q

La qualité des estimations auxquelles conduit l'A.C.P. dépend, de façon évidente, du choix de q c'est-à-dire du nombre de composantes retenues pour reconstituer les données, ou encore de la dimension du sous-espace de représentations. De nombreux critères de choix pour q ont été proposés dans la littérature.
La "qualité globale" des représentations est mesurée par la part d'inertie expliquée :

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{i=1}^{p} \lambda_i}.$$

5 Analyse factorielle des correspondances

L'AFC s'applique aux tableaux de contingences (tableau croisé de co-occurrence) c'est-à-dire aux tableaux de comptage obtenus par croisement de deux variables nominales. Cette méthode sert à déterminer et à hiérarchiser toutes les dépendances entre les lignes et les colonnes du tableau.

Pour visualiser ou regrouper, il faut identifier des données proches les unes des autres : question centreale du choix d'une distance !

5.1 Choix des distances

La distance euclidienne usuelle entre deux points-lignes exprimée sur le tableau d'effectifs bruts ne ferait que traduire les différences d'effectifs entre deux modalités (par exemple yeux bleus VS yeux verts). En revanche, la distance euclidienne usuelle entre deux profils-lignes traduit bien la ressemblance ou la différence entre 2 modalités (par exemple la différence des cheveux entre 2 couleurs des yeux) sans tenir compte des effectifs totaux de ces modalités (yeux bleus VS yeux verts) :

$$d^2(i,i') : \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Cependant, cette distance favorise les colonnes qui ont une masse importante c'est-à-dire la modalité (la couleur des cheveux) qui est bien rerésentée dans la population étudiée.

Distance χ^2 : on pondère chaque écart par l'inverse de la masse de la colonne et l'on calcule une nouvelle distance appelée *la distance du* χ^2 :

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

On définit de la même manière la distance entre les profils-colonnes par :

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

C'est cette distance pondérée, ainsi que le rôle symétrique joué par les lignes et les colonnes du tableau de contingences, qui particularisent l'analyse des correspondances et lui assurent des propriétés remarquables que ne possède pas l'analyse en composantes principales : l'équivalence distributionnelle et les relations de transition.

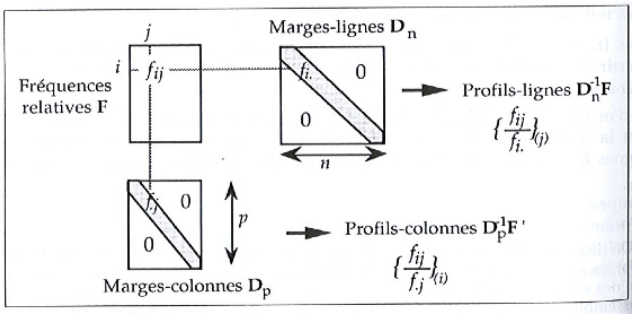
N.B : L'inertie des nuages de points lignes (ou des points colonnes) calculée avec cette distance est proportionnelle au classique χ^2 de Karl Pearson utilisé pour éprouver l'indépendance des lignes et des colonnes d'une table de contingence. D'où le nom de distance du χ^2.

5.2 Méthode

Le but principal de l'AFC reste donc le même que celui de l'ACP : lire l'information contenue dans un espace multidimensionnel par une réduction de la dimension de cet espace tout en conservant un maximum de l'information contenu dans l'espace de départ. Le tableau des données subit deux transformations, l'une en profil-lignes, l'autre en profil-colonnes, à partir desquelles vont être construits les nuages de points dans $\mathbb{R}^p$ et $\mathbb{R}^n$. Les transformations opérées sur le tableau des données peuvent s'écrire à partir des trois matrices F, D_n et D_p qui définissent les éléments de base de l'analyse.

Nuage de n points-lignes dans l'espace $\mathbb{R}^p$	Éléments de base	Nuage de p points-colonnes dans l'espace $\mathbb{R}^n$
$\mathbf{X} = \mathbf{D}_n{}^{-1}\mathbf{F}$ p coordonnées (point-ligne i) $\frac{f_{ij}}{f_{i.}}$, pour $j = 1, 2, \ldots, p$.	Analyse du tableau $\mathbf{X}$	$\mathbf{X} = \mathbf{D}_p{}^{-1}\mathbf{F'}$ n coordonnées (point-colonne j) $\frac{f_{ij}}{f_{.j}}$, pour $i = 1, 2, \ldots, n$.
$\mathbf{M} = \mathbf{D}_p{}^{-1}$ $d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} (\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}})^2$	avec la métrique $\mathbf{M}$	$\mathbf{M} = \mathbf{D}_n{}^{-1}$ $d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} (\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}})^2$
$\mathbf{N} = \mathbf{D}_n$ masse du point i : $f_{i.}$	et le critère $\mathbf{N}$	$\mathbf{N} = \mathbf{D}_p$ masse du point j : $f_{.j}$

Table 1: Elements de base de l'analyse



Fréquences, marges, profils

F de dimensions $n \times p$ désigne le tableau des fréquences relatives;
D_n de dimensions $n \times n$ est la matrice diagonale dont les éléments diagonaux sont les marges en lignes $f_{i.}$;
D_p de dimensions $p \times p$ est la matrice diagonale des marges en colonnes $f_{.j}$.

Les masses dans un espace sont liées aux pondérations dans l'autre.

Dans $\mathbb{R}^p$	Éléments de Construction	Dans $\mathbb{R}^n$
$\mathbf{S} = \mathbf{F'D}_n{}^{-1}\mathbf{FD}_p^{-1}$	Matrice à diagonaliser	$\mathbf{T} = \mathbf{FD}_p{}^{-1}\mathbf{F'D}_n^{-1}$
$\mathbf{S}u_\alpha = \lambda_\alpha u_\alpha$	Axe factoriel	$\mathbf{T}v_\alpha = \lambda_\alpha v_\alpha$
$\psi_\alpha = \mathbf{D}_n^{-1}\mathbf{FD}_p^{-1}u_\alpha$	Coordonnées	$\phi_\alpha = \mathbf{D}_p{}^{-1}\mathbf{F'D}_n^{-1}v_\alpha$
$\psi_{\alpha i} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}f_{.j}} u_{\alpha j}$	factorielles	$\phi_{\alpha j} = \sum_{i=1}^{n} \frac{f_{ij}}{f_{i.}f_{.j}} v_{\alpha i}$

Table 2: Elements de construction de l'analyse

L'AFC peut être considérée comme une ACP particulière dotée de la métrique du χ^2 qui ne dépend que du profil des colonnes du tableau.

Si on se place dans l'espace des colonnes $\mathbb{R}^p$ (manipulation des nuages de points lignes), on peut chercher l'axe (factoriel) d'inertie maximum du nuages passant par l'origine O et engendré par un vecteur unitaire u pour la pondération (métrique) D_p^{-1} :
Maximisation de la somme pondérée des carrés des projetctions sur l'axe :

$$\max_u \sum_i f_{i.} d^2(i, 0)$$

ce qui revient à rendre maximal :

$$u^T D_p^{-1} F^T D_n^{-1} F D_p^{-1} u$$

avec $u^T D_p^{-1} u = 1$. et u vecteur propre de $S = F^T D_n^{-1} F D_p^{-1}$ associé à la plus grande valeur propre $\lambda \neq 1$.

La matrice S à diagonaliser est de terme général :

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j'}}$$