A F C

L'Analyse factoreille des Correspondances est une méthode statistique que permet de révelers des liens possibles entre différentes données présenté »es dans un tableau statistiques à double entrées comme ceux qui sont utilisés dans les études sociales ou commerciales. Vous trouverez sur le web des présentations théoriques qui vous expliquent cette méthode d'analyse. Pour expliquer le but de cette méthode je prend un exemple simple.

Partons d'un exemple simple proposé comme exemple dans cet espace. celui portant sur l'orientation des étudiants entre les filières de l'enseignement supérieur en fonction du type de bac obtenu. L'exemple est publié par Philippe CIBOIS dans la publications des Presses Universitaires de France (PUF – Juillet 1991) intitulé « L'ANALYSE FACTORIELLE ».

Les données collectées se présentent sous forme d'un tableau à double entrées En colonnes sont représentées les filières de l'enseignement supérieur : Université, Classe Préparatoire aux Grandes Ecoles (CPGE) et Autres filières. En lignes sont représentés les types de bac : Lettres, Economique et social, Scientifiques et enfin techniques.

	Universite	**CPGE**	**Autre**
Lettres	1300	2300	500
Eco et sociale	2000	200	800
Scientifiques	1000	500	500
Technique	700	100	2200

Le but est d'identifier l'orientation préférée des étudiants de ces filières en fonction du type de bac. Nous constatons ici que le lien entre les bacs techniques et la filière autre (avec 2200 étudiants) est évident, l'analyse va d'ailleurs le révéler de manière significative. Mais ces liens entre les caractéristiques (ou attributs présentés en lignes) et les variables présentées en colonnes ne se voient pas toujours aussi aisement. Ce cas peut même figurer comme une exception car cette « attraction » - « répulsion » entre les lignes et les colonnes n'est pas toujours visible directement dans le tableau.

Le principe de la méthode consiste a tenter d'extraire des facteurs qui permettent de résumer le tableau. Ce dernier sera décomposé en plusieurs tableaux qui se complètent. C'est le principe de la mise en facteur utilisée en algébre.

T = > Tableau des observations.
T0 => Tableau qui représente une répartition des données entre les lignes et les colonnes si aucune « attraction » ni « répulsion » n'existe. Une sorte de tableau neutre (appelé aussi inertie).

Les écarts (EC1) entre T et T0 forment un tableau qui est soumis à l'algorythme d'extraction des vecteurs ligne et colonne qui peuvent reconstituer au plus près ce tableau des écarts.

Ce tableau des écarts est retravaillé de manière à pondérer ses données pour tenir compte du poids relatif de chaque ligne et chaque colonne. L'appréciation des écarts entre la distribution « idéale » et la distribution réelle ainsi réajustés permet d'obtenir le poids total des écarts entre le tableau des observations réélles et la tableau neutre. Dans l'exemple le poids total est de 2491,7 soit 24,9 % des données obervées (2491, 7 / 10 000)

Les vecteurs ainsi obtenus permettent de reconstituer un tableau T1. Les données de ces vecteurs ligne et colonne forment le premier facteur. Pour information et pour les intéressés, le critère de fin des calculs (la convergence) est la rapprochement entre la norme du vecteur ligne et celle du vecteur colonne. La convergencve est obtenue quand les deux normes sont éagles ou que leur différence est inférieurs au seuil fixé, par défaut fixé ici à 0,05. Il est modifiable.

Un nouveau tableau des écarts (EC1) est obtenu en soustrayant le tableau ainsi reconstitué du tableau des éacrts initial (ECGlobal). Ce second tableau des écarts est soumis au même traitement que le premier tableau des écarts pour obtenir son poids relatif dans l'écart global. Le poids de ce tableau EC1 est 1997,9 soit 80 % des 2491,7 du tableau EC Global

Ce second tableau des écarts est soumis à son tour au calcul d'extration des vecteurs qui peuvent le reconstituer au plus près. A l'issue de ces calculs itératifs les vecteurs ainsi obtenus peuvent recontituer un tableau T2 sensé reconstituer au plus près le tableau des écarts EC1. Si le tableau T2 n'est pas identique au tableau EC1 alors la reconstituion est incomplète car il reste un résidu. Le poids de ce tableau des écarts (EC2) est 493,3 soit 20 % de EC Global.

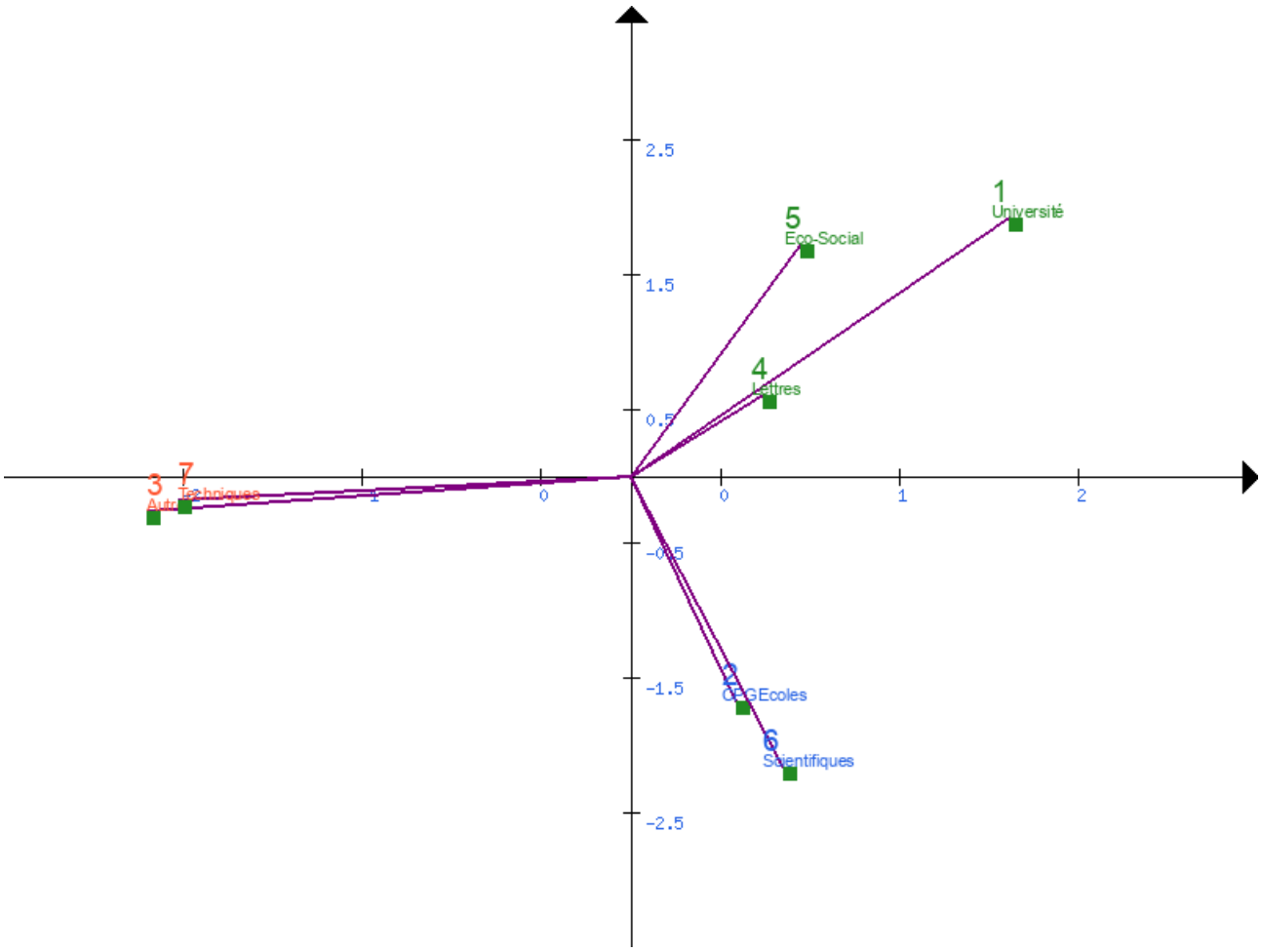
Les 2 facteurs reconstituent l'intégralité des écarts. Nous avons donc T = T0 + T1 + T2 + 0

EC Global => 2491,7
EC1 => 1997,9
EC2 => 493,3

La somme de 1997,9 et 493,3 est 2 491,2 . La différence avec le poids de ECGlobal s'explique par les arrondis (des décimales retenues dans les calculs).

En théorie lorsque l'écart n'est par reconstitué avec les deux premiers facteurs il faudrait poursuivre les les calculs d'extraction des vecteurs qui reconstituent le nouvel écart ainsi obtenu. Cependant si les deux premiers facteurs ne reconstituent pas correctement les tableaux des écarts c'est sans doute que cette analyse n'est pas pertinente pour le cas de figure étudié. L'interprétation des résultats devra en tenir compte quoiqu'il en soit.

Je me contente des deux facteurs qui permettent une représentation graphique en 2 dimensions. Un axe horizontal sur lequel seont représentées les valeurs du premier facteur. Un axe vertical sur lequel seront représentées les données du second facteur. La combinaison du facteur 1 avec le facteur 2 donne les coordonnées de chaque ligne et chaque colonne sur ce plan à 2 dimensions.



Le graphique montre bien cette attraction les bacheliers techniques (7) et la filière technique (3) – De même que les bachliers scientifiques s'orientent davantage vers les Classe préparatoires des Grandes Ecoles, Les lignes et les colonnes ont été numérotées par commodité.