Chapitre 3

Analyse Factorielle des Correspondances

3.1 DONNÉES, NOTATIONS, HYPOTHÈSE D'INDÉPENDANCE

À l'origine, l'Analyse Factorielle des Correspondances (AFC) a été conçue pour étudier des tableaux appelés couramment tableaux de contingence (ou tableaux croisés). Il s'agit de tableaux d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de n individus. Dans l'exemple commenté au chapitre 10, la population est constituée par l'ensemble des individus qui ont quitté le système scolaire français en 1972 et qui occupent un emploi en 1973 ; pour chaque individu, on connaît son niveau de diplôme et sa catégorie d'emploi. La **figure 3.1** résume les principales notations.

On parle indifféremment de la modalité i (par exemple le baccalauréat) ou de la classe i, c'est-à-dire de la classe des individus qui possèdent la modalité i (par exemple les bacheliers).

Dans ce chapitre, nous nous limitons à l'étude d'un tableau de contingence. Cependant, la plupart des notions introduites et des résultats présentés peuvent être généralisés à des tableaux qui ne sont pas strictement de ce type. Le cas très important du tableau disjonctif complet fait l'objet d'un chapitre particulier : l'Analyse des Correspondances Multiples. La conclusion du présent chapitre donne quelques points de repère sur l'application de l'AFC à d'autres tableaux que les tableaux de contingence.

On considère souvent le tableau des fréquences relatives F, obtenu en divisant chaque effectif k_{ij} par l'effectif total n. Ce nouveau tableau définit une mesure de probabilité sur l'ensemble produit $I \times J$. Ses marges, ou probabilités marginales,

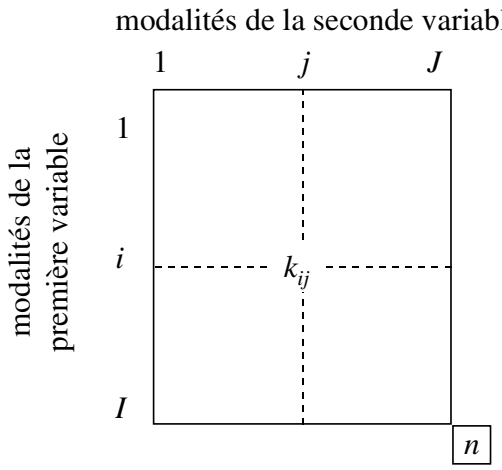
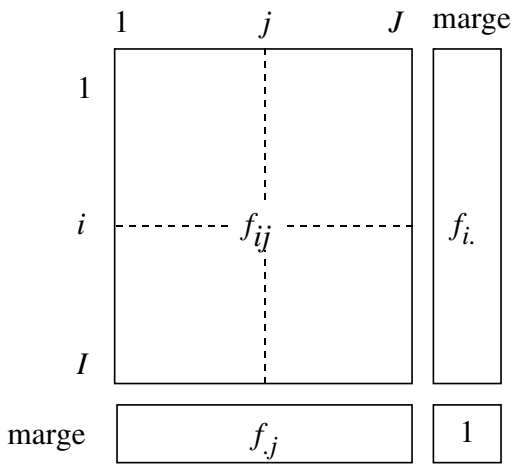


Figure 3.1 Tableau des données brutes. I : ensemble des lignes et nombre de lignes (8 niveaux de diplôme). J : ensemble des colonnes et nombre de colonnes (9 catégories d'emploi). k_{ij} : nombre d'individus possédant à la fois la modalité i de la première variable et la modalité j de la seconde (i.e. qui ont le niveau de diplôme i et qui occupent un emploi de la catégorie j). $\sum_i \sum_j k_{ij} = n$ (nombre total d'individus).



$$
\begin{aligned}
f_{ij} &= k_{ij}/n \\
f_{i.} &= \sum_j f_{ij} \\
f_{.j} &= \sum_i f_{ij} \\
\sum_i f_{i.} &= \sum_j f_{.j} = \sum_i \sum_j f_{ij} = 1
\end{aligned}
$$

Figure 3.2 Tableau F des fréquences relatives et ses marges.

ont pour terme général $f_{i.}$ pour la marge-colonne et $f_{.j}$ pour la marge-ligne (*cf.* **Figure 3.2**).

Un tableau de contingence exprime la liaison entre deux variables qualitatives. Classiquement, pour une mesure de probabilité, on dit qu'il y a **indépendance** entre les deux variables lorsque, pour tout i et pour tout j, on a l'égalité :

$$f_{ij} = f_{i.} f_{.j}$$

Il y a **liaison** entre les deux variables dès que certaines cases du tableau f_{ij} diffèrent du produit $f_{i.} f_{.j}$. Si f_{ij} est supérieur à ce produit, les modalités i et j s'associent plus qu'elles ne le font dans l'hypothèse d'indépendance : on dit que i et j s'attirent. Au contraire, si f_{ij} est inférieur au produit des marges, i et j s'associent moins que dans l'hypothèse d'indépendance : on dit qu'il y a répulsion entre ces deux modalités.

L'**indépendance** s'exprime aussi en considérant le tableau comme un ensemble de lignes. En effet, l'égalité ci-dessus est équivalente à l'égalité :

$$\frac{f_{ij}}{f_{i.}} = f_{.j}$$

La quantité $f_{.j}$ représente le pourcentage de la population totale qui possède la modalité j tandis que $f_{ij}/f_{i.}$ représente ce même pourcentage dans la sous-population possédant la modalité i. Lorsqu'il y a indépendance, les I sous-populations caractérisées par les modalités i de la première variable se répartissent selon les J modalités j de la deuxième variable avec les mêmes pourcentages. Toutes les lignes sont alors proportionnelles. La réciproque est vraie : lorsque toutes les lignes sont proportionnelles, elles sont proportionnelles à la marge $f_{.j}$ et les deux variables sont indépendantes. Il y a donc **liaison** dès lors que les lignes ne sont pas toutes proportionnelles à la marge, c'est-à-dire lorsqu'elles ne sont pas identiques du point de vue de leur association avec l'ensemble des colonnes.

Remarquons enfin que, dans un tableau de contingence, les lignes et les colonnes jouent un rôle absolument symétrique : l'indépendance s'exprime de la même façon sur l'ensemble des colonnes. Les deux égalités ci-dessus sont en effet équivalentes à la suivante :

$$\frac{f_{ij}}{f_{.j}} = f_{\mathrm{i.}}$$

Il y a indépendance lorsque tous les pourcentages en colonnes sont égaux à la marge $f_{i.}$ c'est-à-dire lorsque les colonnes sont proportionnelles. Il y a liaison lorsqu'elles ne le sont pas.

3.2 OBJECTIFS

Bien que le tableau étudié soit de nature très différente de celui étudié en ACP, les objectifs de l'AFC peuvent s'exprimer de manière analogue à ceux de l'ACP : on cherche à obtenir une typologie des lignes, une typologie des colonnes et à relier ces deux typologies entre elles ; mais la notion de ressemblance entre deux lignes, ou entre deux colonnes, est différente de celle de l'ACP.

Dans un tableau de contingence, la ressemblance, entre deux lignes d'une part et entre deux colonnes d'autre part, s'exprime de manière totalement symétrique. Deux lignes sont considérées comme proches si elles s'associent de la même façon à l'ensemble des colonnes, c'est-à-dire si elles s'associent trop (ou trop peu) aux mêmes colonnes ; les termes « trop » et « trop peu » sont pris en référence à la situation d'indépendance. Symétriquement, deux colonnes sont proches si elles s'associent de la même façon à l'ensemble des lignes.

Schématiquement, l'étude de l'ensemble des lignes revient à mettre en évidence une typologie dans laquelle on cherche les lignes dont la répartition s'écarte le plus de celle de l'ensemble de la population, celles qui se ressemblent entre elles (dans le sens précisé ci-dessus) et celles qui s'opposent. Pour mettre en relation la typologie des lignes avec l'ensemble des colonnes, on caractérise chaque groupe de lignes par les colonnes auxquelles ce groupe s'associe trop ou trop peu.

L'étude de l'ensemble des colonnes est absolument analogue.

Cette approche, grâce à la notion de ressemblance utilisée, permet d'étudier la liaison entre les deux variables, c'est-à-dire l'écart du tableau à l'hypothèse d'indépendance. L'analyse de cette liaison est l'objectif fondamental de l'AFC.

Une approche complémentaire de la précédente, fait intervenir conjointement l'ensemble des lignes et celui des colonnes en ne privilégiant ni l'un ni l'autre. Prenons l'exemple du tableau croisant les catégories d'emploi et les niveaux de diplôme. L'ensemble des diplômes est ordonné par la longueur des études tandis que celui des catégories d'emploi l'est par le salaire moyen. La relation entre ces deux ordres (un salaire élevé correspond généralement à un diplôme élevé) explique clairement une bonne part de la liaison entre emplois et diplômes. Mais ce lien ne se restreint peut-être pas à cet unique aspect ; il peut exister d'autres phénomènes comme l'association presque exclusive de certains diplômes avec certains emplois. L'objectif de l'AFC est de décomposer la liaison entre deux variables en une somme (ou une superposition) de tendances simples et interprétables comme celles qui viennent d'être évoquées et de mesurer leur importance relative afin de les ordonner.

Enfin, bien qu'il y soit fait peu référence par la suite, il faut signaler que l'AFC, comme toute Analyse Factorielle, est utilisée aussi dans le but de réduire la dimension des données en conservant le plus d'information possible. Ceci en vue d'un traitement statistique ultérieur (classification, régression, analyse discriminante, etc.) ou d'une transmission d'information.

3.3 TRANSFORMATIONS DES DONNÉES EN PROFILS

En AFC, le tableau brut n'est pas analysé directement. Dans l'étude des lignes, le tableau des données est transformé en divisant chaque terme f_{ij} de la ligne i par la marge $f_{i.}$ de cette ligne i. La nouvelle ligne est appelée profil-ligne (*cf.* **Figure 3.3**).

Cette transformation découle de l'objectif qui vise à étudier la liaison entre les deux variables au travers de l'écart entre les pourcentages en lignes. Elle se justifie aussi de façon directe puisque la comparaison de deux lignes du tableau brut risque d'être influencée principalement par leurs effectifs marginaux. Ainsi, dans le tableau croisant emplois et diplômes, la différence entre les lignes brutes *Bac technique* et *Bac général* traduit essentiellement une différence entre les effectifs globaux de ces deux diplômes.

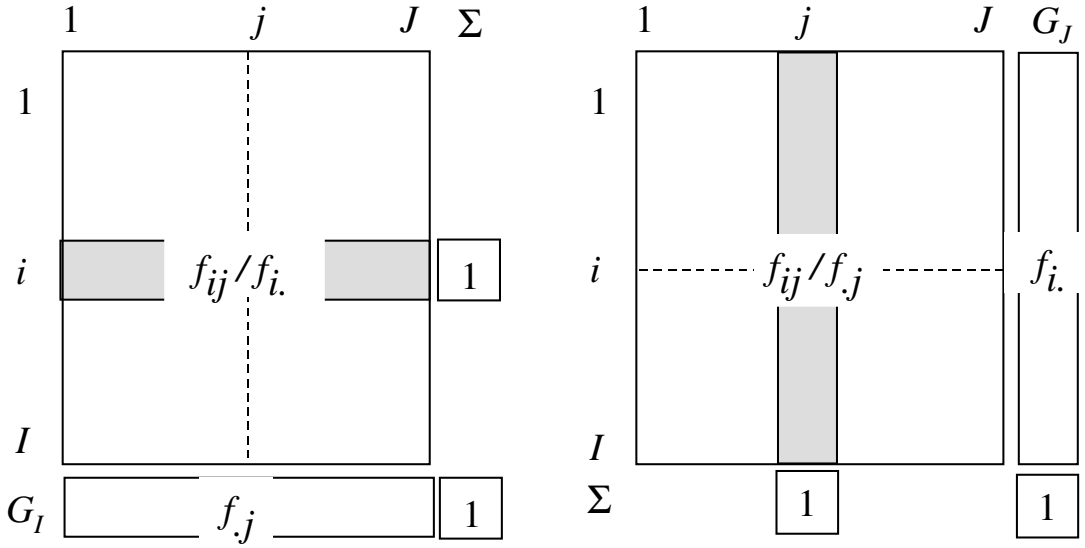


Figure 3.3 Profil-ligne (à gauche) et profil-colonne (à droite). G_I et G_J : profils marginaux.

Le nombre $f_{ij}/f_{i.}$ représente, dans notre exemple, la probabilité d'occuper un emploi de la catégorie j sachant que l'on détient le niveau de diplôme i. Le profil-ligne i n'est rien d'autre que la loi de probabilité conditionnelle définie par i sur l'ensemble des colonnes. Pour analyser l'écart à l'indépendance, on confronte ces profils au profil ligne marginal (= établi sur l'ensemble de la population) de terme général $f_{.j}$ et noté G_I.

Du fait du rôle symétrique joué par les lignes et les colonnes, un raisonnement analogue peut être mené à propos des colonnes. Il conduit à la notion de profil-colonne (*cf.* **Figure 3.3**).

Ainsi, en AFC, selon que l'on s'intéresse aux lignes ou aux colonnes, on ne considère pas le même tableau transformé. Toutefois, les deux transformations en profils possèdent la même signification vis-à-vis des objets qu'elles concernent. Ces transformations sont intéressantes en elles-mêmes indépendamment de tout contexte d'analyse factorielle. Lorsqu'un tableau croisé est commenté, il est presque toujours présenté sous la forme de pourcentages, par rapport aux lignes ou aux colonnes selon les aspects que l'on cherche à mettre en évidence.

3.4 RESSEMBLANCE ENTRE PROFILS : DISTANCE DU χ^2

En AFC, la ressemblance entre deux lignes ou entre deux colonnes est définie par une distance entre leurs profils connue sous le nom de distance du χ^2. Elle est définie de façon symétrique pour les lignes et pour les colonnes. Soit :

$$\begin{aligned} \mathrm{d}\chi^2(\text{profil-ligne } i, \text{profil-ligne } l) &= \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2 \\ \mathrm{d}\chi^2(\text{profil-colonne } j, \text{profil-colonne } k) &= \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ik}}{f_{.k}} \right)^2 \end{aligned}$$

Dans ces relations, la distance entre deux lignes dépend essentiellement des différences terme à terme entre les deux profils dont elle fait une somme des carrés pondérés. La pondération $1/f_{.j}$ équilibre l'influence des colonnes sur la distance entre les lignes : elle augmente les termes, *a priori* plus faibles, concernant les modalités rares ; elle joue, jusqu'à un certain point, un rôle analogue à celui de la division par l'écart-type dans le cas des variables numériques.

La distance du χ^2 jouit d'une propriété fondamentale appelée **équivalence distributionnelle**. Selon cette propriété, si deux colonnes proportionnelles d'un tableau sont cumulées en une seule, la distance entre les profils-lignes est inchangée. Le cas d'une proportionnalité parfaite entre deux colonnes ne se rencontre guère en pratique mais constitue une situation limite dont on peut être assez proche. La propriété mathématique est alors utilisée sous la forme d'une règle pragmatique : remplacer, par leur somme, deux colonnes ou deux lignes presque proportionnelles ne modifie pas sensiblement les résultats d'une AFC. On se réfère surtout à cette règle lorsque plusieurs ensembles de modalités sont possibles pour définir une même variable. Ainsi, la variable *catégorie d'emploi* peut être plus ou moins détaillée : par exemple, on peut se demander si les catégories *ouvrier qualifié* et *ouvrier non qualifié* peuvent être regroupées en une seule catégorie. Du fait de l'équivalence distributionnelle, si ces deux catégories ont des profils voisins, le choix entre les deux solutions n'est pas fondamental puisque les AFC des deux tableaux aboutissent à des résultats analogues.

3.5 LES DEUX NUAGES

3.5.1 Nuage des profils-lignes

S'intéresser aux modalités de la première variable revient à considérer les données comme une juxtaposition de profils-lignes. Chaque profil-ligne est une suite de J valeurs numériques et peut être représenté par un point de l'espace R^J dont chacune des J dimensions est associée à une modalité de la seconde variable. La distance

du χ^2 définissant la ressemblance entre profils-lignes (*cf.* section 3.4) possède les propriétés d'une distance euclidienne et confère à R^J la structure d'espace euclidien. Cette distance revient à affecter le poids $1/f_{.j}$ à la j^{e} dimension de R^J. La somme des coordonnées de chaque profil-ligne vaut 1 ; il en résulte que le nuage N_I appartient à un hyperplan, noté H_I (*cf.* **Figure 3.4**).

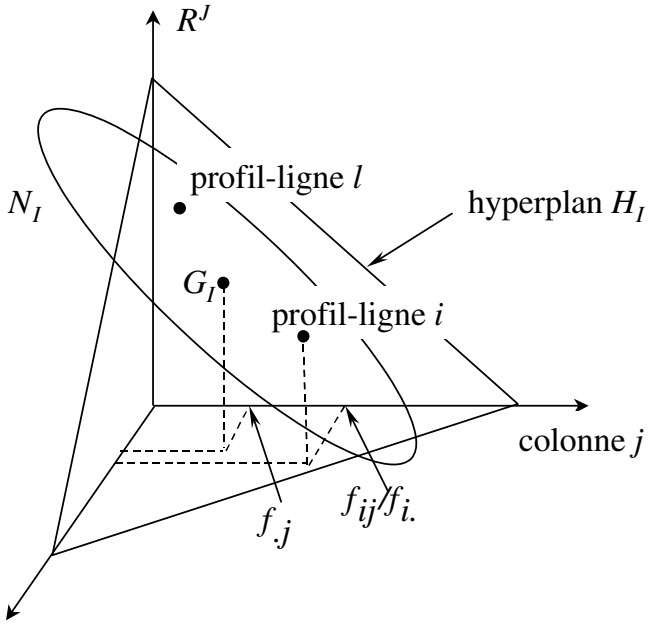


Figure 3.4 Le nuage N_I des profils-lignes dans R^J. Le point i a pour coordonnée sur l'axe j : $f_{ij}/f_{i.}$; son poids est $f_{i.}$; la distance entre deux profils est la distance du χ^2 ; Le barycentre G_I du nuage N_I a pour coordonnée sur l'axe j la fréquence marginale $f_{.j}$; le nuage N_I appartient à un hyperplan noté H_I.

En AFC, les poids affectés à chaque point du nuage sont imposés. Le point i a un poids égal à la fréquence marginale $f_{i.}$ (ce poids est proportionnel à l'effectif de la classe d'individus représentée par le point i).

Le **barycentre** des points de N_I munis de ces poids est noté G_I. Sa j^{e} coordonnée est égale à la fréquence marginale $f_{.j}$.

$$f_{.j} = \sum_i f_{i.} \frac{f_{ij}}{f_{i.}}$$

Il s'interprète comme un profil moyen. Dans l'exemple du tableau qui croise les niveaux de diplôme et les catégories d'emploi, G_I est le profil d'emplois de l'ensemble de la population, tous les diplômes étant cumulés. Il sert constamment de référence dans l'étude des lignes du tableau ; ainsi, étudier dans quelle mesure et de quelle façon une classe d'individus i diffère de l'ensemble de la population revient à étudier l'écart entre le profil de cette classe i et le profil moyen. Étudier la dispersion du nuage

autour de son barycentre revient à étudier l'écart entre les profils des lignes et le profil marginal, et donc la liaison entre les deux variables (*cf.* section 3.1).

3.5.2 Nuage des profils-colonnes

Compte tenu du rôle symétrique joué par les lignes et les colonnes en AFC, la construction du nuage des profils-colonnes s'effectue selon une démarche strictement identique à celle du nuage des profils-lignes. Il est toutefois utile de la décrire, ne serait-ce que pour fixer les notations.

S'intéresser aux modalités de la seconde variable revient à considérer les données comme une juxtaposition de profils-colonnes. Chaque profil-colonne est une suite de I valeurs numériques et peut être représenté par un point de l'espace R^I dont chacune des dimensions est associée à une modalité de la première variable. R^I est muni d'une structure euclidienne par la distance du χ^2 : à la i^{e} dimension on affecte le poids $1/f_{i.}$ (*cf.* **Figure 3.5**).

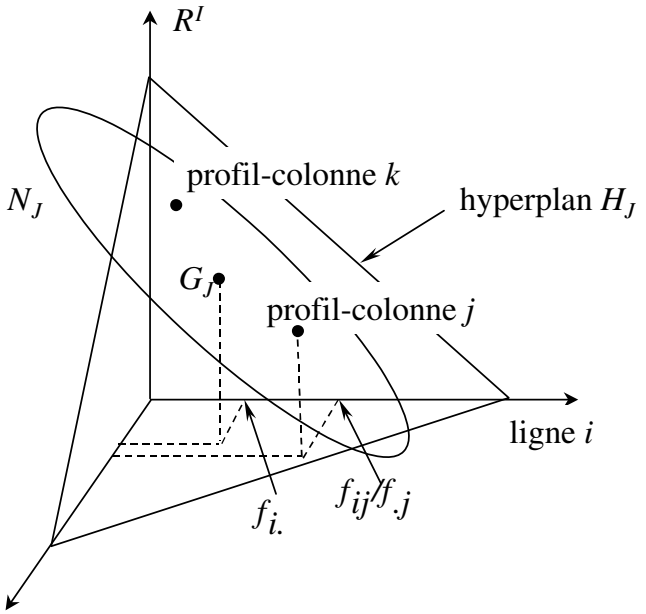


Figure 3.5 Le nuage N_J des profils-colonnes dans R^I. Le point j a pour coordonnée sur l'axe i : $f_{ij}/f_{.j}$; son poids est $f_{.j}$; la distance entre deux profils est la distance du χ^2 ; le barycentre G_J du nuage N_J a pour coordonnée sur l'axe i la fréquence marginale $f_{i.}$; le nuage N_J appartient à un hyperplan noté H_J.

Le point G_J représente la marge $\{f_{i.} | i = 1, ..., I\}$; c'est le barycentre de N_J lorsque l'on munit chaque profil-colonne j du poids $f_{.j}$; en tant que profil moyen, il sert constamment de référence dans l'étude de N_J.

3.6 AJUSTEMENT DES DEUX NUAGES

3.6.1 Ajustement du nuage des profils-lignes

Dans R^J, l'ajustement vise à obtenir une suite d'images planes approchées du nuage N_I. De la même façon que l'ACP, l'AFC procède en recherchant une suite d'axes orthogonaux sur lesquels le nuage N_I est projeté. Chaque axe possède la propriété de rendre maximum l'inertie projetée du nuage N_I avec la contrainte d'être orthogonal aux axes déjà trouvés.

Les images planes de N_I doivent être telles que les distances entre les points de l'image ressemblent le plus possible aux distances entre les points de N_I. Cet objectif est tout à fait analogue à celui de l'ajustement du nuage des individus en ACP : pratiquement, il implique que le nuage analysé soit centré, c'est-à-dire que son barycentre soit choisi comme origine des axes (*cf.* section 3.5).

Dans le nuage centré, la classe définie par la modalité i est représentée par un point dont la coordonnée sur le j^{e} axe vaut : $f_{ij}/f_{i.} - f_{.j}$. La position de ce point exprime la différence entre la répartition, sur l'ensemble des modalités de la seconde variable, des individus de la classe i et celle de la population totale. Ainsi, rechercher les directions d'inertie maximum du nuage centré revient à mettre en évidence les classes qui s'écartent le plus du profil de l'ensemble de la population.

Chaque profil est muni d'un poids égal à sa fréquence marginale $f_{i.}$. Ce poids intervient en premier lieu dans le calcul du barycentre du nuage. Il intervient aussi dans l'inertie et donc dans le critère d'ajustement satisfait par les axes (*cf.* **Figure 3.6**).

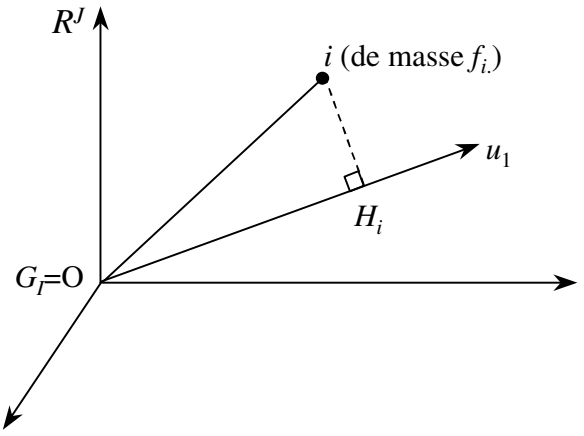


Figure 3.6 Ajustement dans R^J du nuage des profils-lignes. i : point associé au profil-ligne i. u_1 : vecteur unitaire du premier axe factoriel. H_i : projection de i sur u_1. u_1 rend maximum $\sum_i f_{i.}\text{O}H_i^2$.

Du fait de l'introduction des poids $f_{i.}$ dans le critère d'ajustement, chaque modalité possède un poids proportionnel à la population qu'elle représente. Ainsi, à disparité de profil égale, les axes factoriels mettent plutôt en évidence des phénomènes concernant

une fraction importante de la population totale. Selon un autre point de vue, les modalités d'effectif faible, pour lesquelles les profils risquent d'être moins fiables, interviennent moins dans la construction des axes.

En résumé, l'ajustement du nuage N_I en AFC est analogue à celui du nuage des individus en ACP. Il en diffère par trois points :

1. les lignes interviennent au travers de leur profil ;
2. la distance entre les profils est celle du χ^2 ;
3. chaque ligne i est affectée du poids $f_{i.}$.

3.6.2 Ajustement du nuage des profils-colonnes.

Du fait du rôle symétrique joué par les lignes et les colonnes en AFC, l'ajustement de N_J dans R^I se pose dans les mêmes termes et possède les mêmes propriétés que l'ajustement de N_I dans R^J. Nous les résumons ci-dessous.

1. Les images planes de N_J doivent être telles que les distances entre les profils projetés ressemblent le plus possible aux distances entre les profils dans R^I. Il en résulte la nécessité d'analyser le nuage N_J par rapport à son barycentre G_J. L'inertie totale de N_J par rapport à G_J provient des différences entre les profils des différentes classes j et le profil de l'ensemble de la population.
2. Chaque colonne j est affectée d'un poids égal à sa fréquence marginale $f_{.j}$. Avec des notations analogues à celles de la **figure 3.6**, H_j étant la projection sur v_1 (vecteur unitaire du premier axe factoriel dans R^I) du point j associé au profil-colonne j, v_1 rend maximum la quantité : $\sum_j f_{.j}(\mathrm{O}H_j)^2$. La justification de ce poids $f_{.j}$ est strictement analogue à celle développée à propos des profils-lignes.

3.6.3 Un aspect technique du centrage en AFC

Du point de vue technique, on peut montrer (*cf.* **section 5.5 page 121**) qu'il n'est pas nécessaire de centrer explicitement le nuage N_I avant de l'analyser. En effet, mis à part le premier facteur, l'analyse du nuage par rapport à O sans centrage conduit aux mêmes facteurs que l'analyse du nuage centré.

Lorsque l'on réalise l'AFC du nuage N_I non centré (c'est-à-dire par rapport à l'origine O sans centrage), le premier axe factoriel possède les propriétés suivantes (*cf.* **Figure 3.7**) :

1. il relie l'origine O au barycentre G_I du nuage N_I ;
2. cet axe est orthogonal, au sens de la distance utilisée (i.e. distance du χ^2), à l'hyperplan H_I contenant le nuage N_I;
3. l'inertie projetée de N_I dans cette direction vaut 1.

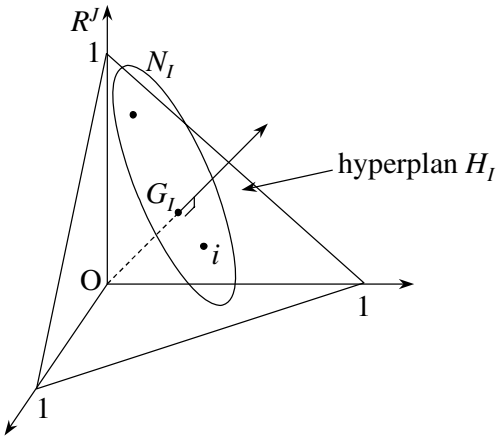


Figure 3.7 Le premier axe factoriel du nuage N_I non centré est le facteur trivial OG_I orthogonal à H_I. L'inertie projetée de N_I sur OG_I vaut 1.

Naturellement, cet axe ne présente pas d'intérêt en lui-même : la projection sur OG_I de chaque point de N_I est confondue avec G_I. Cette projection de N_I sur l'axe OG_I est appelée **facteur trivial** ou facteur constant.

L'orthogonalité du premier axe OG_I avec l'hyperplan H_I présente une conséquence importante. Les axes suivants étant par définition orthogonaux à OG_I, l'analyse peut être poursuivie indifféremment par rapport à O ou à G_I (*cf.* **Figure 3.8**).

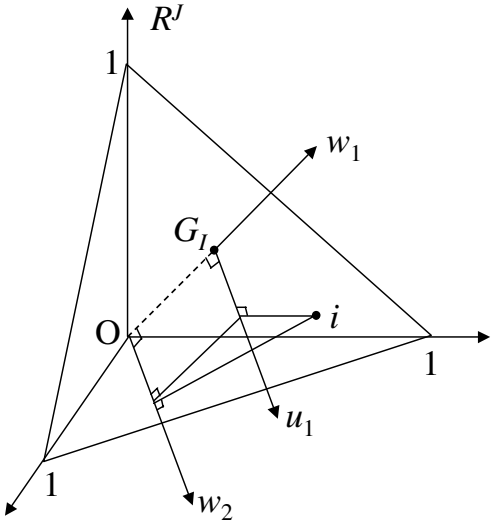


Figure 3.8 Analyse par rapport au barycentre et par rapport à l'origine. w_1 : premier axe factoriel du nuage N_I lorsque l'origine des axes est en O. w_2 : deuxième axe factoriel du nuage N_I lorsque l'origine des axes est en O (orthogonal à u_1). u_1 : premier axe factoriel du nuage N_I lorsque l'origine des axes est en G_I. Les projections de N_I sur w_2 et u_1 sont identiques.

3.7 LA DUALITÉ

Les deux nuages N_I et N_J constituent deux représentations d'un même tableau, l'une à travers ses profils-lignes, l'autre à travers ses profils-colonnes. Il s'ensuit que les analyses de ces deux nuages ne sont pas indépendantes : les relations entre ces deux analyses sont communément regroupées sous le terme de dualité.

Cette dualité est plus fondamentale et plus riche en AFC qu'en ACP car les lignes et les colonnes représentent des objets de même nature, ce qui n'est pas le cas en ACP.

3.7.1 Statistique du χ^2 et inertie des deux nuages N_I et N_J

Lorsque l'on étudie un tableau de contingence, c'est-à-dire une population de n individus au travers de deux variables qualitatives, il est classique de mesurer la significativité de la liaison entre ces deux variables à l'aide de la statistique χ^2. Appliquée à un tableau d'effectifs, cette statistique mesure l'écart entre les effectifs observés et les effectifs théoriques que l'on obtiendrait en moyenne si les deux variables étaient indépendantes. Elle s'écrit :

$$\chi^2 = \sum_{ij} \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}} = \sum_{ij} \frac{(n\ f_{ij} - n\ f_{i.} f_{.j})^2}{n\ f_{i.} f_{.j}}$$

La statistique χ^2 est égale, au coefficient n près, à l'inertie totale par rapport à leur barycentre de l'un ou l'autre des nuages N_I et N_J. En effet, dans R^I, l'inertie totale de N_I par rapport à G_I s'écrit :

$$\text{Inertie}(N_I) = \sum_i \text{Inertie}(i) = \sum_i f_{i.} d^2(i, G_I) = \sum_i f_{i.} \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$$

Soit :

$$\chi^2 = n[\text{Inertie}(N_I)] = n[\text{Inertie}(N_J)]$$

Cette double égalité montre que l'inertie totale de chacun des deux nuages N_I et N_J représente, sous deux formes différentes, la liaison entre les deux variables.

> ***Remarque :*** La quantité χ^2/n , notée Φ^2, mesure l'intensité de la liaison entre deux variables qualitatives (cette liaison est d'autant plus intense que les modalités de l'une s'associent exclusivement aux modalités de l'autre) et non sa significativité (elle ne dépend pas de l'effectif total) ; l'indicateur χ^2 , lui, mesure la significativité (une liaison forte peut ne pas être significative si elle est observée sur très peu d'individus ; une liaison faible peut être significative si elle est observée sur beaucoup d'individus).

3.7.2 Dualité entre les facteurs sur *I* et les facteurs sur *J*

De même qu'en ACP, on appelle *facteur* l'ensemble des coordonnées des projections des points d'un nuage sur l'un de ses axes factoriels ; les facteurs sur les lignes sont les projections de N_I et les facteurs sur les colonnes les projections de N_J. Le rang d'un facteur est le rang de l'axe factoriel correspondant. Outre leur inertie totale identique, les nuages N_I et N_J possèdent une propriété remarquable : leur ajustement conduit à deux suites de facteurs « duaux ». Plus précisément, nous montrons au chapitre 5 que :

1. les inerties associées aux axes de même rang dans chacun des nuages sont égales ;
2. les facteurs (de même rang) sur les lignes et ceux sur les colonnes sont liés par des relations dites de transition (elles permettent de transiter de R^I dans R^J et inversement).

Les deux paragraphes suivants détaillent cette dualité dont la conséquence essentielle est la suivante : les facteurs sur I et sur J de même rang doivent être interprétés conjointement car ils mettent en évidence la même part de liaison, exprimée pour l'un en termes de profils-lignes et pour l'autre en termes de profils-colonnes.

a) Relations de transition

Les formules de transition précisent les relations entre les points représentant d'une part les lignes et d'autre part les colonnes. Avec les notations suivantes :

1. $F_s(i)$: projection de la ligne i sur l'axe de rang s de N_I,
2. $G_s(j)$: projection de la colonne j sur l'axe de rang s de N_J,
3. λ_s : valeur commune de l'inertie associée à chacun de ces deux axes,

les deux relations de transition s'écrivent :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j)$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{.j}} F_s(i)$$

Ces deux propriétés, qui expriment les résultats de l'analyse d'un nuage en fonction des résultats de l'analyse de l'autre nuage, conduisent à une économie de calcul. Mais surtout, elles donnent un sens à une représentation simultanée des lignes et des colonnes.

b) Représentation simultanée des lignes et des colonnes ; relations barycentriques

La représentation simultanée s'obtient en superposant les projections de chacun des deux nuages N_I et N_J sur des plans engendrés par des axes de même rang pour

les deux nuages. Sur les graphiques ainsi obtenus, les rapports entre la position des points lignes et des points colonnes dus aux relations de transition peuvent être décrits ainsi : au coefficient $1/\sqrt{\lambda_s}$ près, la projection, notée $F_s(i)$, de la ligne i sur l'axe de rang s (dans R^J) est le barycentre des projections, notées $G_s(j)$, des colonnes j sur l'axe de rang s (dans R^I), chaque colonne j étant affectée du poids $f_{ij}/f_{i.}$ (cette expression d'une formule de transition est appelée propriété barycentrique). Les éléments « lourds » attirant le barycentre, une colonne j attire d'autant plus une ligne i que la valeur de $f_{ij}/f_{i.}$ est élevée. Sur les plans factoriels, les points éloignés de l'origine retiennent particulièrement l'attention car ce sont les profils les plus différents du profil moyen. On trouve donc, pour un facteur, du même côté qu'une ligne i les colonnes j auxquelles elle s'associe le plus et, à l'opposé, celles auxquelles elle s'associe le moins. Il est ainsi possible d'interpréter la position d'**une ligne** par rapport à **l'ensemble des colonnes**, ce qui justifie l'intérêt pratique de la représentation simultanée.

La formulation symétrique vaut, en inversant les rôles joués par les lignes et les colonnes. D'où le nom de double propriété barycentrique donnée à ce qui est **la principale règle d'interprétation des graphiques de l'AFC**. Cette double propriété est non seulement spécifique de l'AFC, mais la caractérise : on démontre que l'on retrouve les facteurs de l'AFC en cherchant à construire des fonctions définies sur les lignes et les colonnes d'un tableau de contingence telles que la double propriété barycentrique soit vérifiée.

La représentation simultanée en AFC est universellement adoptée, ce qui n'est pas le cas de celle de l'ACP. On peut citer deux arguments importants en faveur de cette superposition.

1. Alors qu'en ACP les lignes et les colonnes représentent des objets de nature bien différentes (individus et variables), les lignes et les colonnes, dans l'AFC d'un tableau de contingence, sont de même nature, à savoir des classes d'individus. Selon ce simple point de vue, cela ne pose aucun problème de figurer toutes ces classes sur un même graphique.
2. Il existe d'autres présentations de l'AFC dans lesquelles les classes d'individus que constituent les lignes et les colonnes d'un tableau de contingence sont situées dans un même espace : leur représentation simultanée est alors naturelle.

En résumé, sur les graphiques de la représentation simultanée des lignes et des colonnes, la position relative de deux points d'un même ensemble (lignes ou colonnes) s'interprète en tant que distance tandis que la position d'**un** point d'un ensemble par rapport à celle de **tous** les points de l'autre ensemble s'interprète en tant que barycentre. Toute association entre **une** ligne et **une** colonne suggérée par une proximité sur le graphique doit être contrôlée sur le tableau de données.

3.7.3 Interprétation de l'inertie des axes

L'inertie d'un point (ou d'un nuage de points) dans un espace euclidien se décompose sur toute base orthogonale : c'est la somme de ses inerties sur chacun des axes de cette base.

L'ajustement des nuages N_I et N_J décompose leur inertie selon des directions privilégiées : du fait de l'orthogonalité des axes, la somme des inerties d'un nuage sur chacun des axes est égale à l'inertie totale du nuage.

Contrairement au cas de l'ACP, dans laquelle l'inertie des nuages est égale au nombre de variables, cette inertie en AFC traduit la structure du tableau : l'inertie de chacun des deux nuages, des profils-lignes et des profils-colonnes, est égale à la statistique Φ^2 (*cf.* section 3.7.1). L'AFC propose donc une décomposition de cette statistique et chaque facteur représente une part de la liaison entre les variables. L'inertie d'un facteur a donc une signification en absolu, et pas seulement en pourcentage de l'inertie totale du nuage : elle mesure en absolu l'importance de la part de liaison qu'il représente. Nous donnons l'interprétation des deux valeurs limites entre lesquelles elle se situe.

Lorsqu'un tableau vérifie les relations d'indépendance, les nuages sont concentrés en un point (leur barycentre) ; tous les profils-lignes sont identiques et égaux à la marge ligne $\{f_{.j}; j = 1, \ldots, J\}$ et tous les profils-colonnes sont identiques et égaux à la marge-colonne $\{f_{i.}; i = 1, \ldots, I\}$. L'inertie des nuages N_I et N_J relativement à leur centre de gravité est nulle et l'AFC ne donne aucun facteur (ou plutôt toute direction est associée à une inertie projetée nulle).

Il découle de la double propriété barycentrique que l'inertie associée à un axe factoriel vaut au maximum 1. Lorsque cette inertie vaut 1, l'axe factoriel met en évidence une situation « d'extrême dépendance » au sens suivant : l'ensemble des lignes d'une part, et celui des colonnes d'autre part, peuvent être divisés en au moins deux groupes, chaque groupe de lignes ne s'associant qu'à un groupe de colonnes (et réciproquement) selon le schéma de la **figure 3.9**. Dans ce cas, les facteurs définis par ces axes ont la même valeur pour chaque élément d'un même groupe de lignes ainsi que pour chaque élément du groupe de colonnes qui s'y associe. Une inertie proche de 1 indique que la structure du tableau est proche de cette situation limite : il existe une partition de I et de J telle que chaque classe de I s'associe presque exclusivement à une classe de J et réciproquement.

Lorsque deux axes factoriels ont une inertie égale à 1, les lignes d'une part et les colonnes d'autre part peuvent être divisées en au moins trois groupes qui ne s'associent qu'à un seul groupe de l'autre ensemble, etc. La situation de plus extrême dépendance entre deux variables qualitatives présentant le même nombre de modalités est celle où chaque modalité de l'une des variables ne s'associe qu'à l'une des modalités de l'autre. En ce cas, le tableau de contingence ne possède des effectifs non nuls que sur

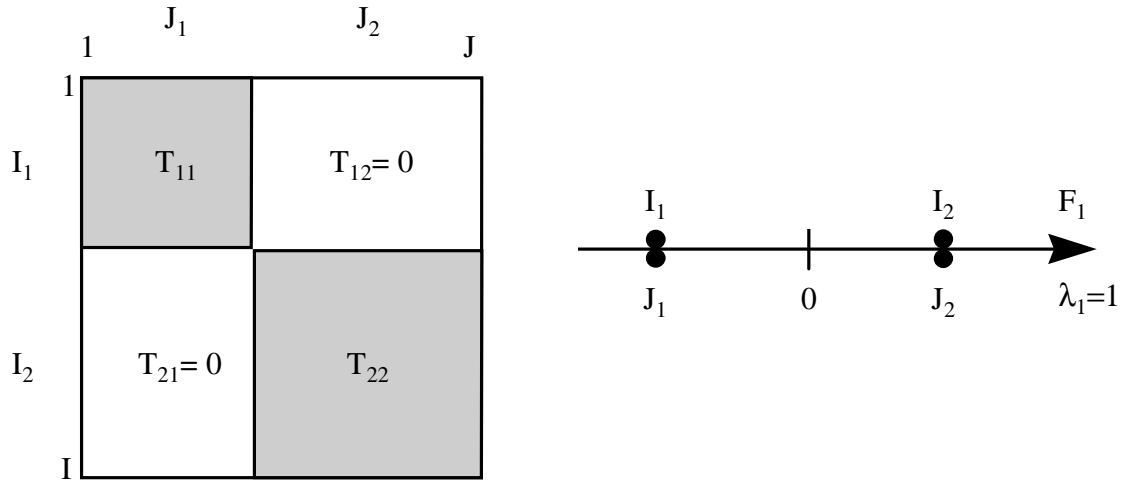


Figure 3.9 Cas d'une inertie associée à un axe égale à 1. Partitions, des lignes d'une part et des colonnes d'autre part, mises en évidence par un axe factoriel associé à une inertie égale à 1. Tous les effectifs des sous-tableaux T_{12} et T_{21} sont nuls.

la diagonale. Il résulte de ce qui précède que, dans ce cas, chaque axe de l'AFC est associé à une inertie de 1.

3.7.4 Formule de reconstitution des données

À la décomposition de l'inertie, on peut associer une décomposition du tableau lui-même. En effet, on peut montrer (*cf.* section 5.6) que :

$$f_{ij} - f_{i.}f_{.j} = f_{i.}f_{.j}\sum_{s} F_s(i)G_s(j)/\sqrt{\lambda_s}$$

Cette formule, appelée formule de reconstitution des données, permet de recalculer les valeurs du tableau initial en fonction des marges et des facteurs. Lorsque l'on dépouille les résultats d'une AFC, on limite généralement l'interprétation aux premiers facteurs. Cela revient à considérer non pas le tableau des données mais son approximation obtenue à l'aide des premiers termes de la somme ci-dessus.

Cette relation met en évidence une décomposition de l'écart du tableau relativement à l'hypothèse d'indépendance en une somme de tableaux dont chacun ne dépend que d'un couple de facteurs (F_s, G_s) de même rang. Elle formalise l'aspect de l'objectif annoncé : décomposition de la liaison en éléments simples. En effet, chaque tableau de terme général $f_{i.}f_{.j}F_s(i)G_s(j)$ exprime une liaison simple puisque le terme de la case (i, j) ne dépend que de la ligne i et de la colonne j. Si les valeurs de $F_s(i)$ et de $G_s(j)$ sont de même signe, cette case exprime une attirance entre i et j ; dans le cas contraire, il exprime une répulsion d'autant plus importante que $F_s(i)$ et $G_s(j)$ sont grands en valeur absolue.

Nous illustrons cette décomposition dans la section 10.3.1.a, page 231, à propos d'un exemple.

3.8 NOMBRE D'AXES ET INERTIE TOTALE

Dans l'espace R^J, le nuage N_I est contenu dans un sous-espace de dimension $J-1$; dans cet espace, on peut donc trouver au maximum $J-1$ dimensions orthogonales d'inertie non nulle. De même, dans l'espace R^I , on peut trouver au maximum $I-1$ dimensions orthogonales d'inertie non nulle. Compte tenu de la dualité (même inertie sur les axes de même rang dans les deux espaces), en AFC on peut trouver au maximum $\min\{I-1, J-1\}$ axes d'inertie non nulle.

L'inertie associée à un axe étant au maximum égale à 1, l'inertie totale en AFC est donc comprise entre 0 (indépendance) et $\min\{I-1, J-1\}$ (liaison d'intensité maximum = association stricte entre les modalités des deux variables mises en correspondances).

3.9 AIDES À L'INTERPRÉTATION ET ÉLÉMENTS SUPPLÉMENTAIRES

Les indices d'aide à l'interprétation (qualité de représentation d'un élément par un axe ou un plan et contribution d'un élément à l'inertie d'un axe) définis en ACP (*cf.* section 1.9) sont valables pour un nuage quelconque. Ils s'appliquent donc en AFC. Notons que, si en ACP les poids de tous les éléments sont en général égaux, ce n'est pas le cas en AFC ; or ces poids interviennent dans la contribution d'un point à l'inertie d'un axe.

En AFC, comme en ACP, on utilise presque systématiquement la technique des éléments supplémentaires, qui consiste à projeter sur les axes factoriels des profils de lignes ou de colonnes qui n'interviennent pas dans le calcul de ces axes. Une ligne supplémentaire est reliée aux colonnes actives par la formule barycentrique. De même, une colonne supplémentaire est reliée aux lignes actives par la formule barycentrique. Ces éléments servent très souvent, eux aussi, d'aides à l'interprétation ; dans les tableaux de grande dimension, par exemple, il est très pratique de connaître la position et la qualité de représentation du barycentre de plusieurs lignes ou de plusieurs colonnes.

3.10 SCHÉMA GÉNÉRAL DE L'AFC

Nous résumons les principaux résultats de l'AFC dans un schéma général (*cf.* **Figure 3.10**). Les numéros ci-dessous renvoient à ce schéma.

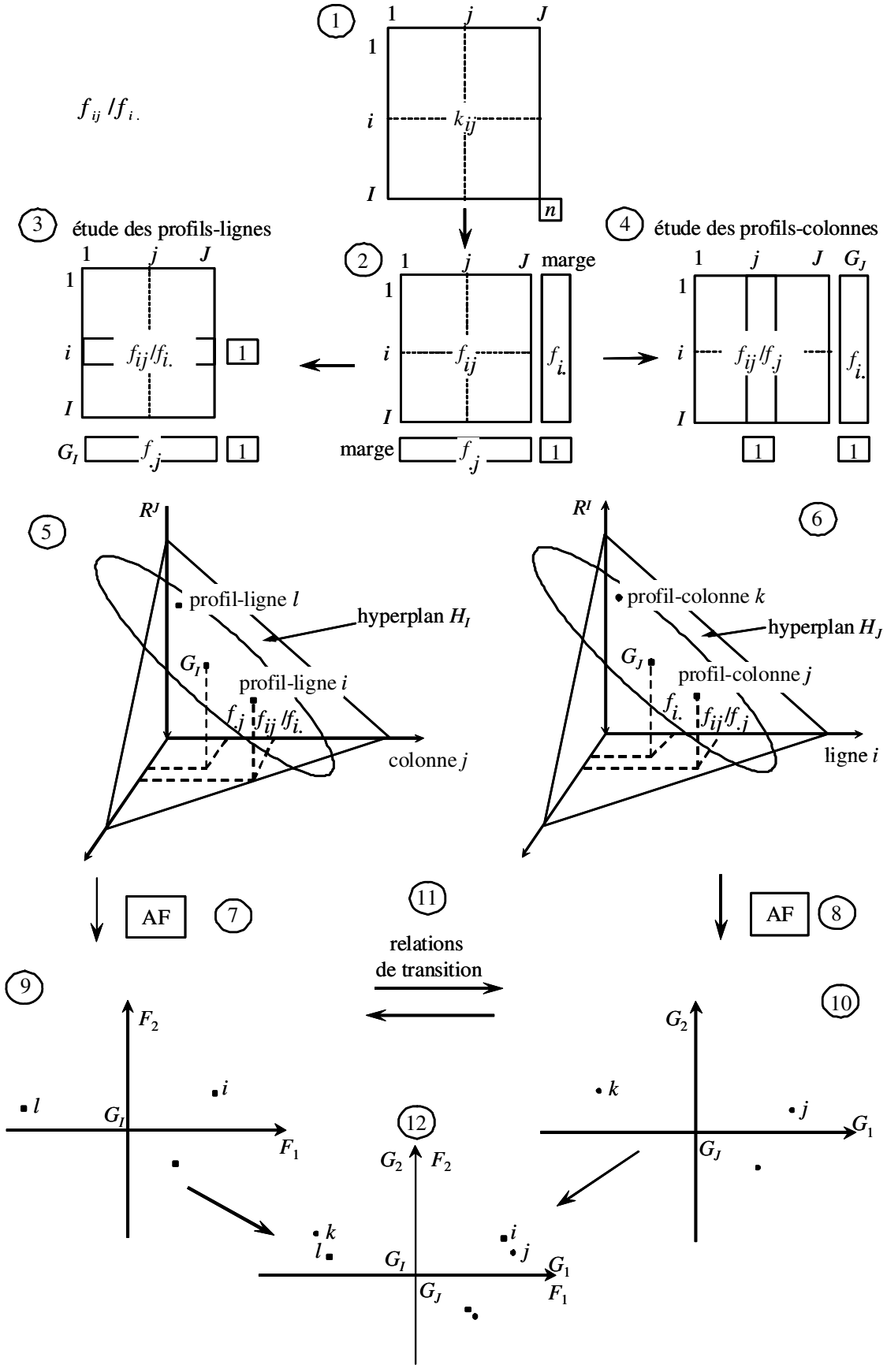


Figure 3.10 Schéma général de l'AFC.

1. Les données brutes. Lignes et colonnes jouent des rôles symétriques : ce sont des modalités de variables. La somme de tous les termes k_{ij} du tableau est n.

2. Ce tableau intermédiaire fait apparaître les données sous forme de loi de probabilité : $f_{ij} = k_{ij}/n$. Les probabilités marginales sont $\{f_{i.}|i \in I\}$ et $\{f_{.j}|j \in J\}$.

3. et **4.** Pour étudier les lignes du tableau, on les transforme en profils-lignes. Pour étudier les colonnes, on les transforme en profils-colonnes. On dispose donc de deux tableaux. Un profil s'interprète comme une probabilité conditionnelle. Les profils moyens G_I et G_J sont les distributions marginales associées au tableau **2**.

5. Un profil-ligne est une suite de J nombres et peut être représenté par un point de R^J. Le nuage N_I des profils-lignes appartient à l'hyperplan H_I des vecteurs dont la somme des coordonnées vaut 1. Chaque profil-ligne i est affecté du poids $f_{i.}$; le nuage N_I ainsi pondéré a pour barycentre le profil moyen G_I. Dans le nuage N_I, on s'intéresse à la ressemblance entre les profils mesurée au travers de la distance du χ^2.

6. La représentation des profils-colonnes dans R^I appelle des commentaires strictement symétriques à ceux de la représentation des profils-lignes dans R^J.

7. L'Analyse Factorielle (AF) d'un nuage consiste à mettre en évidence une suite de directions orthogonales telles que l'inertie, par rapport à O, de la projection du nuage sur ces directions est maximum. Appliquée à N_I, l'AF fournit une première direction – dite triviale – reliant O à G_I et orthogonale à H_I. Pour les directions suivantes, G_I se projette à l'origine des axes : ces directions suivantes sont les directions d'allongement maximum de N_I. Il est donc équivalent de réaliser l'analyse par rapport à O ou par rapport à G_I.

8. On peut reprendre point par point le commentaire de **7** en le transposant aux colonnes.

9. et **10.** Les plans factoriels, croisant deux facteurs, sur les lignes ou sur les colonnes, fournissent des images approchées des nuages N_I et N_J. Sur ces plans, la distance entre deux points s'interprète comme une ressemblance entre les profils de ces points. L'origine des axes est confondue avec le profil moyen.

11. Les relations de transition expriment les résultats d'une AF (par exemple dans R^I) en fonction des résultats de l'autre (par exemple dans R^J).

12. Du fait des relations de transition, les interprétations des plans factoriels représentant N_I et N_J doivent être menées simultanément. Il est commode de superposer ces représentations. L'interprétation de cette représentation simultanée est régie par la double propriété barycentrique.

3.11 CONCLUSION

Dans ce chapitre, l'AFC est introduite comme une méthode particulièrement bien adaptée à l'étude d'un tableau de contingence. D'un point de vue historique, elle a d'ailleurs été imaginée pour traiter ce type de tableau. Toutefois, les remarquables propriétés de cette méthode ont très tôt incité à l'appliquer à d'autres tableaux : aujourd'hui, la pratique courante de l'AFC dépasse largement le cadre des tableaux de contingence.

Dès l'instant que l'on étudie un tableau qui n'est pas un tableau de contingence, l'objectif de l'AFC ne peut plus être formulé en terme de liaison entre deux variables qualitatives. En revanche, il existe des tableaux dont l'étude nécessite une typologie des lignes d'une part et des colonnes d'autre part, à travers leur profil.

Pour établir l'intérêt de l'AFC dans la réalisation de telles typologies, il convient de s'assurer que les différentes notions mises en jeu par cette méthode (transformation en profils, distance du χ^2 , poids des éléments) sont en accord avec le point de vue que l'on veut avoir sur les données étudiées. Les formules barycentriques, qui relient les projections des lignes et des colonnes et qui permettent à elles seules de caractériser les facteurs, peuvent aussi justifier l'application de l'AFC.

Nous illustrons ces situations à l'aide de deux exemples.

Premier exemple : Dans l'étude de la liaison entre le diplôme obtenu et l'emploi occupé, on peut s'intéresser à deux tableaux de même structure établis l'un en se limitant aux hommes et l'autre en se limitant aux femmes. Le chapitre 10 propose une série d'analyses pour ce couple de tableaux. Dès maintenant, on peut se rendre compte de l'intérêt de l'AFC sur une juxtaposition « en ligne » de plusieurs tableaux. En réalité, ce tableau est encore un tableau de contingence dont l'une des deux variables est obtenue par croisement des deux variables *emploi* et *sexe*.

Second exemple : Les lignes sont les entreprises d'un secteur économique. Les colonnes sont les postes d'actif du bilan. À l'intersection de la ligne i et de la colonne j, se trouve la valeur du poste j pour l'entreprise i. Un tel tableau peut être analysé à l'aide d'une ACP. En ce cas, les postes sont des variables centrées et réduites ; chaque poste est affecté du même poids ainsi que chaque entreprise. Généralement, les entreprises diffèrent assez sensiblement par leur total d'actif, ce qui induit presque toujours un effet taille en tant que premier facteur (*cf.* section 1.6).

Mais ce tableau peut aussi être analysé à l'aide d'une AFC. Tout d'abord, ses marges (qui servent de référence) ont une signification claire : la somme des termes de la i^{e} ligne est le total des actifs de l'entreprise i ; la somme des termes de la j^{e} colonne est la valeur du poste j pour l'entreprise fictive que constitue l'ensemble du secteur. Sans entrer dans les détails, les principales caractéristiques impliquées par l'AFC de ce tableau sont les suivantes.

1. Chaque entreprise est analysée au travers de son profil : chacun de ses postes est exprimé par rapport au total des actifs. Un éventuel effet taille est éliminé.
2. Chaque entreprise a un poids proportionnel à son total d'actif.
3. Chaque poste de bilan a un poids proportionnel à son importance pour l'ensemble du secteur.
4. Les postes du bilan sont transformés en profil ; cette harmonisation des données n'est pas très différente du couple centrage-réduction en ACP. À la différence de l'ACP, le nuage des postes est analysé à partir de son barycentre : on étudie les différences entre postes. Ce qui est commun à l'ensemble des postes est éliminé : on ne peut observer d'effet taille.

Ce second exemple montre que certains tableaux peuvent être analysés par ACP ou AFC. Ces deux analyses ne sont pas équivalentes et peuvent fournir des éclairages assez différents. On examinera les pondérations induites par l'AFC aussi bien pour choisir entre les deux méthodes que pour interpréter conjointement leurs résultats.