

## Chapitre 1

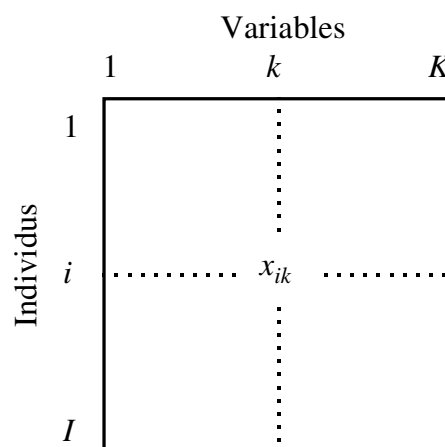
---

# Analyse en Composantes Principales

### 1.1 DONNÉES ET OBJECTIFS DE L'ÉTUDE

L'Analyse en Composantes Principales (ACP) s'applique à des tableaux croisant des individus et des variables quantitatives, appelés de façon concise tableaux *Individus*  $\times$  *Variables quantitatives*.

Selon un usage bien établi, les lignes du tableau représentent les individus et les colonnes représentent les variables. A l'intersection de la ligne  $i$  et de la colonne  $k$  se trouve la valeur de la variable  $k$  pour l'individu  $i$ . La **figure 1.1** illustre ces notions et complète les notations. Le tableau 2.1 page 32 en est un exemple.



**Figure 1.1** Tableau des données en ACP.  $x_{ik}$  : valeur de la variable  $k$  pour l'individu  $i$ .  $I$  : nombre d'individus et ensemble des individus.  $K$  : nombre de variables et ensemble des variables.

Les termes *individu* et *variable* recouvrent des notions différentes. Par exemple, dans le tableau étudié au chapitre 6, les individus sont des vins et les variables sont des critères décrivant ces vins (acidité, astringence, etc.). Les questions que l'on se pose sur les individus et celles que l'on se pose sur les variables ne sont pas de même nature.

À propos de deux **individus**, on essaie d'évaluer leur **ressemblance** : deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables. En ACP, la distance  $d(i, l)$  entre deux individus  $i$  et  $l$  est définie par :

$$d^2(i, l) = \sum_{k \in K} (x_{ik} - x_{lk})^2$$

À propos de deux **variables**, on essaie d'évaluer leur **liaison**. En ACP, la liaison entre deux variables est mesurée par le coefficient de corrélation linéaire (dans de rares situations, on utilise la covariance), noté usuellement  $r$ . Soit :

$$\begin{aligned} r(k, h) &= \frac{\text{covariance}(k, h)}{\sqrt{\text{variance}(k) \times \text{variance}(h)}} \\ &= \frac{1}{I} \sum_{i \in I} \left( \frac{x_{ik} - \bar{x}_k}{s_k} \right) \left( \frac{x_{ih} - \bar{x}_h}{s_h} \right) \end{aligned}$$

avec  $\bar{x}_k$  et  $s_k$  la moyenne et l'écart-type de la variable  $k$ .

Appliquée à un tel tableau, l'objectif général de l'ACP est une étude exploratoire. Les deux voies principales de cette exploration sont :

**Un bilan des ressemblances entre individus.** On cherche alors à répondre à des questions du type suivant : quels sont les individus qui se ressemblent ? Quels sont ceux qui diffèrent ? Plus généralement, on souhaite décrire la variabilité des individus. Pour cela, on cherche à mettre en évidence des groupes homogènes d'individus dans le cadre d'une **typologie des individus**. Selon un autre point de vue, on cherche les **principales dimensions de variabilité** des individus.

**Un bilan des liaisons entre variables.** Les questions sont alors : quelles variables sont corrélées positivement entre elles ? Quelles sont celles qui s'opposent (corrélées négativement) ? Existe-t-il des groupes de variables corrélées entre elles ? Peut-on mettre en évidence une **typologie des variables** ?

Un autre aspect de l'étude des liaisons entre variables consiste à résumer l'ensemble des variables par un petit nombre de **variables synthétiques** appelées ici **composantes principales**. Ce point de vue est très lié au précédent : une composante principale peut être considérée comme le représentant (la synthèse) d'un groupe de variables liées entre elles.

Naturellement, ces deux voies ne sont pas indépendantes du fait de la dualité inhérente à l'étude d'un tableau rectangulaire : la structure du tableau peut être analysée à

la fois par l'intermédiaire de la typologie des individus et de la typologie des variables. Aussi, cherche-t-on en général à relier ces deux typologies. Pour cela, on caractérise les classes d'individus par des variables (on sélectionne ainsi les variables pour lesquelles l'ensemble des individus d'une classe possède des valeurs particulièrement grandes ou particulièrement petites). De même, on caractérise un groupe de variables liées entre elles par des individus types (on sélectionne ainsi les individus qui possèdent des valeurs particulièrement grandes ou des valeurs particulièrement petites pour un ensemble de variables liées positivement entre elles). Enfin, dans la situation idéale, les deux typologies peuvent être « superposées » : chaque groupe de variables caractérise un groupe d'individus et chaque groupe d'individus rassemble les individus types d'un groupe de variables. Ajoutons enfin que la notion de principale dimension de variabilité des individus rejoint celle de variable synthétique.

### a) Poids des individus

Dans la plupart des cas, les individus jouent le même rôle. Nous nous sommes situés implicitement dans cette situation jusqu'ici, en affectant le même poids à chaque individu. Par commodité, on choisit ces poids tels que la masse totale de ces individus soit égale à 1 : à chaque individu on associe alors le poids  $1/I$ . Toutefois, dans certains cas, on peut souhaiter attribuer des poids différents aux individus. Cette situation se présente notamment lorsque les individus représentent chacun une sous-population ; on affecte alors à un individu un poids proportionnel à l'effectif de la sous-population qu'il représente. Ce poids intervient dans le calcul de la moyenne de chaque variable (c'est-à-dire dans la définition d'un individu théorique moyen), dans le calcul de la variance de chaque variable et dans celui de la mesure de liaison (le coefficient de corrélation) entre les variables. Soit, en appelant  $p_i$  le poids affecté à l'individu  $i$  ( $\sum_i p_i = 1$ ) :

$$\bar{x}_k = \sum_i p_i x_{ik} \quad s_k^2 = \sum_i p_i (x_{ik} - \bar{x}_k)^2$$

$$r(k, h) = \sum_i p_i \left( \frac{x_{ik} - \bar{x}_k}{s_k} \right) \left( \frac{x_{ih} - \bar{x}_h}{s_h} \right)$$

Les programmes complets d'ACP permettent tous d'introduire des poids d'individus.

### b) Poids des variables

Nous avons accordé jusqu'ici la même importance *a priori* aux différentes variables. On est très rarement conduit, dans la pratique, à souhaiter leur affecter des importances différentes. À tel point que les programmes courants d'ACP ne le permettent pas. Cette importance peut être modulée à l'aide d'un coefficient appelé poids de la variable. En appelant  $m_k$  le poids de la variable  $k$ , la distance entre deux individus  $i$  et  $l$  est définie par :

$$d^2(i, l) = \sum_{k \in K} m_k (x_{ik} - x_{lk})^2$$

Toutefois, comme nous le verrons dans le chapitre 5 qui contient l'ensemble des résultats techniques concernant les analyses factorielles, ces poids ne modifient en rien les principes généraux de l'analyse. Afin de ne pas alourdir l'exposé de ce chapitre, nous considérons dans la suite que les individus possèdent le même poids ( $p_i = 1/I$  quel que soit  $i \in I$ ) ainsi que les variables ( $m_k = 1$  quel que soit  $k \in K$ ).

## 1.2 TRANSFORMATION DES DONNÉES

En ACP, le tableau des données est toujours centré (en pratique, le centrage est inclus dans les programmes d'ACP). A chaque valeur numérique, on soustrait la moyenne de la variable en cause. Le tableau obtenu est alors de terme général :

$$x_{ik} - \bar{x}_k$$

Cette transformation n'a aucune incidence sur les définitions de la ressemblance entre individus et de la liaison entre variables. À ce niveau, elle peut être considérée comme un intermédiaire technique qui présente d'intéressantes propriétés mais qui ne change fondamentalement rien à la problématique.

L'ACP peut être réalisée sur des données seulement centrées. Toutefois, ses résultats sont alors très sensibles au choix des unités de mesure. Généralement, ce choix est arbitraire : ainsi, dans l'exemple classique de mensurations d'animaux, la variable *hauteur* peut être exprimée en mètres ou en centimètres. Or ce choix a une grande influence sur la mesure de ressemblance entre individus. Le passage du mètre au centimètre multiplie par  $100^2$  l'influence de la variable *hauteur* dans le calcul du carré de la distance entre deux individus.

La façon classique de s'affranchir de l'arbitraire des unités de mesure est de réduire les données. Le tableau obtenu a pour terme général  $(x_{ik} - \bar{x}_k)/s_k$ . Ce faisant, on utilise comme unité de mesure pour la variable  $k$ , son écart-type  $s_k$ . Toutes les variables présentent alors la même variabilité et de ce fait la même influence dans le calcul des distances entre individus.

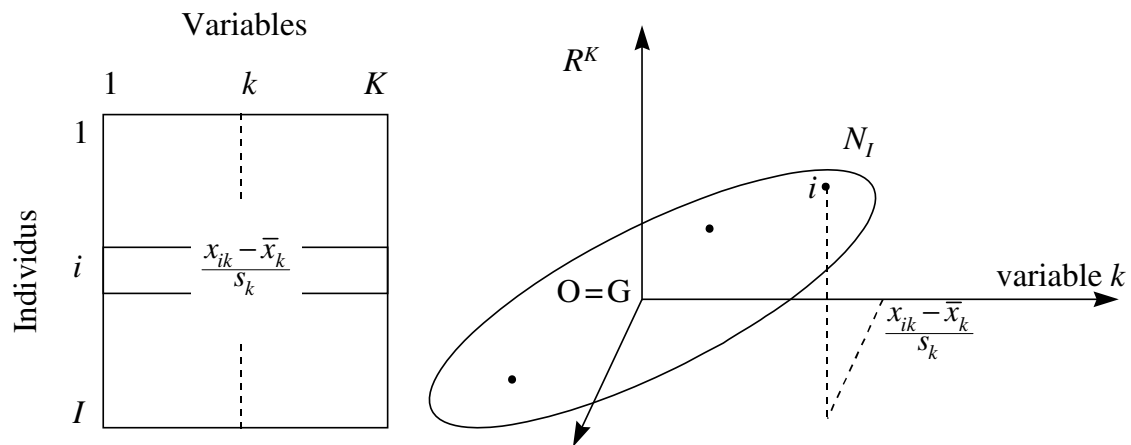
Dans les études où toutes les variables s'expriment dans la même unité, on peut souhaiter ne pas réduire les variables. En procédant ainsi, on accorde à chaque variable réduite un poids égal à sa variance (cf. définition de la distance entre individus). Selon un autre point de vue, la définition de  $d(i, l)$  montre que la variance de la variable  $k$  est égale à la contribution moyenne de la variable  $k$  au carré de la distance entre individus. Cela se déduit de l'écriture suivante de la variance :

$$s_k^2 = \frac{1}{2I^2} \sum_{i,l} (x_{ik} - x_{lk})^2$$

Un exemple de discussion de l'opportunité de la réduction est donné section 2.1.2 page 32. Dans la suite, sauf mention explicite du contraire, les variables sont toujours supposées centrées et réduites.

### 1.3 NUAGE DES INDIVIDUS

S'intéresser aux individus revient à envisager le tableau en tant que juxtaposition de lignes. À chaque individu est associée une suite de  $K$  nombres. Selon ce point de vue, un individu peut être représenté comme un point de l'espace vectoriel à  $K$  dimensions, noté  $R^K$ , dont chaque dimension représente une variable. L'ensemble des individus constitue le nuage  $N_I$  dont le centre de gravité  $G$  est confondu avec l'origine  $O$  des axes du fait du centrage ;  $G$  représente l'individu moyen précédemment cité. Ces notations sont rassemblées **figure 1.2**.



**Figure 1.2** Tableau des données et nuage des individus associé dans l'espace  $R^K$ . Du fait du centrage, l'origine des axes est confondue avec le centre de gravité du nuage.

Dans l'espace  $R^K$ , la notion de ressemblance entre deux individus introduite section 1.1 n'est autre que la distance euclidienne usuelle. Cette interprétation géométrique constitue une justification a posteriori décisive du choix de la mesure de ressemblance : le fait qu'elle soit une distance euclidienne lui confère un grand nombre de propriétés mathématiques indispensables pour la suite.

L'ensemble des distances inter-individuelles constitue ce que l'on appelle la forme du nuage  $N_I$ . Réaliser un bilan de ces distances revient à étudier la forme du nuage  $N_I$ , c'est-à-dire à y déceler une partition des points (la typologie mentionnée dans les objectifs) ou des directions d'allongement remarquables (les principales dimensions de variabilité).

Dès que  $K$  est supérieur à 3, l'étude directe du nuage  $N_I$  est impossible du fait de la limitation à trois dimensions de notre sens visuel. D'où l'intérêt des méthodes

factorielles en général, et dans ce cas particulier de l'ACP, qui fournissent des images planes approchant le mieux possible (au sens d'un critère défini et discuté section 1.5) un nuage de points situé dans un espace de grande dimension.

## 1.4 NUAGE DES VARIABLES

S'intéresser aux variables revient à envisager le tableau en tant que juxtaposition de colonnes. À chaque variable, est associée une suite de  $I$  nombres. Selon ce point de vue, une variable peut être représentée comme un vecteur de l'espace vectoriel à  $I$  dimensions, noté  $R^I$ , dont chaque dimension représente un individu : par exemple, la variable  $k$  est représentée par le vecteur noté lui aussi  $k$  et dont la  $i^{\text{e}}$  composante est  $(x_{ik} - \bar{x}_k)/s_k$ . L'ensemble des extrémités des vecteurs représentant les variables constitue le nuage  $N_K$ . Ces notations sont regroupées dans la **figure 1.3**.

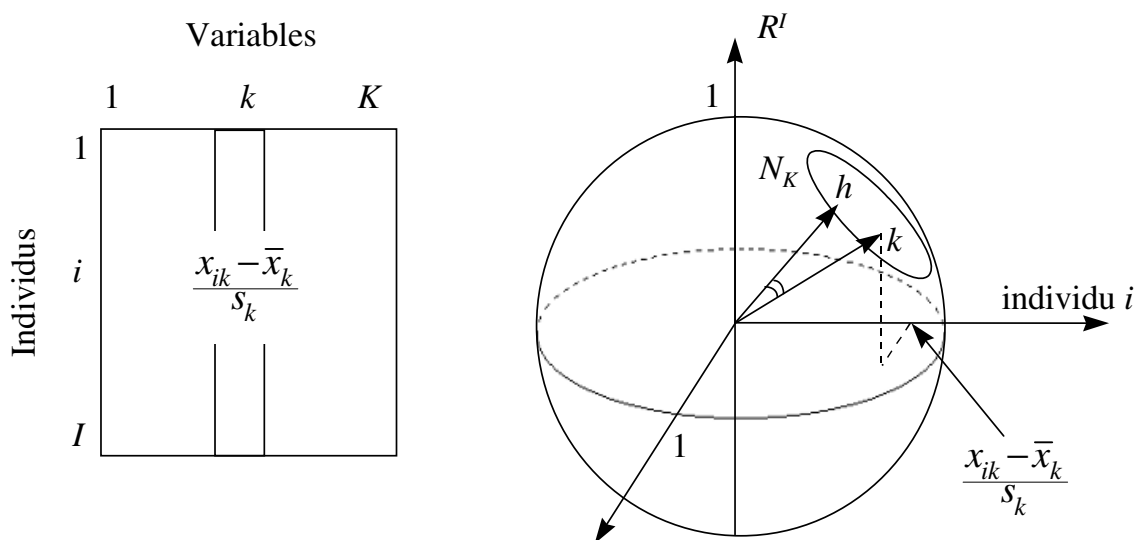


Figure 1.3 Tableau des données et nuage des variables associé dans l'espace  $R^I$ .

$$\cos(\overrightarrow{Oh}, \overrightarrow{Ok}) = r(h, k)$$

Le choix de la distance dans  $R^I$  consiste à affecter à chaque dimension un coefficient égal au poids de chaque individu dans le nuage  $N_I$  de  $R^K$  (on peut avoir l'intuition de ce choix en considérant deux individus absolument identiques que l'on peut remplacer par un seul ayant un poids double). Dans le cas général où ces poids sont identiques, la distance utilisée est, au coefficient  $1/I$  près, la distance euclidienne usuelle. Avec cette distance, les vecteurs représentant les variables centrées ont les propriétés suivantes :

1. La norme de chaque vecteur représentant une variable est égale à son écart-type.  
Soit :

$$\|\text{variable } k\|^2 = \sum_{i=1}^I \frac{1}{I} (x_{ik} - \bar{x}_k)^2$$

Ainsi, lorsque les variables sont centrées réduites, chaque variable a pour longueur 1 : le nuage  $N_K$  est alors situé sur une sphère de rayon 1 (on dit aussi hypersphère pour rappeler que  $R^I$  est de dimension supérieure à 3). Pour cette raison, l'ACP sur données centrées-réduites est dite **ACP normée**. Lorsque les variables sont seulement centrées, leur longueur est égale à leur écart-type et on parle alors d'**ACP non normée**.

2. Le cosinus de l'angle formé par les vecteurs représentant les deux variables  $h$  et  $k$ , obtenu en calculant le produit scalaire noté  $\langle h, k \rangle$  entre ces deux vecteurs normés, est égal au coefficient de corrélation entre ces deux variables. Soit :

$$\cos(h, k) = \langle h, k \rangle = \sum_i \frac{1}{I} \left( \frac{x_{ih} - \bar{x}_h}{s_h} \right) \left( \frac{x_{ik} - \bar{x}_k}{s_k} \right) = \text{corr\u00e9lation}(h, k)$$

L'interpr\u00e9tation d'un coefficient de cor\u00e9lation comme un cosinus est une propri\u00e9t\u00e9 tr\u00e8s importante puisqu'elle donne un support g\u00e9om\u00e9trique, donc visuel, au coefficient de cor\u00e9lation. Cette propri\u00e9t\u00e9 n\u00e9cessite le centrage, ce qui justifie cette transformation pr\u00e9sent\u00e9e section 1.1 comme un interm\u00e9diaire technique. Elle justifie aussi le choix de la distance (on dit aussi *m\u00e9trique*) dans  $R^I$  et implique que, dans la repr\u00e9sentation des variables, on s'int\u00e9resse surtout aux directions d\u00e9termin\u00e9es par les variables, c'est-\u00e0-dire aux vecteurs plut\u00f4t qu'aux leurs extr\u00e9mit\u00e9s.

La longueur des vecteurs repr\u00e9sentant les variables \u00e9tant \u00e9gale \u00e0 1, la coordonn\u00e9e de la projection d'une variable sur une autre s'interpr\u00e8te comme un coefficient de cor\u00e9lation.

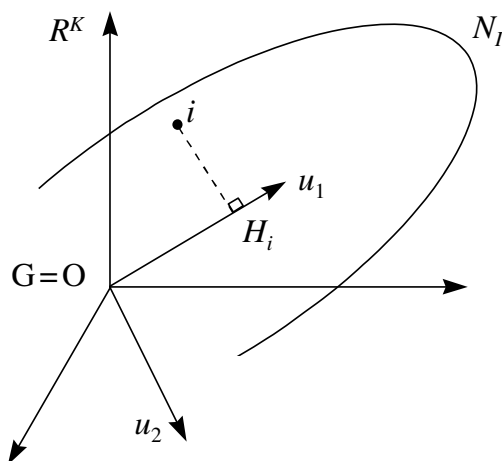
### ► Conclusion

R\u00e9aliser un bilan des coefficients de cor\u00e9lation entre les variables revient \u00e0 \u00e9tudier les angles entre les vecteurs d\u00e9finissant le nuage  $N_K$ . Cette \u00e9tude directe est impossible du fait de la dimension de  $R^I$ . L'int\u00e9r\u00eat de l'ACP est de fournir des variables synth\u00e9tiques qui constituent un r\u00e9sum\u00e9 de l'ensemble des variables initiales et sont la base d'une repr\u00e9sentation plane approch\u00e9e des variables et de leurs angles.

## 1.5 AJUSTEMENT DU NUAGE DES INDIVIDUS

L'objectif est de fournir des images planes approch\u00e9es du nuage  $N_I$  situ\u00e9 dans l'espace  $R^K$  (cf. section 1.3). Pratiquement, on recherche une suite  $\{u_s; s = 1, \dots, S\}$  de  $S$  directions privil\u00e9gi\u00e9es de  $R^K$  appel\u00e9es axes factoriels qui, prises deux \u00e0 deux, d\u00e9finissent des plans factoriels sur lesquels on projette le nuage  $N_I$ . Chaque direction  $u_s$  est choisie de fa\u00e7on \u00e0 rendre maximum l'inertie par rapport \u00e0 l'origine  $O$  (confondue avec le centre de gravit\u00e9  $G$ , du fait du centrage) de la projection de  $N_I$  sur  $u_s$ . Dans la recherche d'une suite, on impose \u00e0 chaque direction d'\u00eatre orthogonale aux directions d\u00e9j\u00e0 trouv\u00e9es (cf. **Figure 1.4**). On peut montrer que le plan engendr\u00e9 par les deux

premiers axes  $u_1$  et  $u_2$  rend maximum l'inertie projetée sur ce plan. Il en est de même pour le sous-espace engendré par les trois premiers axes, etc.



**Figure 1.4** L'ajustement du nuage des individus. L'individu  $i$  se projette en  $H_i$  sur  $u_1$ . On cherche d'abord  $u_1$  qui rend maximum  $\sum_i OH_i^2$ . Puis on cherche  $u_2$ , orthogonal à  $u_1$ , qui satisfait le même critère et ainsi de suite. Lorsque les individus sont munis de poids  $p_i$  différents, le critère consiste à rendre maximum :  $\sum_i p_i OH_i^2$ .

Il est équivalent de rendre maximum  $\sum_i OH_i^2$  ou de rendre minimum  $\sum_i i H_i^2$ . Cette deuxième écriture, forme classique du critère des moindres carrés, montre que les axes factoriels rendent minimum l'écart entre le nuage des individus et sa projection.

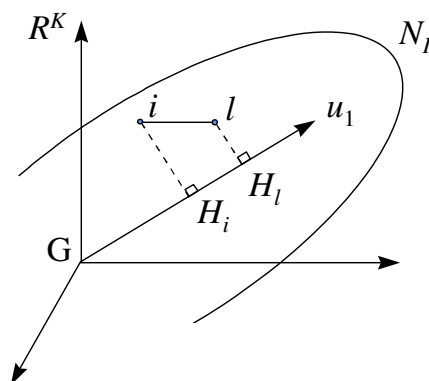
Du fait du centrage, le critère (inertie maximum par rapport au centre de gravité  $G$ ) permet d'interpréter les axes factoriels comme des directions d'allongement maximum du nuage  $N_K$ . On parle aussi de **principales dimensions de variabilité**, dans la mesure où ils rendent compte le plus possible de la diversité des individus.

On peut montrer que, toujours du fait du centrage, rendre maximum  $\sum_i OH_i^2$  est équivalent à rendre maximum  $\sum_i \sum_l (OH_i - OH_l)^2$ . Cette dernière forme fait apparaître les distances entre points projetés. La projection ne pouvant que réduire la distance entre points, les axes factoriels apparaissent comme les directions telles que les distances entre points projetés ressemblent le plus possible aux distances entre les points homologues de  $N_I$  (cf. **Figure 1.5**).

Selon les objectifs d'une analyse, on mettra en avant l'une ou l'autre des interprétations du critère.

► Individus supplémentaires (= illustratifs)

Fréquemment, on souhaite que certains individus n'interviennent pas dans la détermination des axes ; par contre, on souhaite connaître la position de leur projection sur les



**Figure 1.5** La représentation des distances inter-individuelles. L'axe  $u_1$  rend  $\sum_i \sum_l (OH_i - OH_l)^2$  maximum, c'est-à-dire est tel que  $\sum_i \sum_l d^2(H_i - H_l)$  est le plus proche possible de  $\sum_i \sum_l d^2(i, l)$ .

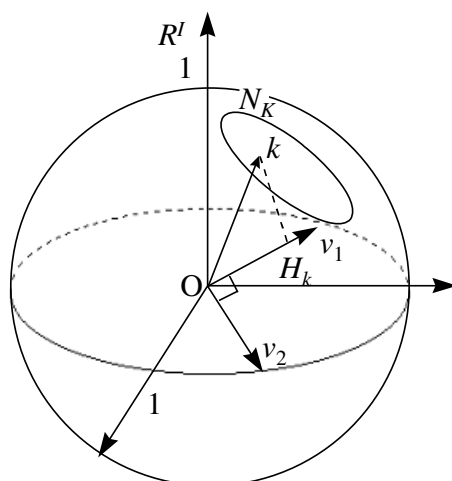
axes déterminés par les autres individus (dits *actifs*). Tous les programmes prévoient cette situation ce qui revient à mettre un poids nul à certains individus au niveau du critère d'ajustement.

Ces individus sont appelés **individus supplémentaires** (ou illustratifs). On introduit un individu en supplémentaire lorsque l'on souhaite qu'il participe à l'interprétation des plans factoriels mais non à leur construction. C'est le cas lorsque l'on dispose d'individus présentant des caractères exceptionnels, ou suspectés d'avoir été l'objet d'erreurs de mesures, ou enfin n'appartenant pas au champ strict de l'étude mais à un domaine voisin.

## 1.6 AJUSTEMENT DU NUAGE DES VARIABLES

Pour obtenir une suite de  $S$  variables synthétiques  $\{v_s; s = 1, \dots, S\}$  et une représentation approchée des corrélations entre les variables, l'ACP applique au nuage  $N_K$  des variables la même démarche qu'au nuage des individus (cf. **Figure 1.6**).

Le critère (inertie projetée maximum) satisfait dans le choix des axes est exactement le même que pour le nuage d'individus. Mais il prend une signification différente du fait que le nuage n'est pas centré (son centre de gravité n'est pas à l'origine) et que tous les points sont situés sur la sphère unité : ce sont les angles entre les vecteurs représentant les variables qui sont peu déformés par les projections et non pas les distances entre les points du nuage. En effet, le plan  $(v_1, v_2)$ , en maximisant l'inertie à l'origine du nuage projeté, rend maximum la somme des cosinus carrés des angles entre les vecteurs et leur projection : il ajuste les vecteurs et déforme donc le moins possible leurs angles.



**Figure 1.6** L'ajustement du nuage des variables.  $H_k$  : projection du point représentant la variable  $k$  sur  $v_1$ . On cherche d'abord  $v_1$  qui rend maximum :  $\sum_k OH_k^2$ . Puis on cherche  $v_2$ , orthogonal à  $v_1$ , qui satisfait le même critère et ainsi de suite.

### ► Composantes principales

Le vecteur  $v_1$  qui caractérise la direction d'inertie maximum définit une nouvelle variable. Les variables étudiées étant centrées et réduites, leur projection sur  $v_1$  est égale à leur coefficient de corrélation avec cette variable (*cf.* section 1.4). De ce fait, rechercher le vecteur  $v_1$  qui rend maximum  $\sum_k OH_k^2$  équivaut à rechercher la combinaison linéaire la plus liée à l'ensemble des variables (au sens du critère : somme des carrés des corrélations maximum). En ce sens,  $v_1$  est la variable qui synthétise le mieux l'ensemble des variables initiales. Les axes factoriels étant orthogonaux deux à deux, on met en évidence une suite de variables synthétiques, les composantes principales, non corrélées entre elles, qui résument au mieux l'ensemble des variables initiales.

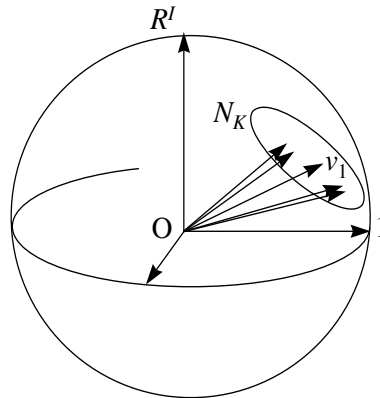
### ► Variables supplémentaires (= illustratives)

Les variables, comme les individus, peuvent être traitées en éléments supplémentaires. Les variables supplémentaires sont simplement projetées sur les axes déterminés par les autres variables, dites actives. Cela permet de visualiser les corrélations entre n'importe quelle variable, même extérieure au domaine étudié, et les composantes principales.

### ► L'effet taille

Si, dans un jeu de données, les variables sont toutes corrélées positivement deux à deux, alors le nuage  $N_K$  est loin de l'origine. Le premier axe factoriel rend alors surtout compte de la position de  $N_K$  par rapport à l'origine : parallèlement, la forme

du nuage  $N_K$  est mal représentée en ce sens que les projections des variables sont proches les unes des autres (cf. **Figure 1.7**).



**Figure 1.7** L'effet taille dans  $R^I$ . Les variables, étant corrélées positivement deux à deux, forment entre elles des angles aigus. Le nuage  $N_K$  est concentré sur un petit secteur de la sphère. La projection des variables sur le premier axe factoriel, défini par  $v_1$ , rend compte principalement de la position de  $N_K$  par rapport à O.

Ce cas de figure est couramment appelé « effet taille » : il correspond à la situation dans laquelle certains individus ont des petites valeurs pour l'ensemble des variables, d'autres de grandes valeurs pour l'ensemble des variables, les autres occupant une situation intermédiaire entre ces extrêmes. Il existe donc dans ce cas une structure commune à l'ensemble des variables : c'est ce que traduit la première composante principale.

## 1.7 DUALITÉ ET FORMULES DE TRANSITION EN ACP

Le nuage  $N_I$  des individus et le nuage  $N_K$  des variables sont deux représentations du même tableau, l'une à travers ses lignes et l'autre à travers ses colonnes. Des relations très fortes, dites relations de dualité (démontrées en section 5.4) lient ces deux nuages.

### 1.7.1 Inerties

Tout d'abord, leur inertie totale est la même ; elle est égale au nombre de variables (lorsque les variables sont réduites) :

$$\text{Inertie totale de } N_I \text{ (ou de } N_K) = \frac{1}{I} \sum_k \sum_i \left( \frac{x_{ik} - \bar{x}_k}{s_k} \right)^2 = K$$

La projection de chacun de ces deux nuages sur une suite d'axes orthogonaux correspond à une décomposition de l'inertie totale. On peut montrer que les deux

décompositions sont identiques : les inerties des nuages  $N_I$  et  $N_K$  projetés sur les axes factoriels de même rang sont égales (et notées  $\lambda_s$ ). Soit, pour les axes de rang  $s$  :

$$\text{Inertie}(N_I/u_s) = \text{Inertie}(N_K/v_s) = \lambda_s$$

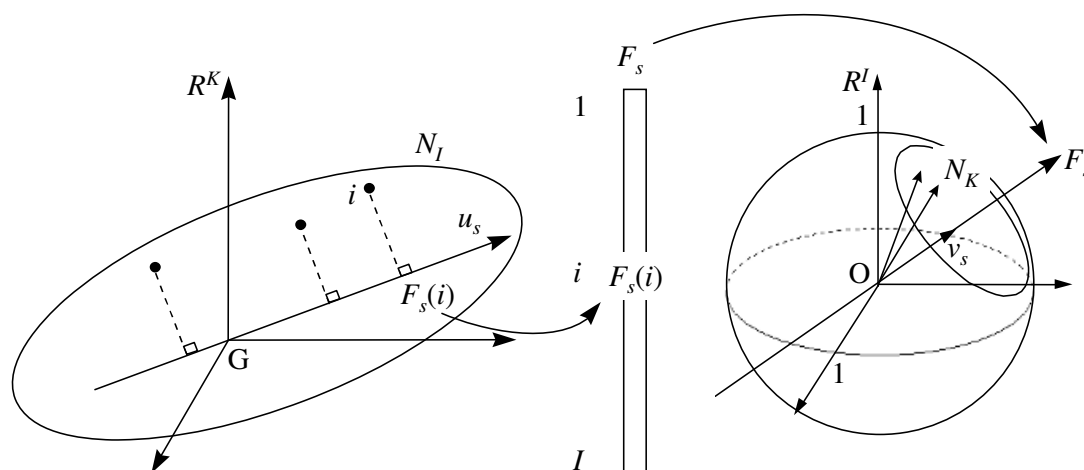
### 1.7.2 Facteurs

L'ensemble des projections de tous les points du nuage d'individus  $N_I$  sur le  $s^{\text{e}}$  axe factoriel  $u_s$ , appelé  $s^{\text{e}}$  facteur sur les individus, constitue une nouvelle variable notée  $F_s$ . On montre, dans la section 5.4.1, que cette variable se confond, à la norme près, avec la  $s^{\text{e}}$  composante principale  $v_s$  obtenue dans l'analyse du nuage des variables. Plus précisément, le carré de la norme du facteur  $F_s$  (vecteur de  $R^I$ ), étant la somme des carrés de ses coordonnées, vaut  $\lambda_s$  ; la relation entre le  $s^{\text{e}}$  facteur sur  $I$  et le  $s^{\text{e}}$  axe factoriel de  $R^I$  s'écrit donc :

$$v_s = \frac{1}{\sqrt{\lambda_s}} F_s$$

Ces résultats sont illustrés dans la **figure 1.8**.

Ainsi, les projections planes des individus dans  $R^K$  sont des représentations graphiques des couples de variables synthétiques obtenues dans  $R^I$ . Les résultats issus de l'étude de chacun des deux nuages possèdent fondamentalement la même signification, même s'ils s'expriment en termes d'individus pour l'un et en termes de variables pour l'autre.



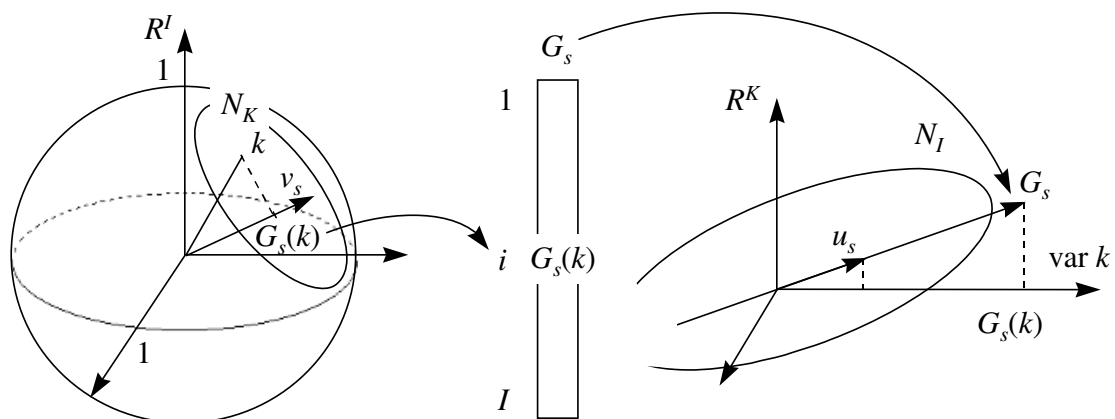
**Figure 1.8** Une des deux formes de la dualité. Les coordonnées de  $N_I$  sur  $u_s$  ( $s^{\text{e}}$  axe factoriel de  $N_I$ ) constituent le  $s^{\text{e}}$  facteur sur les individus (noté  $F_s$ ). Le vecteur  $F_s$  dans  $R^I$  est colinéaire à  $v_s$  ( $s^{\text{e}}$  axe factoriel de  $N_K$ ).

Le rôle du nuage des individus et celui du nuage des variables sont, dans une certaine mesure, symétriques et la dualité se formule de manière analogue en échangeant le rôle des deux nuages : la projection des  $K$  variables sur le  $s^e$  axe factoriel  $v_s$  de leur nuage  $N_K$  définit une valeur pour chacune des  $K$  variables : ces valeurs constituent le  $s^e$  facteur sur les variables (noté  $G_s$ ) qui est en quelque sorte un « individu » nouveau. Cette notion d'individu « type » est moins classique que celle de composante principale (pratiquement, on prend plutôt des individus réels comme individus types). Cependant, dans quelques cas particuliers, comme celui où les individus sont des courbes et les variables leurs valeurs en  $K$  points de discrétisation, ces individus sont représentables et de ce fait utilisés.

On montre que le point représentant dans  $R^K$  cet individu type est situé sur le  $s^e$  axe du nuage des individus. Plus précisément :

$$u_s = \frac{1}{\sqrt{\lambda_s}} G_s$$

Cette relation montre que, au coefficient  $\sqrt{\lambda_s}$  près, les coordonnées des variables sur  $v_s$  sont les coefficients de la combinaison linéaire des variables que constitue l'axe  $u_s$  de  $R^K$ . Ainsi, la coordonnée de la variable  $k$  sur  $v_s$  s'interprète à la fois comme le coefficient de corrélation entre  $k$  et  $v_s$  et comme le coefficient de  $k$  dans  $u_s$  ; cette double interprétation est caractéristique des axes principaux et essentielle dans l'interprétation (à l'inverse, penser aux difficultés d'interprétation des coefficients de la régression multiple quand ils ne sont pas de même signe que les coefficients de corrélation associés). Ce résultat est illustré dans la **figure 1.9**.



**Figure 1.9** La deuxième forme de la dualité. Les coordonnées de  $N_K$  sur  $v_s$  ( $s^e$  axe factoriel de  $N_K$ ) constituent le  $s^e$  facteur sur les variables (noté  $G_s$ ). Le vecteur  $G_s$  dans  $R^K$  est colinéaire au  $s^e$  axe factoriel  $u_s$  de  $N_I$ .

### 1.7.3 Relations de transition

On appelle relations de transition entre les facteurs de rang  $s$ ,  $F_s$  et  $G_s$ , l'écriture algébrique des propriétés illustrées par les **figures 1.8** et **1.9**. Ces relations s'écrivent, en notant  $\lambda_s$  l'inertie projetée de  $N_I$  (ou de  $N_K$ ) sur l'axe de rang  $s$  :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{x_{ik} - \bar{x}_k}{s_k} G_s(k)$$

$$G_s(k) = \frac{1}{I} \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{x_{ik} - \bar{x}_k}{s_k} F_s(i)$$

La première relation exprime le fait que la projection  $F_s(i)$  d'un individu  $i$ , est une combinaison linéaire des projections  $G_s(k)$  de toutes les variables. Dans cette combinaison linéaire, le coefficient d'une variable  $k$  est positif si la valeur  $x_{ik}$  de cette variable pour l'individu  $i$  dépasse la moyenne  $\bar{x}_k$ . Dans le cas contraire, ce coefficient est négatif. Ainsi, lorsque l'on regarde simultanément les deux graphiques, un individu est du côté des variables pour lesquelles il a de fortes valeurs **et** à l'opposé des variables pour lesquelles il a de faibles valeurs.

Le graphique des individus est une représentation approchée des distances inter-individuelles. Celui des variables peut être considéré en tant qu'élément explicatif de cette représentation : deux individus situés à une même extrémité d'un axe sont proches car ils ont tous deux généralement de fortes valeurs pour les variables situées du même côté qu'eux **et** de faibles valeurs pour les variables situées à l'opposé.

Réciproquement, le graphique des individus peut intervenir en tant qu'aide à l'interprétation du graphique des variables : si deux variables sont très corrélées positivement, elles sont situées du même côté sur un axe. Sur l'axe correspondant du nuage d'individus, les individus qui ont de fortes valeurs pour ces deux variables se situent du même côté qu'elles et ceux qui ont de faibles valeurs se situent à l'opposé. Les individus extrêmes pour ces variables sont loin de l'origine. Les éventuels individus particuliers induisant à eux seuls des corrélations fortes sont ainsi repérés facilement.

Ainsi, en ACP, le graphique des individus et celui des variables sont à la fois optimaux en eux-mêmes (ils représentent le mieux possible l'un les individus l'autre les variables) **et** se servent mutuellement d'aides à l'interprétation. Cette propriété liant les représentations des lignes et des colonnes vaut pour toutes les analyses factorielles et leur est spécifique.

### 1.7.4 Représentation superposée

La nécessité d'une interprétation conjointe des représentations des individus et des variables conduit certains utilisateurs à les superposer. Il importe de souligner que la justification d'une telle représentation simultanée des individus et des variables est

essentiellement pragmatique : la représentation des variables aide l'interprétation de celle des individus et réciproquement. Elle pose toutefois le problème de la représentation sur un même graphique de points de natures différentes, évoluant dans des espaces différents. Cette difficulté n'est pas seulement de principe : la présence simultanée d'individus et de variables sur un même plan engendre des proximités entre individus et variables qui, à leur tour, peuvent suggérer des idées qui ne se vérifient pas dans les données. C'est pourquoi cette représentation est déconseillée. Toutefois, en conservant à l'esprit les points de repère suivants, on pourra utiliser sans danger la représentation simultanée en ACP.

1. Les formules de transition relient la coordonnée sur un axe d'un individu avec l'ensemble des coordonnées de toutes les variables sur l'axe de même rang. On ne peut interpréter la position d'un individu par rapport à une seule variable (et réciproquement).
2. Fondamentalement, les variables sont des vecteurs et non des points. Ce n'est pas la proximité entre un individu et un ensemble de points représentant des variables qui est importante mais l'éloignement de l'individu dans la direction de cet ensemble de variables.

### 1.7.5 Projection des vecteurs unitaires de la représentation des individus

Une autre idée, en vue de la représentation superposée des individus et des variables, consiste à projeter les vecteurs unitaires de  $R^K$  sur les axes  $u_s$ . On obtient ainsi une représentation superposée plus naturelle que la précédente, en ce sens que les objets représentés proviennent du même espace.

Du fait de la relation entre  $u_s$  et  $G_s$ , et en remarquant que la  $k^e$  coordonnée de  $u_s$  est égale à la projection sur  $u_s$  du vecteur unitaire du  $k^e$  axe de  $R^K$ , cette nouvelle représentation des variables est homothétique de la précédente axe par axe dans le rapport  $\sqrt{\lambda_s}$ .

Notre préférence va à la 1<sup>e</sup> représentation superposée, fondée sur les relations de transition données plus haut, car elle permet d'inclure les variables supplémentaires.

## 1.8 SCHÉMA GÉNÉRAL DE L'ACP

Nous résumons les principaux résultats de ce chapitre dans un schéma général (*cf. Figure 1.10*). Les numéros ci-dessous renvoient à ce schéma.

1. Les données brutes. Lignes (individus) et colonnes (variables) ne jouent pas des rôles symétriques : les moyennes et les variances n'ont généralement de sens que pour les colonnes.

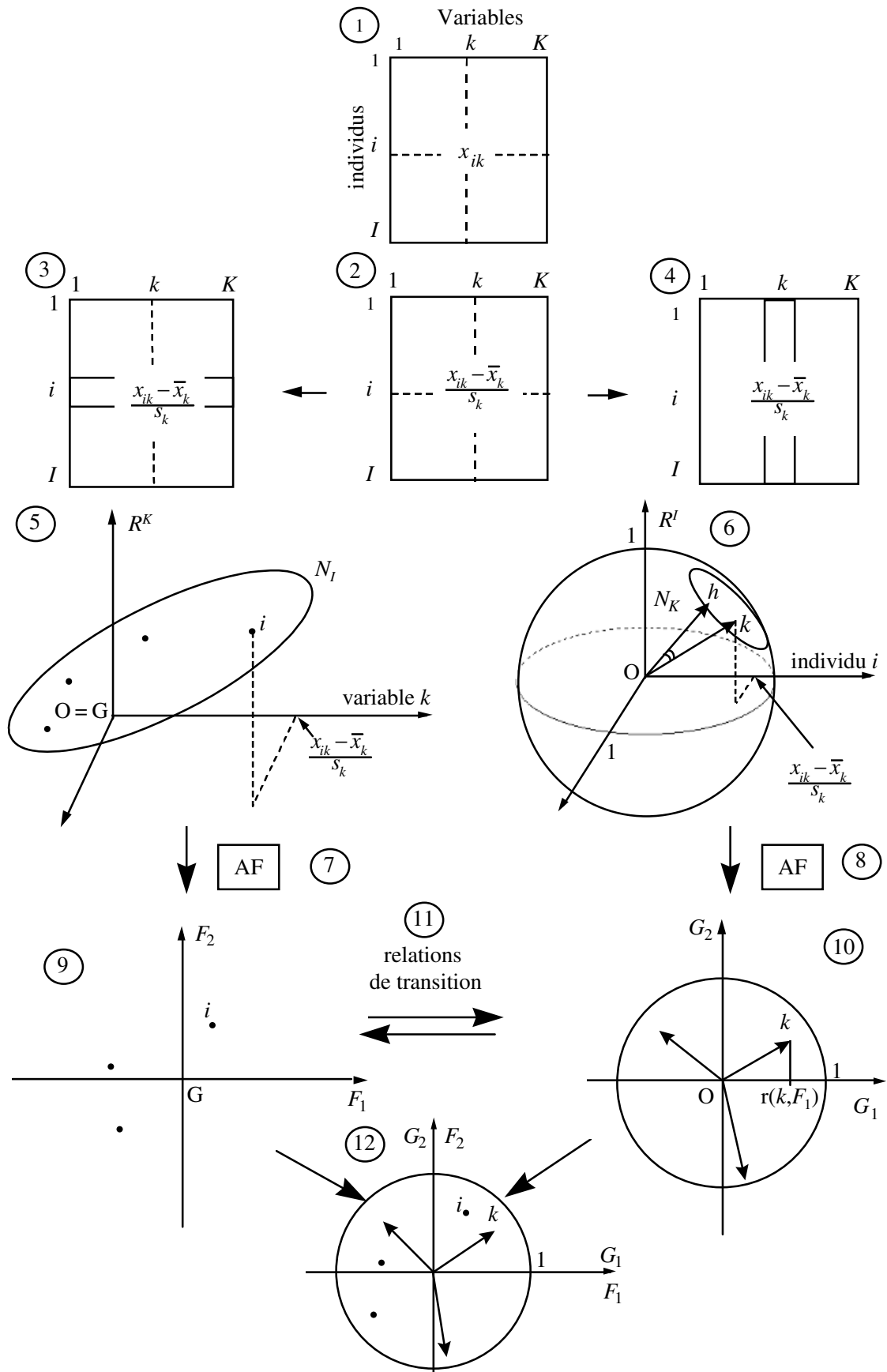


Figure 1.10 Schéma général de l'ACP.

2. Les données centrées et réduites. Que l'on s'intéresse aux individus ou aux variables, le tableau est transformé de la même façon. Le centrage est surtout technique. La réduction permet de s'affranchir de l'arbitraire des unités de mesure.
- 3 et 4. Dans l'étude des individus, le tableau est considéré comme une juxtaposition de lignes. Dans l'étude des variables, le tableau est considéré comme une juxtaposition de colonnes. C'est le même tableau qui est considéré de deux façons différentes.
5. Un individu est une suite de  $K$  nombres et peut être représenté par un point de  $R^K$ . Dans le nuage  $N_I$ , on s'intéresse aux distances inter-individuelles qui s'interprètent comme des ressemblances. Du fait du centrage, l'origine des axes est confondue avec le centre de gravité de  $N_I$ . Dans la plupart des cas, on affecte à chaque individu le même poids :  $1/I$ .
6. Une variable est une suite de  $I$  nombres et peut être représentée par un vecteur de  $R^I$ . Dans le nuage  $N_K$ , on s'intéresse surtout aux angles entre variables. Le cosinus d'un angle entre deux variables s'interprète comme le coefficient de corrélation entre les deux variables. Du fait de la réduction, toutes les variables sont équidistantes de l'origine et donc situées sur une hypersphère de rayon 1.
- 7 et 8. L'Analyse Factorielle (AF) d'un nuage consiste à mettre en évidence une suite de directions telles que l'inertie, par rapport à O, de la projection du nuage sur ces directions est maximum. Dans  $R^K$ , où l'origine O est confondue avec le centre de gravité G, les axes factoriels sont les directions d'allongement maximum de  $N_I$ . Dans  $R^I$ , où la projection d'une variable sur une autre s'interprète comme un coefficient de corrélation, les axes factoriels sont les variables synthétiques les plus liées à l'ensemble des variables initiales.
9. Le plan factoriel croisant deux facteurs sur les individus -ici  $F_1(I)$  et  $F_2(I)$ - fournit une image approchée de  $N_I$  dans  $R^K$ . La distance entre deux points s'interprète comme une ressemblance.
10. Le plan factoriel croisant deux facteurs sur les variables -ici  $G_1(K)$  et  $G_2(K)$ - fournit une image approchée de  $N_K$  dans  $R^I$ . Les coordonnées d'une variable s'interprètent comme des coefficients de corrélation avec les facteurs sur les individus.
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  $R^K$ ).
12. Du fait des relations de transition, les interprétations des axes factoriels doivent être menées simultanément. Il peut être commode de superposer ces deux représentations.

## 1.9 AIDES À L'INTERPRÉTATION

Les axes factoriels fournissent des images approchées d'un nuage de points. Il est donc nécessaire de mesurer la qualité de l'approximation, tant pour chacun des points que pour l'ensemble du nuage. En outre, les plans factoriels représentent les coordonnées des points et non les inerties qui ont présidé à leur détermination. Il est souvent utile de consulter ces inerties. Il en résulte que l'étude d'un plan est toujours réalisée conjointement avec la consultation d'un ensemble d'indicateurs regroupés sous le terme d'aides à l'interprétation. Ce paragraphe définit les principales aides à l'interprétation : le chapitre 2 contient le traitement d'un exemple se référant largement à ces aides ; le chapitre 11 montre comment elles s'insèrent dans une démarche générale d'interprétation.

### 1.9.1 Définitions

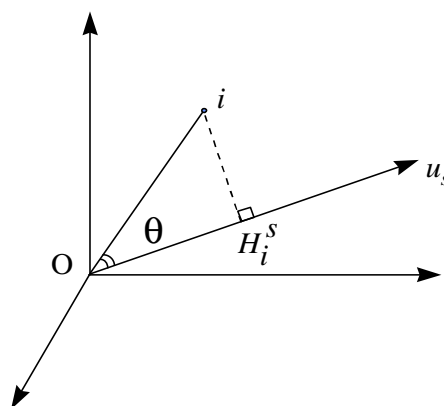
#### a) Qualité de représentation d'un élément par un axe

La qualité de représentation de l'élément  $i$  (individu ou variable) par l'axe  $s$  est mesurée par le rapport :

$$QLT_s(i) = \frac{[\text{inertie de la projection de l'élément } i \text{ sur l'axe } s]}{[\text{inertie totale de } i]}$$

C'est aussi le cosinus carré de l'angle  $\theta$  entre  $Oi$  et l'axe  $s$  (cf. **Figure 1.11**).

$$QLT_s(i) = \frac{(OH_i^s)^2}{(Oi)^2} = \cos^2 \theta$$



**Figure 1.11** Qualité de représentation d'un élément par un axe.  $H_i^s$  : projection de  $i$  sur l'axe de rang  $s$

Cette définition se généralise au cas d'un plan. En outre, du fait de l'orthogonalité des axes factoriels, la qualité de représentation de l'élément  $i$  par le plan (axe  $s$ , axe  $t$ ) est la somme des qualités de représentation de  $i$  par l'axe  $s$  et par l'axe  $t$ . C'est aussi le cosinus carré de l'angle entre le vecteur  $Oi$  et le plan de projection. Si la qualité de représentation d'un point sur un axe ou un plan est proche de 1, ce point est très proche de l'axe ou du plan. S'il s'agit d'un individu, sa distance au centre de gravité (qui est le point moyen) est alors bien visible sur la projection. Elle ne l'est pas dans le cas contraire (lorsque sa qualité de représentation est proche de 0). De même, la distance entre deux points sur un plan ne traduit bien leur distance dans le nuage que si ces deux points sont bien représentés. S'il s'agit d'une variable centrée-réduite, le vecteur a pour norme 1 et sa qualité de représentation est le carré de la longueur de sa projection. Sur un plan, elle s'apprécie directement par sa proximité au cercle de rayon 1, trace de l'hypersphère de rayon 1 sur le plan factoriel. Ce cercle est appelé couramment cercle des corrélations.

#### b) Qualité de représentation d'un nuage par un axe

La définition précédente se généralise à l'ensemble d'un nuage par le rapport :

$$\frac{\text{inertie de la projection du nuage sur l'axe}}{\text{inertie totale du nuage}}$$

Cet indicateur, appelé pourcentage d'inertie associé à un axe, mesure en outre « l'importance » relative d'un axe factoriel dans la variabilité des données.

Comme dans le cas d'un seul élément, ces pourcentages peuvent être cumulés sur plusieurs axes ; on parle alors du pourcentage d'inertie extrait par un plan ou par les  $S$  premiers facteurs. Du fait de la dualité (cf. section 1.7), il est équivalent de calculer ces pourcentages d'inertie à partir du nuage des individus ou de celui des variables.

#### c) Contribution d'un élément à l'inertie d'un axe

Un axe factoriel rend maximum (sous contrainte d'orthogonalité avec les axes précédents) l'inertie projetée d'un nuage. Cette inertie projetée du nuage peut être décomposée point par point. Le quotient de l'inertie de la projection de l'élément  $i$  (de poids  $p_i$ ) sur l'axe  $s$  [soit  $p_i(OH_i^s)^2$ ] par l'inertie de la projection de l'ensemble du nuage sur l'axe  $s$  (soit  $\lambda_s$ ) représente la contribution de l'élément  $i$  à l'inertie de l'axe  $s$ . Soit, en notant  $CTR_s(i)$  la contribution de l'élément  $i$  à l'axe de rang  $s$  :

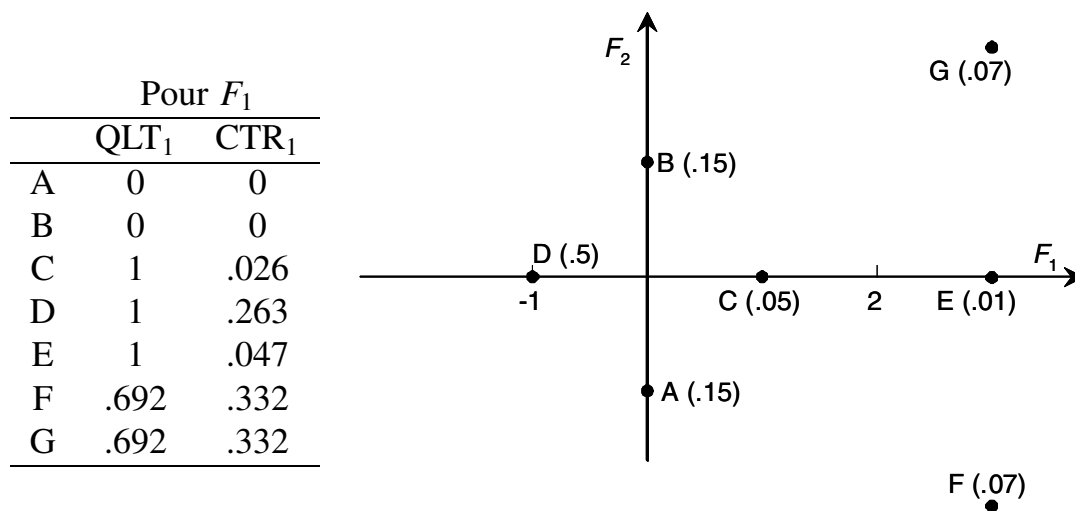
$$CTR_s(i) = \frac{p_i (OH_i^s)^2}{\lambda_s}$$

Cet indicateur se généralise à un sous-ensemble d'éléments. La contribution d'un ensemble de points à l'inertie d'un axe est la somme des contributions des points

qui le composent. Ce rapport est précieux pour mettre en évidence le sous-ensemble d'éléments qui ont contribué principalement à la construction de l'axe et sur lequel s'appuiera en premier lieu l'interprétation.

### 1.9.2 Exemple numérique

Nous présentons ici, sur un exemple artificiel, la façon dont les coordonnées des points et les aides à l'interprétation interviennent dans l'analyse d'un facteur. Sept points du plan, munis de poids, sont représentés dans leurs axes principaux (cf. **Figure 1.12**).



**Figure 1.12** Nuage plan pondéré représenté dans ses axes principaux. Les poids figurent entre parenthèses. QLT<sub>1</sub>, CTR<sub>1</sub> : qualité de représentation et contribution (pour le premier axe).

#### a) Coordonnées sur $F_1$

Les points A, B, et C sont moyens ; D, E, F et G sont extrêmes, D étant opposé à E, F et G. Quelle que soit la qualité de représentation de ces points et leur contribution à l'inertie, cette structure traduite par le premier facteur n'est pas à mettre en doute.

#### b) Qualité de représentation sur $F_1$

$$\text{Exemple : } \text{QLT}_1(G) = \frac{\text{inertie projetée de G}}{\text{inertie totale de G}} = \frac{3^2}{(3^2 + 2^2)} = .692$$

Les points D, C et E, situés sur l'axe, ont une qualité de représentation égale à 1. Leurs distances dans le plan (à l'origine et entre eux) sont complètement traduites dans leur projection sur  $F_1$ . Les points D et E, à la fois extrêmes et bien représentés, sont caractéristiques de l'axe : l'examen de leurs différences avec la moyenne et entre eux permet de préciser l'opposition traduite par  $F_1$ . Réciproquement, toute valeur de E et de D qui s'écarte de la moyenne s'interprète par  $F_1$ .

Les points A et B, situés dans une direction orthogonale à l'axe 1, ont une qualité de représentation sur le premier axe égale à 0 : ni leur écart par rapport à l'origine, ni leur distance dans le plan ne sont visibles sur le premier facteur.

Les points F et G, extrêmes, ont une qualité de représentation moyenne : bien que très marqué pour le facteur  $F_1$ , leur écart à la moyenne n'est qu'en partie traduit par lui.

► Contribution à l'inertie de  $F_1$

Exemple : inertie du nuage ( $\lambda_1$ ) :  $.5(-1)^2 + .05(1)^2 + .01(3)^2 + .07(3)^2 + .07(3)^2 = 1.9$

$$CTR_1(F) = \text{inertie du point F} / \text{inertie du nuage} = (.07 \times 3^2) / 1.9 = .332$$

Les points A et B ont une coordonnée nulle, donc une contribution nulle. Le point C est proche de O et a un petit poids : sa contribution est extrêmement faible. La suppression de ces trois points ne modifierait pas la direction du premier facteur.

Les points E et F ont la même coordonnée mais E, ayant un poids 7 fois plus faible que F, a une contribution 7 fois plus faible. La suppression de E risque moins de modifier le facteur que celle de F, pourtant moins bien représenté.

Le point D, malgré son poids égal à plus de 7 fois celui de F, a une contribution plus faible car il est situé plus près de l'origine (dans la contribution à l'inertie, la distance intervient par son carré alors que le poids intervient tel quel).

## 1.10 VARIABLES QUALITATIVES ILLUSTRATIVES EN ACP

On est souvent conduit à vouloir relier les résultats d'une ACP à des variables qualitatives définies sur les individus.

**Exemple :** On étudie les notes obtenues à différentes épreuves par un ensemble d'élèves. L'ACP de ce tableau met en évidence les principales dimensions de variabilité des élèves, par exemple une opposition entre les élèves plutôt meilleurs dans les matières scientifiques et ceux plutôt meilleurs dans les matières littéraires. On dispose par ailleurs d'informations sur ces élèves sous forme de variables qualitatives, par exemple leur genre (fille/garçon), la catégorie socio-professionnelle des parents, etc. Il est utile de relier ces variables qualitatives aux axes factoriels, avec en perspective des questions du type : observe-t-on, sur ces données, l'idée souvent émise selon laquelle les filles obtiennent des résultats plutôt meilleurs dans les matières littéraires et les garçons des résultats plutôt meilleurs dans les matières scientifiques ?

Pour cela, on dispose de deux outils graphiques simples et efficaces :

- identification, sur les plans factoriels, des individus par leur modalité à l'aide d'un code, de couleur ou de forme (dans l'exemple on pourra identifier les filles par un point rose et les garçons par un point bleu !); cela permet d'étudier finement

la relation entre une variable qualitative et le plan factoriel mais nécessite un graphique par variable ;

- représentation, sur les plans factoriels, des centres de gravité des ensembles d'individus possédant une même modalité (dans l'exemple, le centre de gravité des filles et celui des garçons) ; à la différence de la technique précédente, un seul graphique permet d'examiner plusieurs variables qualitatives simultanément, mais, en revanche, ne donne pas d'informations quant à la variabilité des individus présentant une même modalité.

On peut chercher à traduire la variabilité des individus autour des centres de gravité des variables qualitatives en terme de variabilité des centres de gravité eux-mêmes. Pour cela, on construit, autour de chaque centre de gravité, une ellipse de confiance, analogue bi-dimensionnel de l'intervalle de confiance que l'on calcule usuellement autour d'une moyenne. Pour produire ces ellipses, on procède de la façon suivante :

1. On considère l'ensemble  $I$  des  $I$  individus observés, comme un échantillon d'une population plus vaste, dite de référence ; dans cette perspective, la variabilité des individus se traduit par une variabilité des centres de gravité induite par le fait que l'ensemble  $I$  observé n'est que l'un des ensembles possibles de  $I$  individus parmi la population de référence.
2. La variabilité des centres de gravité pourrait être obtenue en extrayant d'autres échantillons de la population de référence mais cela est généralement impossible ; aussi approxime-t-on la population de référence par l'ensemble  $I$  et l'on tire, au hasard avec remise, plusieurs échantillons de  $I$  individus dans cet ensemble ; cette procédure est appelée « bootstrap ».
3. Pour chaque échantillon « bootstrap », on calcule les centres de gravité des différentes modalités et l'on projette ces centres de gravité (dits bootstrap) en supplémentaire sur les plans de l'ACP (initiale) des  $I$ .
4. Si l'on effectue  $n$  tirages bootstrap, on obtient, pour une modalité donnée,  $n$  points ; on pourrait se contenter de représenter ces  $n$  points mais les graphiques obtenus sont peu lisibles dès lors que le nombre de modalités étudiées est un tant soit peu grand ; pour simplifier les représentations, on construit l'ellipse centrée sur le centre de gravité initial et contenant 95 % des  $n$  centres de gravité bootstrap ; ces ellipses sont dites « ellipses de confiance bootstrap ». L'expérience montre que schématiser la distribution des  $n$  points par une ellipse n'est pas gênant (au sens où, en pratique, l'observation du nuage des  $n$  points ne conduit pas à des interprétations plus riches) dès lors que l'effectif par modalité est assez grand (disons une vingtaine d'individus pour fixer les idées).

► Remarque

La taille d'une ellipse ainsi obtenue dépend de la variabilité (dans le plan factoriel) des individus présentant la modalité étudiée mais aussi de son effectif.

L'utilisation pratique des ellipses de confiance s'articule autour de deux questions relatives aux modalités.

**La modalité  $m$  est-elle caractérisée par le plan factoriel ?** Autrement dit, les individus possédant la modalité  $m$  occupent-ils (dans l'ensemble) une position excentrée sur le plan ? Pour cela, on examine la position de l'origine des axes, centre de gravité de l'ensemble  $I$ , par rapport à l'ellipse de confiance de  $m$ . Si cette ellipse englobe l'origine, on décidera que la modalité  $m$  (*i.e.* les individus possédant cette modalité) n'est pas caractérisée par le plan.

**Les deux modalités  $m$  et  $m'$  sont-elles différenciées par le plan ?** Autrement dit, les individus possédant la modalité  $m$  occupent-ils, dans l'ensemble, la même région du plan que ceux possédant la modalité  $m'$  ? Pour cela, on examine le recouvrement entre les deux ellipses associées aux modalités  $m$  et  $m'$ . Une absence de recouvrement conduit à décider que le plan différencie les deux modalités et, à l'inverse, un fort recouvrement conduit à décider d'une non différenciation. Un recouvrement faible laisse la place au doute : pour aider sa décision, l'utilisateur peut calculer la probabilité critique du test statistique «  $T^2$  de Hotelling » appliqué à la comparaison des deux modalités du point de vue des deux composantes principales étudiées considérées simultanément.

Les questions concernant la position des modalités sur un plan peuvent être posées pour chaque axe. Pour cela, en projetant les ellipses sur chaque axe, on obtient un intervalle de confiance que l'on peut utiliser comme un intervalle de confiance usuel. Il existe aussi un indicateur, appelé valeur-test et introduit initialement dans le logiciel SPAD, qui permet de juger, pour un axe factoriel (et, plus généralement pour n'importe quelle variable), de l'écart entre le centre de gravité d'une classe et le centre de gravité général (*cf.* section 2.4.4 page 54).