Brigitte Escofier, Jérôme Pagès

Analyses factorielles simples et multiples

Cours et études de cas

5e ÉDITION

DUNOD

Illustration de couverture :
african seamless pattern © alexvv – fotolia.com

11 rue Paul Bert, 92240 Malakoff
www.dunod.com
EAN 9782100859580

Table des matières

Introduction

L'analyse des données : outil de connaissance dans les domaines les plus divers

Les méthodes d'analyse des données ont largement démontré leur efficacité dans l'étude de grandes masses complexes d'informations. Ce sont des méthodes dites multidimensionnelles en opposition aux méthodes de la statistique descriptive qui ne traitent qu'une ou deux variables à la fois. Elles permettent donc la confrontation entre de nombreuses informations, ce qui est infiniment plus riche que leur examen séparé. Les représentations simplifiées de grands tableaux de données que ces méthodes permettent d'obtenir s'avèrent un outil de synthèse remarquable. De données trop nombreuses pour être appréhendées directement, elles extraient les tendances les plus marquantes, les hiérarchisent et éliminent les effets marginaux ou ponctuels qui perturbent la perception globale des faits.

Nées à l'université, elles ont d'abord été connues essentiellement des chercheurs et appliquées à des domaines scientifiques comme l'écologie, la linguistique, l'économie, etc. Elles ont permis d'aborder des études nouvelles plus riches et plus complexes. Mais leur domaine d'application déborde depuis longtemps ce cadre universitaire, surtout depuis que l'acquisition et le stockage des informations sont facilités par le développement de l'informatique. Dans tous les domaines (marketing, assurance, banque, etc.), d'importants fichiers de données sont accumulés. Le premier objectif est de conserver les informations et de pouvoir les consulter facilement. Mais on s'aperçoit vite que pour exploiter l'ensemble de l'information contenue dans ces fichiers, dont le recueil est souvent coûteux, il est nécessaire de disposer d'outils statistiques adaptés.

Puissance des représentations géométriques de l'analyse factorielle

Parmi les méthodes de l'analyse des données, l'analyse factorielle tient une place primordiale. Elle est utilisée soit seule, soit conjointement avec des méthodes de classification (alors que ces dernières sont rarement appliquées seules). Cette place de choix tient en partie aux représentations géométriques des données, qui transforment en distances euclidiennes des proximités statistiques entre éléments.

Elles permettent d'utiliser les facultés de perception dont nous usons quotidiennement : sur les graphiques de l'analyse factorielle, on voit, au sens propre du terme

(avec les yeux et l'analyse assez mystérieuse que notre cerveau fait d'une image), des regroupements, des oppositions, des tendances, impossibles à discerner directement sur un grand tableau de nombres, même après un examen prolongé.

Ces représentations graphiques sont aussi un moyen de communication remarquable car point n'est besoin d'être statisticien pour comprendre que la proximité entre deux points traduit la ressemblance entre les objets qu'ils représentent.

L'analyse factorielle ou les analyses factorielles ?

Les deux expressions se justifient.

1. Il existe plusieurs méthodes adaptées à différents types de données : ainsi, pour citer les plus connues, l'analyse en composantes principales (ACP) traite des tableaux croisant des individus et des variables quantitatives, l'analyse factorielle des correspondances (AFC) traite des tableaux de fréquence et l'analyse des correspondances multiples (ACM) s'applique à des tableaux croisant des individus et des variables qualitatives.
2. Le principe de ces méthodes est unique. Deux nuages de points, représentant respectivement les lignes et les colonnes du tableau étudié, sont construits et représentés sur des graphiques. Les représentations des lignes et des colonnes sont fortement liées entre elles.

Rigueur et souplesse des méthodes d'analyse factorielle

Le fait que l'analyse factorielle ne s'applique qu'à des tableaux rectangulaires peut paraître au premier abord une limitation importante à la fois sur le type de données et sur la manière de les aborder. En réalité, la plupart des études de données peuvent être formalisées comme une analyse de tableaux rectangulaires. D'autre part, un même fichier de données peut conduire à un grand nombre de tableaux différents et donc à des analyses différentes qui permettent chacune d'étudier un des aspects du problème.

La construction de tableaux à partir d'un fichier initial est appelée codage. Ce terme de codage inclut la transformation de données brutes en variables quantitatives ou qualitatives, le choix des lignes et des colonnes du tableau, celui des éléments à traiter en actif, etc. Dans cette étape de codage, la marge de manœuvre est presque infinie. Le résultat d'une analyse factorielle est unique, ce qui en assure la rigueur, mais les analyses possibles sont nombreuses, ce qui en assure la souplesse et la faculté d'adaptation.

Les tableaux multiples

Les analyses factorielles ont été conçues pour étudier un tableau de données unique. Or, les personnes qui analysent des données sont de plus en plus fréquemment confrontées à l'étude simultanée de plusieurs tableaux rectangulaires. Il s'agit le plus souvent :

1. d'une suite de tableaux indicés par le temps ;
2. d'un ensemble de tableaux rectangulaires provenant d'un unique tableau de dimension trois ;
3. d'un tableau initialement unique mais dans lequel on distingue des sous–tableaux (ce cas général inclut le cas particulier dans lequel un ensemble d'individus est décrit à la fois par des variables quantitatives et des variables qualitatives).

Au fil des ans, des méthodologies ont été mises au point. On se ramène généralement à l'analyse d'un tableau complexe formé par la juxtaposition des différents tableaux. Ces méthodes fondées sur les méthodes d'analyse classique, elles-mêmes conçues pour l'étude d'un tableau simple, utilisent largement la technique dite des « éléments supplémentaires ». Mais ces techniques ont leurs limites et les objectifs spécifiques de l'analyse des « tableaux multiples » ne sont pas tous atteints. Aussi, de nouvelles méthodes, utilisant les mêmes principes fondamentaux que les analyses factorielles « classiques » mais prenant en compte le caractère « multiple » des tableaux, ont été mises au point.

Esprit du livre

Cet ouvrage est destiné avant tout aux utilisateurs d'analyse des données. C'est pourquoi il présente des méthodes d'analyse factorielle en tentant de dégager leurs objectifs et les interprétations de leurs résultats. Pour en faciliter la lecture aux non-spécialistes, nous avons pris le parti de séparer le plus possible les aspects intuitifs des méthodes (objectifs, principe général et représentations géométriques), des aspects mathématiques et théoriques. Les aspects intuitifs ne nécessitent qu'un très faible bagage statistique et mathématique et sont donc abordables par beaucoup. Ils sont largement commentés sur plusieurs exemples.

Les aspects théoriques sont regroupés essentiellement dans deux chapitres. Leur but est de fournir les justifications des méthodes en précisant les critères optimisés et les algorithmes de calcul. La bibliographie est restreinte au minimum : lorsqu'une démonstration risque d'alourdir trop le texte, une note en bas de page renvoie à une référence plus complète.

Les objectifs. Devant un jeu de données à analyser, se pose le problème du choix du traitement statistique, c'est-à-dire du choix du couple indissociable codage-méthode. Pour bien choisir, il est nécessaire de connaître les moyens dont on dispose, donc les possibilités des méthodes qui peuvent répondre chacune à un certain nombre d'objectifs précis. La réflexion sur les objectifs d'une étude est fondamentale. Elle est plus efficace si elle se fait dans le cadre des possibilités techniques. Cette réflexion doit toujours intervenir le plus tôt possible car elle influe non seulement sur le traitement statistique mais aussi sur le recueil même des données.

L'interprétation. L'analyse effectuée, le travail du statisticien n'est pas terminé : il faut interpréter les résultats. Cette phase fait intervenir à la fois la connaissance du problème et celle des méthodes.

Contenu du livre

Ce livre contient à la fois un rappel des méthodes classiques, des exposés des méthodologies d'analyse des tableaux multiples basées sur ces dernières et une introduction aux méthodes d'analyse spécifiques de ces tableaux. Ces dernières ont été conçues par les auteurs et exposées dans le cadre de leurs recherches, mais cet ouvrage est le premier qui en contient une présentation générale destinée aux utilisateurs. L'interprétation des résultats d'une analyse factorielle, qui est avec le codage la phase la plus délicate de l'étude, est illustrée par plusieurs exemples tout le long du texte ; elle fait aussi l'objet d'une réflexion générale.

La première partie du livre, qui comprend cinq chapitres, présente les méthodes classiques d'analyse factorielle : l'ACP, l'AFC et l'ACM. Le traitement d'un exemple par ACP donne l'occasion de présenter une méthode de classification et son dépouillement conjointement avec celui d'une analyse factorielle. La fin du chapitre consacré à l'ACM est dédiée à l'analyse factorielle de données mixtes (AFDM), méthode peu connue qui traite des mélanges de variables quantitatives et qualitatives. Enfin, une présentation formalisée de l'ACP, de l'AFC et de l'ACM, incluant les démonstrations essentielles, est faite dans un cadre commun à ces trois méthodes.

La deuxième partie est consacrée aux tableaux multiples. Les chapitres 6, 7, 8 et 9 concernent l'étude simultanée de plusieurs tableaux croisant les mêmes individus et différents groupes de variables numériques ou qualitatives. Le chapitre 6 commente plusieurs traitements de la même enquête par les méthodes classiques. C'est à la fois une illustration des méthodes présentées dans les premiers chapitres, une réflexion sur les objectifs généraux de l'étude de tableaux comprenant plusieurs groupes de variables, et un bilan sur l'intérêt et les limites des méthodologies basées sur ces méthodes. L'analyse factorielle multiple (AFM), conçue pour ce type de données, est introduite dans le chapitre 7 à partir des résultats issus de son application à un second exemple ; sa présentation complète constitue le chapitre 8 ; une réflexion sur son utilisation constitue le chapitre 9. Le chapitre 10 traite des tableaux de fréquence ternaires et plus généralement de l'étude simultanée de plusieurs tableaux de fréquence binaires. Bien qu'il s'agisse comme dans les quatre chapitres précédents de tableaux multiples, la nature des données (fréquences au lieu de variables) implique des objectifs fondamentalement différents. Ce chapitre tente d'en dégager les principaux et illustre sur un même exemple les méthodologies dérivées de l'AFC et une technique nouvelle, baptisée *analyse intra*, qui permet d'étudier un aspect spécifique des tableaux de fréquence ternaire : les liaisons conditionnelles.

La dernière partie commence par un chapitre entièrement consacrée à l'interprétation des résultats en analyse factorielle. Elle est issue en partie des réflexions d'un groupe de travail[1] réuni par l'ADDAD[2] dans le cadre d'un contrat avec la Société THOMSON. À partir des expériences confrontées et du regroupement de commentaires épars d'applications d'analyse factorielle, nous avons construit un guide. Ce guide propose une démarche générale d'interprétation en analyse factorielle en différenciant ACP, AFC, ACM et AFM. Cette démarche est appliquée dans le chapitre suivant qui regroupe deux études de cas. Cette partie se termine par les corrigés des exercices proposés à la fin de certains chapitres.

Il est conseillé aux lecteurs novices en analyse des données de commencer la lecture de cet ouvrage par les deux premières fiches techniques incluses dans le chapitre 13. Ces deux fiches détaillent les représentations géométriques des nuages d'individus et de variables utilisées systématiquement en analyse factorielle. La troisième fiche, plus technique, est destinée plutôt aux lecteurs qui souhaitent approfondir les aspects mathématiques et théoriques développés dans les chapitres 5 et 8.

L'index systématique reprend l'ensemble des notions essentielles.

Note sur la cinquième édition

Par rapport à la précédente, cette cinquième édition comporte principalement deux nouveautés. Des exercices corrigés ont été ajoutés. À partir de données particulièrement simples, on met en œuvre la plupart des méthodes (ACP, ACM, AFDM et AFM) sans le recours à un programme, en travaillant « à la main » en quelque sorte. Ceci est possible parce que ces données très simples possèdent des propriétés géométriques très particulières. C'est l'occasion de voir les méthodes « de l'intérieur ». Pour réaliser ces analyses, le lecteur est guidé pas à pas au moyen d'une série de questions permettant une auto-évaluation tout en conférant à l'ensemble un aspect presque ludique. Une mention particulière peut être faite pour l'AFDM : l'exercice permet de comprendre en profondeur comment analyser simultanément des variables quantitatives et qualitatives. Chaque énoncé d'exercice est situé à la fin du chapitre présentant la méthode utilisée. Les corrigés sont regroupés dans un chapitre final. Un autre nouveau chapitre présente de façon détaillée deux études de cas utilisant respectivement l'AFM et l'AFMH. Chaque étude de cas est présentée sous la forme d'un exercice comportant une série de questions guidant le lecteur, d'abord dans la formulation des objectifs, puis dans la définition de la méthodologie et, enfin, dans l'exploration progressive des données. Pour le lecteur désireux de réaliser lui-même ces analyses, les données complètes sont

1. Ch. Bastin, Ch. Bourgarit, J. Confais, B. Escofier, B. Gomel, J.P. Fénelon, J.Pagès.

2. L'Association pour le Développement et la Diffusion de l'Analyse des Données, créée par J.-P. Fénelon et aujourd'hui dissoute, fédérait bon nombre des fondateurs de l'École française d'analyse des données.

disponibles sur le site du laboratoire de mathématiques appliquées d'Agrocampus (http://math.agrocampus-ouest.fr).

Sans oublier Gabriel Jalam, ingénieur informaticien à Agrocampus, qui mit en forme la précédente édition, pour cette cinquième édition, nous sommes redevable à Magalie Houée-Bigot, ingénieure au laboratoire de mathématiques appliquées d'Agrocampus, qui a assuré la mise en forme des nouveautés. Qu'elle soit ici remerciée pour son efficacité et son humeur toujours égale.

Logiciels

L'analyse factorielle multiple, qui est au cœur de cet ouvrage, est maintenant disponible dans de nombreux logiciels. Citons, par ordre alphabétique : ADE4 (Package R), FactoMineR (Package R), SPAD, Uniwin (Statgraphics), XLStat et %AFMULT (macro SAS écrite par B. Gelein et O. Sautory).

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*×*Variables quantitatives*.

Selon un usage bien établi, les lignes du tableau représentent les individus et les colonnes représentent les variables. À 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 36 en est un exemple.

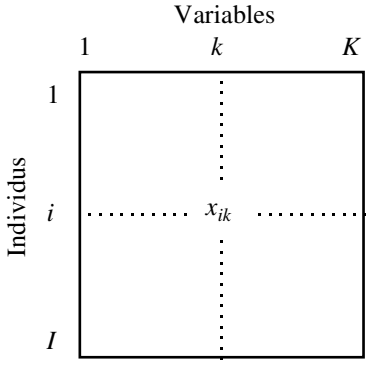

Figure 1.1 Tableau des données en ACP. x_{ik} : valeur de la variable k pour l'individu i. I : nombre d'individus et ensemble des individus. K : nombre de variables et ensemble des variables.

Les termes *individu* et *variable* recouvrent des notions différentes. Par exemple, dans le tableau étudié au chapitre 7, 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 :

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

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 :

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

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 variablité** 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 variablité 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 :

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

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). À 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 $\mathrm{d}(i,l)$ 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 36. 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**.

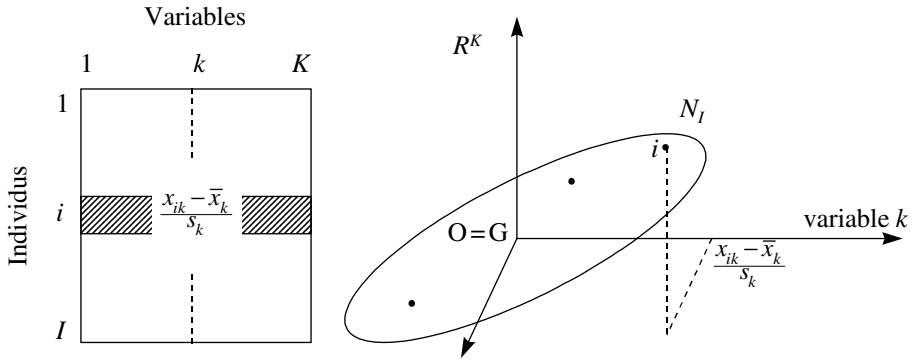

Figure 1.2 Tableau des données et nuage des individus associé dans l'espace R^K. Du fait du centrage, l'origine des axes est confondue avec le centre de gravité du nuage.

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^{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**.

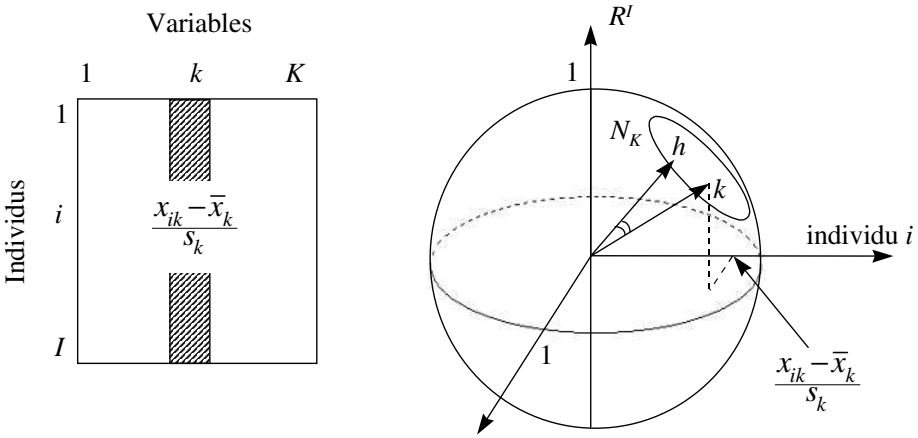

Figure 1.3 Tableau des données et nuage des variables associé dans l'espace R^I.
$\cos(\widehat{Oh}, \widehat{Ok}) = r(h,k)$

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 :

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

L'interprétation d'un coefficient de corrélation comme un cosinus est une propriété très importante puisqu'elle donne un support géométrique, donc visuel, au coefficient de corrélation. Cette propriété nécessite le centrage, ce qui justifie cette transformation présentée section 1.1 comme un intermédiaire technique. Elle justifie aussi le choix de la distance (on dit aussi *métrique*) dans R^I et implique que, dans la représentation des variables, on s'intéresse surtout aux directions déterminées par les variables, c'est-à-dire aux vecteurs plutôt qu'à leurs extrémités.

La longueur des vecteurs représentant les variables étant égale à 1, la coordonnée de la projection d'une variable sur une autre s'interprète comme un coefficient de corrélation.

➤ Conclusion

Réaliser un bilan des coefficients de corrélation entre les variables revient à étudier les angles entre les vecteurs définissant le nuage N_K. Cette étude directe est impossible du fait de la dimension de R^I. L'intérêt de l'ACP est de fournir des variables synthétiques qui constituent un résumé de l'ensemble des variables initiales et sont la base d'une représentation plane approchée des variables et de leurs angles.

1.5 AJUSTEMENT DU NUAGE DES INDIVIDUS

L'objectif est de fournir des images planes approchées du nuage N_I situé dans l'espace R^K (cf. section 1.3). Pratiquement, on recherche une suite $\{u_s; s = 1, ..., S\}$ de S directions privilégiées de R^K appelées axes factoriels qui, prises deux à deux, définissent des plans factoriels sur lesquels on projette le nuage N_I. Chaque direction u_s est choisie de façon à rendre maximum l'inertie par rapport à l'origine O (confondue avec le centre de gravité G, du fait du centrage) de la projection de N_I sur u_s. Dans la recherche d'une suite, on impose à chaque direction d'être orthogonale aux directions déjà trouvées (cf. **Figure 1.4**). On peut montrer que le plan engendré par les deux