Pratique R

Analyse factorielle multiple avec R

Jérôme Pagès

edp sciences

Analyse factorielle multiple avec R

This page intentionally left blank

Jérôme Pagès

Analyse factorielle multiple avec R

edp sciences

ISBN : 978-2-7598-0963-9

Imprimé en France

Collection Pratique R

dirigée par Pierre-André Cornillon et Eric Matzner-Løber

Département MASS
Université Rennes-2-Haute-Bretagne
France

Comité éditorial

Déjà paru dans la même collection :

Psychologie statistique avec R

Yvonnick Noël, 2013
ISBN : 978-2-8178-0425-5 – Springer

Séries temporelles avec R

Yves Aragon, 2011
ISBN : 978-2-8178-0208-4 – Springer

Régression avec R

Pierre-André Cornillon, Eric Matzner-Løber, 2011
ISBN : 978-2-8178-0184-1 – Springer

Méthodes de Monte-Carlo avec R

Christian P. Robert, George Casella, 2011
ISBN : 978-2-8178-0181-0 – Springer

This page intentionally left blank

AVANT-PROPOS

Du fait de son large domaine d'application, l'analyse factorielle multiple (AFM) est maintenant largement utilisée. Le présent livre est un exposé complet de la méthode : il rassemble les aspects théoriques et méthodologiques, des exemples d'application et la mise en œuvre logicielle *via* un package R (**FactoMineR**).

De même que l'analyse en composantes principales (ACP) ou l'analyse des correspondances multiples (ACM), l'AFM s'applique à des tableaux structurés ainsi : pour un ensemble d'individus (un individu = une ligne), on dispose des valeurs de chacun pour un ensemble de variables (une variable = une colonne). La spécificité de l'AFM tient à la prise en compte, au sein des variables actives, d'une structure en groupes définie par l'utilisateur. De telles données sont désignées par « tableau individus × variables structurées en groupes ».

Ce format de données est fréquent. En premier lieu parce qu'il correspond bien à la démarche de l'utilisateur lorsqu'il conçoit un recueil de données. Ainsi, le rédacteur d'un questionnaire d'opinion procède par thèmes, qu'il décline chacun selon plusieurs questions (les questions seront les variables et les thèmes les groupes de variables). Cette structure du questionnaire doit bien sûr être présente lors de l'analyse des résultats. En second lieu, parce que l'utilisateur souhaite souvent rapprocher des données recueillies sur les mêmes individus statistiques mais dans des contextes (par exemple géographiques et/ou temporels) différents. Ainsi, dans l'industrie agroalimentaire, on dispose souvent, pour un même ensemble de produits, de profils sensoriels issus de dégustations réalisées dans plusieurs pays et/ou par des (types de) dégustateurs différents. Il est nécessaire d'analyser simultanément ces ensembles de données tout en préservant leur individualité, lors de l'analyse statistique d'abord et lors de l'interprétation ensuite.

L'expérience, issue de nombreux contacts avec des utilisateurs variés, montre que les tableaux multiples constituent en fait le format standard des données auxquelles on est confronté aujourd'hui lorsque l'on applique des méthodes factorielles. A cette complexité de structure (en groupes de variables), s'ajoute une complexité de nature, les variables pouvant être quantitatives et/ou qualitatives. Il est donc nécessaire de disposer d'une méthodologie d'analyse de tableaux individus × variables dans lesquels les variables sont structurées en groupes quantitatifs, qualitatifs ou mixtes. Tel est précisément le champ d'application de l'AFM.

L'AFM est le résultat d'un travail conjoint de Brigitte Escofier et de Jérôme Pagès au début des années 1980. Cette méthode est maintenant bien établie si l'on en juge par sa disponibilité logicielle croissante. Citons, sans prétendre à l'exhaustivité, quelques logiciels incluant une procédure d'AFM : SPAD, **FactoMiner** (Package R), **ade4** (Package R), Uniwin (Statgraphics), XLStat.

La disponibilité de la méthode étant acquise, la fréquence du format des données justifiant sa mise en œuvre s'imposant d'elle-même, il reste encore à aider l'utilisateur à appréhender ses données dans leur complexité. Pour cela, une question est centrale : que signifie précisément « prendre en compte la structure en groupes de variables dans une analyse d'ensemble » ? Autrement dit, pourquoi ne

pas mettre en œuvre une analyse factorielle usuelle, par exemple une analyse en composantes principales, et tenir compte de la structure en groupes de variables uniquement dans l'interprétation. En étant un peu réducteur, on pourrait dire que ce livre ne répond qu'à cette seule question.

Les deux premiers chapitres reprennent les méthodes de base de l'analyse factorielle d'un tableau individus × variables, ACP et ACM.

Le chapitre 3 traite de l'analyse factorielle simultanée de variables quantitatives et qualitatives, sans distinction de groupes. La méthode décrite, dite AFDM (analyse factorielle de données mixtes), est peu connue ; elle est l'occasion d'introduire les éléments techniques permettant de prendre en compte les deux types de variables au sein d'une analyse unique.

Les chapitres suivants, numérotés de 4 à 10, décrivent l'analyse factorielle multiple. Les quatre premiers abordent successivement les points clés de l'AFM dans le cadre de variables quantitatives. Un chapitre est dédié aux données qualitatives et mixtes. Enfin, deux chapitres comparent chacun l'AFM à une méthode de référence pour des questions spécifiques : la méthode Statis et l'analyse procustéenne.

Le chapitre 11 présente une extension naturelle de l'AFM : l'AFM hiérarchique (AFMH). Dans cette méthode, les variables ne sont pas structurées par une simple partition, mais par une hiérarchie ou, si l'on préfère, une suite de partitions emboîtées. Un exemple typique de ces données est fourni par les enquêtes dont le questionnaire est structuré en thèmes et sous-thèmes.

Enfin, le chapitre 12 présente, sous la forme de deux fiches, quelques éléments de calcul matriciel et d'espaces métriques utilisés dans ce livre.

***

Au terme de cet ouvrage, il m'est agréable de remercier Sophie Puyo, ingénieure statisticienne, qui a assuré l'essentiel de la mise en forme de ce livre. Première lectrice de ce travail, elle a été aussi à l'origine de bon nombre d'améliorations. Ces remerciements s'adressent aussi à Magalie Houée-Bigot, ingénieure statisticienne, qui a pris le relais de Sophie après l'intervention des relecteurs. Je remercie aussi tout particulièrement Eric Matzner-Løber pour l'accueil qu'il a su réserver à ce livre et les échanges que cela a occasionnés. Il est juste enfin de remercier Annie, mon épouse, qui éclaire ma vie et donc, indirectement, ce livre.

***

Les données utilisées dans ce livre sont disponibles sur le site du laboratoire de mathématiques appliquées d'Agrocampus Ouest.

Les chapitres 3, 8, 9 et 10 reprennent, en les adaptant au format d'un livre, des travaux initialement publiés dans la Revue de statistique appliquée (dont la publication s'est arrêtée en 2006). C'est là une excellente occasion de remercier Pierre Cazes, directeur de cette revue, d'abord pour l'excellent accueil qu'il fit à ces travaux et ensuite pour son encouragement à les reprendre dans un livre.

Les calligraphies sont dues au talent de Richard Delécolle.

Table des matières

1 Analyse en composantes principales **1**

1.1 Données, notations 1
1.2 Pourquoi analyser un tableau par ACP ? 2
1.3 Nuages des individus et des variables 3
1.4 Centrage et réduction 6
1.5 Ajustement des nuages N_I et N_K 7
1.5.1 Principe général et formalisation des critères 8
1.5.2 Interprétation des critères 9
1.5.3 Solution 10
1.5.4 Relations entre les analyses des deux nuages 12
1.5.5 Représentation des variables 14
1.5.6 Nombre d'axes 15
1.6 Aides à l'interprétation 15
1.6.1 Pourcentage d'inertie associé à un axe 15
1.6.2 Contribution d'un point à l'inertie d'un axe 16
1.6.3 Qualité de représentation d'un point par un axe 16
1.7 Premier exemple : 909 candidats au bac 17
1.7.1 Inerties projetées 17
1.7.2 Interprétation des axes 18
1.7.3 Remarques méthodologiques 20
1.8 Eléments supplémentaires 22
1.9 Variables qualitatives en ACP 24
1.10 Second exemple : six jus d'orange 27
1.11 ACP dans FactoMineR 29

2 Analyse des correspondances multiples **37**

2.1 Données 37
2.2 Tableau disjonctif complet 38
2.3 Questionnement 39
2.4 Nuages des individus et des variables 40
2.4.1 Nuage des individus 41
2.4.2 Nuage des modalités 43

2.4.3 Variables qualitatives 44
2.5 Ajustement des nuages N_I et N_K 46
2.5.1 Nuage des individus 46
2.5.2 Nuage des modalités 48
2.5.3 Relations entre les deux analyses 49
2.6 Représentation des individus, des modalités et des variables 50
2.7 Aides à l'interprétation 52
2.8 Exemple : 25 étudiants évaluent 5 outils pédagogiques 53
2.8.1 Données 53
2.8.2 Analyse et représentations 54
2.8.3 Comparaison ACM/ACP pour des variables ordinales . . . 57
2.9 ACM dans FactoMineR 59

3 Analyse factorielle de données mixtes 65
3.1 Données, notations 66
3.2 Représentation des variables 66
3.3 Représentation des individus 68
3.4 Relations de transition 69
3.5 Mise en œuvre 70
3.6 Exemple : biométrie de six individus 70
3.7 AFDM dans FactoMineR 73

4 Pondération des groupes de variables 77
4.1 Problématique 77
4.2 Exemple numérique introductif 79
4.3 Pondération des variables en AFM 80
4.4 Application aux six jus d'orange 84
4.5 Relations avec les analyses partielles 86
4.6 Conclusion 88
4.7 AFM dans FactoMineR (premiers résultats) 89

5 Comparaison de nuages d'individus partiels 97
5.1 Problématique 97
5.2 Méthode 100
5.3 Application aux six jus d'orange 102
5.4 Aides à l'interprétation 104
5.5 Distorsions dans la représentation superposée 106
5.5.1 Exemple 106
5.5.2 Interprétation géométrique 108
5.5.3 Approche algébrique 110
5.6 Conclusion sur la représentation superposée 112
5.7 Nuages partiels de l'AFM dans FactoMineR 112

6 Facteurs communs **115**
6.1 Problématique . . . 115
6.1.1 Mesure de liaison entre une variable et un groupe . . . 116
6.1.2 Facteur commun à plusieurs groupes de variables . . . 117
6.1.3 Retour sur les six jus d'orange . . . 117
6.1.4 Analyse canonique . . . 119
6.2 Liaison entre variable et groupe de variables . . . 119
6.3 Recherche de facteurs communs . . . 121
6.4 Recherche de variables canoniques . . . 122
6.5 Aides à l'interprétation . . . 123
6.5.1 Mesure de liaison Lg . . . 123
6.5.2 Coefficients de corrélation canoniques . . . 123

7 Comparaison des groupes de variables **125**
7.1 Nuage N_J des groupes de variables . . . 125
7.2 Produit scalaire, liaison entre groupes de variables . . . 127
7.3 Norme dans l'espace des groupes de variables . . . 129
7.4 Représentation approchée du nuage N_J . . . 130
7.4.1 Principe . . . 130
7.4.2 Critère . . . 132
7.5 Aides à l'interprétation . . . 133
7.6 Modèle Indscal . . . 134
7.6.1 Modèle . . . 135
7.6.2 Estimation des paramètres et propriétés . . . 136
7.6.3 Exemple d'application du modèle Indscal *via* l'AFM . . . 138
7.6.4 Dix vins blancs de Touraine . . . 141
7.7 AFM dans FactoMineR (groupes) . . . 146

8 Groupes qualitatifs et mixtes **149**
8.1 ACM pondérée . . . 149
8.1.1 Nuage des modalités en ACM pondérée . . . 150
8.1.2 Relations de transition en ACM pondérée . . . 151
8.2 AFM de variables qualitatives . . . 151
8.2.1 Point de vue de l'analyse factorielle . . . 151
8.2.2 Point de vue de l'analyse multicanonique . . . 153
8.2.3 Représentation des individus partiels . . . 154
8.2.4 Représentation des modalités partielles . . . 155
8.2.5 Analyse dans l'espace des groupes de variables ($\mathbb{R}^{I^2}$) . . . 155
8.3 Cas des données mixtes . . . 157
8.3.1 Pondération des variables . . . 157
8.3.2 Propriétés . . . 158
8.4 Application . . . 160
8.4.1 Analyses séparées . . . 161
8.4.2 Inerties dans l'analyse globale . . . 162

8.4.3 Coordonnées des facteurs des analyses séparées 163
8.4.4 Premier facteur 164
8.4.5 Deuxième facteur 166
8.4.6 Troisième facteur 167
8.4.7 Représentation des groupes de variables 168
8.4.8 Conclusion 169
8.5 AFM de données mixtes dans FactoMineR 170

9 AFM et Statis **175**
9.1 Notations 175
9.2 Principes communs aux deux méthodes 176
9.3 Pondération des variables 176
9.3.1 Comparaison des deux méthodes 176
9.3.2 Illustration 177
9.4 Représentations superposées 180
9.4.1 Comparaison des deux méthodes 180
9.4.2 Illustration à l'aide des données 2^{6-3} 181
9.5 Mesure de liaison entre groupes de variables 183
9.5.1 Comparaison des deux méthodes 183
9.6 Représentation des groupes de variables 185
9.6.1 Comparaison des deux méthodes 185
9.6.2 Illustration à l'aide des données 2^{6-3} 186
9.7 Conclusion 189
9.8 Statis dans ade4 190

10 AFM et analyse procustéenne **193**
10.1 Analyse procustéenne 193
10.1.1 Données, notations 193
10.1.2 Objectifs 194
10.1.3 Méthodes et variantes 195
10.2 Comparaison entre les deux méthodes 196
10.2.1 Représentation des N_I^j 196
10.2.2 Nuage moyen 197
10.2.3 Objectif, critère, algorithme 198
10.2.4 Propriétés des représentations des N_I^j 199
10.2.5 Premier bilan 199
10.2.6 Harmonisation de l'inertie des N_I^j 200
10.2.7 Relations entre les facteurs homologues 200
10.2.8 Représentation des individus 201
10.2.9 Aides à l'interprétation 202
10.2.10 Représentation des variables 203
10.3 Etude d'un jeu de données choisies (2^{3-1}) 203
10.3.1 Données 2^{3-1} 203
10.3.2 Résultats de l'AFM 205

10.3.3 Résultats de l'APG 207
10.4 Application aux dix vins de Touraine 209
10.5 Conclusion 212
10.6 APG dans FactoMineR 212

11 Analyse factorielle multiple hiérarchique 215
11.1 Données, exemples 215
11.2 Hiérarchie et partitions 217
11.3 Pondération des variables 218
11.4 Représentation des individus partiels 219
11.4.1 Méthode 219
11.4.2 Application aux six jus d'orange 221
11.5 Coefficients de corrélation canoniques 223
11.6 Représentation des nœuds 223
11.7 Application à des données mixtes : le napping® catégorisé 225
11.7.1 Données et méthodologie 225
11.7.2 Analyse intermédiaire : AFM sur une nappe catégorisée . . 227
11.7.3 Décompositions de l'inertie 228
11.7.4 Représentations des individus, moyens et partiels 229
11.8 AFMH dans FactoMineR 234

A Calcul matriciel et espace euclidien 241
A.1 Fiche 1 : éléments de calcul matriciel 241
A.2 Fiche 2 : espace vectoriel euclidien 245
A.2.1 Espace vectoriel muni de la distance usuelle 245
A.2.2 Espace euclidien muni d'une métrique diagonale 247
A.2.3 Visualisation d'un nuage 248

Bibliographie 253

This page intentionally left blank

Chapitre 1

Analyse en composantes principales

L'analyse en composantes principales est la plus répandue des méthodes factorielles. Elle s'applique à un tableau dans lequel un ensemble d'individus (statistiques) est décrit par un ensemble de variables quantitatives. Le présent chapitre décrit de façon détaillée cette méthode, tant dans son principe que dans son application. C'est l'occasion d'introduire bon nombre de concepts qui seront utilisés lors de l'analyse de tableaux multiples, mais qui valent pour des tableaux simples. Cela permettra, dans la présentation de l'analyse factorielle multiple, de faire apparaître ses spécificités sans ambiguïtés.

1.1 Données, notations

On étudie un tableau ayant les caractéristiques décrites ci-après :
- chaque ligne représente un individu statistique ; on note I le nombre d'individus ; I désigne aussi l'ensemble des individus ; l'utilisation d'une même lettre, pour désigner un ensemble et son cardinal, n'est pas gênante car le contexte permet toujours de lever l'ambigüité ;
- chaque colonne représente une variable quantitative ; on note K le nombre de variables (ainsi que l'ensemble des variables) ;
- à l'intersection de la ligne i et de la colonne k, se trouve x_{ik}, valeur (numérique) de l'individu i pour la variable k.

Ajoutons deux notations classiques.

$\bar{x}_k$: moyenne de la variable k ; elle sera peu utilisée car les variables seront supposées centrées, mais il est quelquefois utile de faire apparaître explicitement le centrage ;

s_k : l'écart-type de la variable k.

Ces notations sont regroupées dans la figure 1.1.

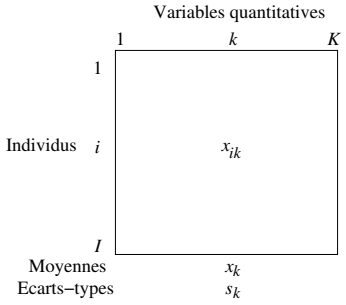

Fig. 1.1 – Structure des données et notations.

Les exemples de données susceptibles d'être analysées par ACP sont innombrables. Dans ce qui suit, nous utiliserons un exemple scolaire, riche et facile à comprendre. On dispose, pour 909 élèves de terminale scientifique ($I = 909$), de leurs notes au bac dans 5 matières ($K = 5$) : mathématiques, physique, sciences naturelles, histoire-géographie et philosophie.

1.2 Pourquoi analyser un tableau par ACP ?

Reprenons l'exemple précité. Après avoir examiné les moyennes, à un niveau très général, l'objet de l'étude statistique de ce tableau est d'étudier la diversité des élèves « intramatière » et « intermatières ». Cette diversité doit d'abord être examinée par matière, à l'aide d'indicateurs (principalement les écarts-types) et de graphiques (principalement boîtes à moustaches et histogrammes).
Le recours à l'analyse en composantes principales est motivé principalement par deux objectifs.

– On considère chaque élève non pas du point de vue de telle ou telle note particulière, mais de celui de l'ensemble de ses notes, ce que l'on appelle son « profil scolaire ». Cela conduit à étudier la diversité de ces profils (dans leur ensemble et non pas note par note). En ACP, cette diversité des profils est étudiée en mettant en évidence leurs principales dimensions de variabilité. Ainsi, dans l'exemple, on peut s'attendre à ce que la principale dimension de variabilité oppose les bons élèves (*i.e.* qui ont de bonnes notes dans toutes les matières) aux mauvais (*i.e.* qui ont de mauvaises notes dans toutes les matières).
– On s'intéresse aux liaisons entre les variables. En ACP, on ne considère que les liaisons linéaires ; l'intensité de ce type de liaison entre deux variables est mesurée, comme usuellement, par le coefficient de corrélation. En outre, ces liaisons sont étudiées à l'aide de variables synthétiques (dites composantes principales), combinaisons linéaires de variables initiales liées le plus possible (en un sens à préciser) à ces variables initiales. Idéalement, chaque variable synthétique est étroitement corrélée à un groupe de variables et non corrélée aux autres, mettant ainsi en évidence des groupes de variables (corrélées « intragroupe » et non

corrélées « intergroupes »).
Nous verrons que ces variables synthétiques coïncident (en un sens à préciser) avec les dimensions de variabilité du point de vue précédent. Ce qui (dé)montre que les deux objectifs évoqués sont étroitement liés, voire deux aspects d'un même questionnement. Cela peut être illustré dans le cadre de l'exemple : dire que la principale dimension de variabilité oppose les bons et les mauvais élèves (optique étude des individus *via* leur profil scolaire) est équivalent à dire que toutes les variables (*i.e.* les notes) sont corrélées positivement deux à deux (optique liaisons entre variables).
Cette idée peut, après tout, paraître évidente : les lignes, d'une part, et les colonnes, d'autre part, d'un tableau sont nécessairement deux faces d'une même réalité (*i.e.* le tableau lui-même). D'où le terme de dualité (*i.e.* caractère double) souvent utilisé pour désigner cette liaison entre les deux objectifs d'une part, et entre les résultats de l'ACP les concernant d'autre part. Elle n'en est pas moins fondamentale : elle aide à mieux comprendre ce que nous cherchons ; elle montre aussi l'adéquation de l'ACP avec une problématique très générale, à savoir analyser un tableau. Remarquons au passage que l'on retrouve cette dualité (des problématiques et des résultats) dans toutes les analyses factorielles (en particulier celles étudiées dans ce livre soit l'ACP, l'ACM, l'AFDM, l'AFM et l'AFMH), ce qui explique le caractère incontournable de la méthodologie factorielle dans l'analyse statistique d'un tableau.

1.3 Nuages des individus et des variables

Nuage N_I des individus

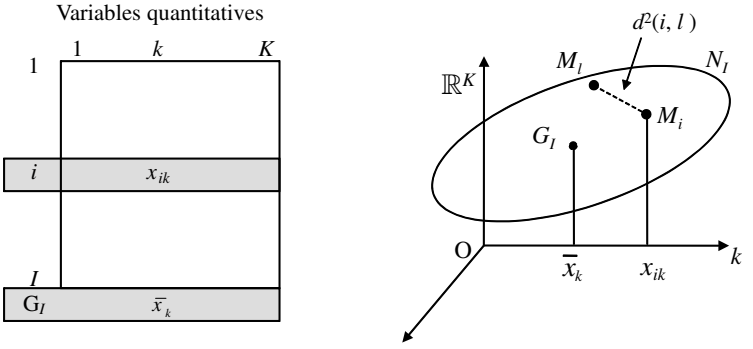

Fig. 1.2 – Le nuage des individus.

A l'individu i, on associe son profil soit $\{x_{ik}; k = 1, K\}$. A ce profil, correspond le point M_i dans l'espace $\mathbb{R}^K$ dont chaque dimension représente une variable (cf. figure 1.2). $\mathbb{R}^K$ est dit « espace des individus ». L'ensemble I des points i constitue un nuage noté N_I. En outre, à chaque individu est associé le poids p_i tel que

$\sum_i p_i = 1$ (généralement $p_i = \frac{1}{I}$).
Le centre de gravité du nuage N_I, noté G_I et dit aussi « point moyen », a pour coordonnées $\{\bar{x}_k; k = 1, K\}$. Lorsque les variables sont centrées, ce qui est toujours le cas en ACP, l'origine des axes dans $\mathbb{R}^K$ est placée en G_I (des compléments sur le centrage sont donnés en 1.4).
Dans le nuage N_I, le carré de la distance entre deux individus i et l s'écrit :

$$d^2(i,l) = \sum_k (x_{ik} - x_{lk})^2.$$

Cette quantité mesure la disparité entre les profils des individus i et l. Etudier la variabilité des individus revient à étudier ces distances, dont l'ensemble constitue la forme du nuage N_I. Cette variabilité peut aussi être appréhendée par les distances entre chaque point M_i et le point moyen G_I, soit, pour l'individu i :

$$d^2(i,G_I) = \sum_k (x_{ik} - \bar{x}_k)^2.$$

Cette distance mesure la particularité de l'individu i. L'ensemble de ces particularités individuelles constitue la variabilité globale des données. Pour mesurer cette variabilité globale, on agrège les carrés des distances au point moyen pour obtenir l'inertie totale de N_I (par rapport à G_I). Soit :

$$\text{Inertie totale de } N_I/G_I = \sum_i p_i d^2(i,G_I) = \sum_k \sum_i p_i (x_{ik} - \bar{x}_k)^2 = \sum_k Var[k].$$

Cette inertie totale est égale à la somme des K variances, notées $Var[k]$, soit, lorsque les variables sont réduites, au nombre de variables. Ce qui montre, de façon flagrante dans le cas centré réduit et aussi dans le cas général, que, en ACP, ce n'est pas l'inertie totale qui est intéressante mais la façon dont elle est répartie. On retrouvera cette propriété en ACM et en AFM. On obtient la même inertie totale en agrégeant les carrés des distances interindividuelles, point de vue adopté au début de cette section. La variance de la variable k en fonction des écarts entre individus s'écrit :

$$Var[k] = \frac{1}{2} \sum_i \sum_l p_i p_l (x_{ik} - x_{lk})^2.$$

On obtient, en combinant les deux équations précédentes :

$$\text{Inertie totale de } N_I/G_I = \frac{1}{2} \sum_{i,l} p_i p_l \sum_k (x_{ik} - x_{lk})^2 = \frac{1}{2} \sum_{i,l} p_i p_l d^2(i,l),$$

ce qui montre que l'inertie de N_I représente la variabilité des individus à la fois du point de vue de leur écart au centre de gravité et du point de vue des distances interindividuelles.

Nuage N_K des variables

A la variable k, on associe ses valeurs pour l'ensemble des individus étudiés soit : $\{x_{ik}; i = 1, I\}$. Cet ensemble correspond au point M_k (et au vecteur v_k) de l'espace $\mathbb{R}^I$ dont chaque dimension correspond à un individu. $\mathbb{R}^I$ est dit « espace des variables » ou, plus généralement, « espace des fonctions sur I » (une fonction sur I associe une valeur numérique à chaque individu i). L'ensemble des points M_k constitue le nuage des variables noté N_K (figure 1.3).

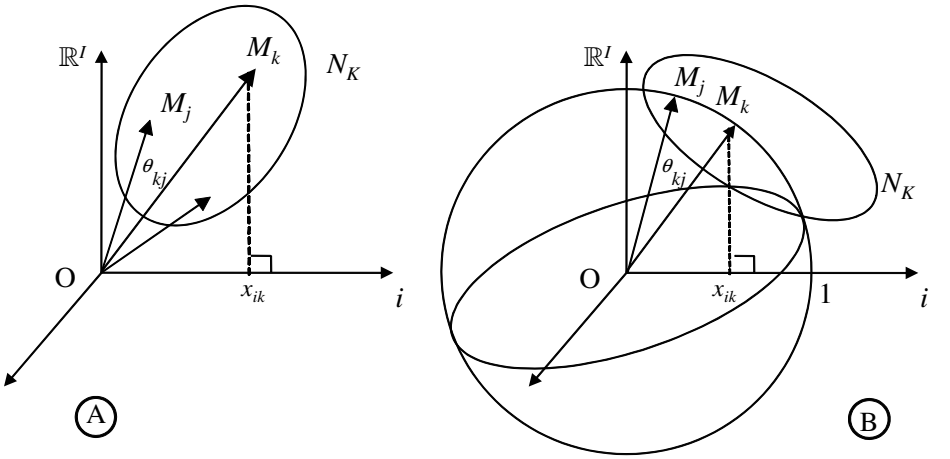

Fig. 1.3 – Le nuage des variables. A : données centrées ; B : données centrées et réduites. θ_{kj} est l'angle formé par les deux vecteurs représentant les variables k et j ($\overrightarrow{OM_k}$ et $\overrightarrow{OM_j}$).

Lorsque les variables sont centrées, c'est-à-dire toujours en ACP, cet espace possède deux propriétés remarquables :

– le cosinus de l'angle θ_{kj} formé par les deux variables k et j est égal à leur coefficient de corrélation. Cette interprétation géométrique du coefficient de corrélation justifie l'intérêt de cet espace dans l'étude des liaisons entre variables. Elle explique aussi que l'on y représente la variable k par le vecteur reliant l'origine au point M_k ;
– la distance entre M_k et O est égale à la variance de la variable k. Interpréter une variance comme un carré de longueur est très précieux en statistique. Cas particulier important : une variable centrée réduite a pour longueur 1 ; le nuage N_K est alors situé sur une hypersphère (de rayon 1).

Pour obtenir ces deux propriétés, il est nécessaire, dans le calcul d'une distance dans $\mathbb{R}^I$, d'accorder à chaque dimension i le poids p_i de l'individu correspondant. Ainsi, on a bien :

$$d^2(O, M_k) = \sum_i p_i \, (x_{ik} - \bar{x}_k)^2 = Var[k].$$

Le tableau 11.6 page 231 rassemble des résultats issus de l'AFMH et des AFM séparées des nœuds reliés au nœud sommital. Le tableau 11.7 page 233 nécessite des calculs dans $\mathbb{R}^{I^2}$. Nous donnons ci-après le code correspondant à ces deux tableaux. Le lecteur peut essayer de le retrouver à titre d'exercice :

```
# Tableau 11.6
# Initialisation
> Tab11_6=matrix(nrow=5,ncol=3)
# Noms des lignes et des colonnes
> row.names(Tab11_6)=c("AFM nappe 1","AFM nappe 2","AFMH nuage moyen"
+ ,"AFMH nuage partiel nappe 1","AFMH nuage partiel nappe 2")
> colnames(Tab11_6)=c("F1","F2","F1/F2")

# Valeurs propres des AFM séparées des nappes catégorisées

> Tab11_6[1,1:2]=resafmnappe1$eig[1:2,1]
> Tab11_6[2,1:2]=resafmnappe2$eig[1:2,1]

# AFMH. Valeurs propres puis variances des nuages partiels
# par dimension de l'AFMH
> Tab11_6[3,1:2]=ResAFMH$eig[1:2,1]
> Tab11_6[4,1:2]=apply(ResAFMH$partial[[2]][,1:2,1],MARGIN=2,
+ FUN=var)*5/6
> Tab11_6[5,1:2]=apply(ResAFMH$partial[[2]][,1:2,2],MARGIN=2,
+ FUN=var)*5/6

> for(i in 1:5){Tab11_6[i,3]=Tab11_6[i,1]/Tab11_6[i,2]}
> round(Tab11_6,3)

# Tableau 11.7
# Initialisation
> Tab11_7=matrix(nrow=6,ncol=4)
# Noms des lignes et des colonnes
> row.names(Tab11_7)=c("Nappe cat. 1","Nappe cat. 2","Nappe ss 1",
+ "Catégorisation 1","Nappe ss 2","Catégorisation 2")
> colnames(Tab11_7)=c("Ng","F1","F2","Plan(1,2)")

# Les normes des groupes avant l'ultime pondération de l'AFMH (Ng)
# sont dans les AFM séparées des nappes catégorisées
> Tab11_7[1,1]=sum(resafmnappe1$eig[,1]^2)/resafmnappe1$eig[1,1]^2
> Tab11_7[2,1]=sum(resafmnappe2$eig[,1]^2)/resafmnappe2$eig[1,1]^2
> Tab11_7[3:4,1]=diag(resafmnappe1$group$Lg)[1:2]
> Tab11_7[5:6,1]=diag(resafmnappe2$group$Lg)[1:2]
```

```
# Cos carré des groupes : carré de longueur projetée (in AFMH)
# sur carré de longueur totale (Ng)
> for(i in 1:2){Tab11_7[1:2,i+1]=ResAFMH$group$coord[[2]][,i]^2/
+ Tab11_7[1:2,1]}
> for(i in 1:2){Tab11_7[3:6,i+1]=ResAFMH$group$coord[[1]][,i]^2/
+ Tab11_7[3:6,1]}
> Tab11_7[,4]=apply(Tab11_7[,2:3],MARGIN=1,FUN=sum)
> round(Tab11_7,3)
```

Nous rassemblons ci-après les lignes de code correspondant à l'AFMH appliquée aux jus d'orange :

```
# Lecture des données et sélection des colonnes utiles ici
# dans le data-frame Orange
> orange5=read.csv2("orange5.csv",header=T,row.names=1)
> orange=orange5[,c(3:17,19:114)]

> library(FactoMineR)

# AFMH
> resAFMH=HMFA(orange,type=c("s","s","s"),H=list(c(8,7,96),c(2,1)),
+ name.group=list(c("Chimie","Sensoriel","Hédonique"),
+ c("Caractérisation","Hédonique")))

# Figure 11.3
> plot.HMFA(resAFMH,choix="ind",invisible="quali",new.plot=TRUE,
+ cex=1.4)
# Cette commande génère 3 graphiques dont celui de la figure 11.3.
# On ferme les autres ce dernier devenant alors actif.
> text(resAFMH$partial[[2]][,1:2,1],labels=rep("c",6),pos=3,
+ offset=.5,cex=1)
> text(resAFMH$partial[[2]][,1:2,2],labels=rep("h",6),pos=3,
+ offset=.5,cex=1)

# Figure 11.4
> plot.HMFA(resAFMH,choix="ind",invisible="quali",new.plot=TRUE,cex=1.4)
# Cette commande génère 3 graphiques dont celui de la figure 11.4.
# On ferme les autres ce dernier devenant alors actif.
> text(resAFMH$partial[[2]][,1:2,1],labels=rep("c",6),pos=3,
+ offset=.5,cex=1)
> text(resAFMH$partial[[2]][,1:2,2],labels=rep("h",6),pos=3,
+ offset=.5,cex=1)
> text(resAFMH$partial[[1]][,1:2,1],labels=rep("ch",6),pos=3,
+ offset=.5,cex=1)
```