

Analyse de données

Denis CLOT

2020-2021

Table des matières simplifiée

Notations	1
I Analyses factorielles dans le cadre multivarié	2
1 Préparation des données	3
2 L'analyse en composantes principales	8
3 L'analyse factorielle des correspondances	36
4 L'analyse factorielle des correspondances multiples	51
5 L'analyse factorielle des données mixtes	61
II Analyses factorielles dans le cadre fonctionnel	69
6 Introduction au cadre fonctionnel	70
7 Transport de l'ACP dans le cadre fonctionnel	75
8 Solution dans le cas monofonctionnel	78
Appendices	86
A Preuves de la partie ACP multivariée dans le cadre euclidien simplifié	86
B ACP dans le cadre euclidien général	90
C Suite des preuves et données de l'analyse Budget/Temps	93
D Algorithme NIPALS	96
E Rappel sur le χ^2	98
F Table de valeurs du χ^2	104
G Prolongements de l'annexe E	106
Table des matières détaillée	109

Table des figures

1.1	Graphiques de synthèse pour variables quantitatives	3
1.2	Graphiques de synthèse pour variables qualitatives	4
1.3	Graphique pour un couple de variables quanti/quali	5
1.4	Représentation 3D	6
1.5	Liaison entre lumière nocture et myopie	7
2.1	Points et ellipse d'inertie	8
2.2	Axes de projections et projections	9
2.3	Points de \mathbb{R}^3	10
2.4	Axes d'inertie de l'ellipsoïde	10
2.5	Projections dans le référentiel de la longueur et la largeur de l'ellipse	11
2.6	Représentations des individus	12
2.7	Représentation physique de la première formulation	13
2.8	Représentation physique de la seconde formulation	14
2.9	Valeurs propres de V et critères	26
2.10	Projections des nuages de variables et d'individus dans le plan porté par u_1 et u_2	28
2.11	Projections des nuages de variables et d'individus dans le plan porté par u_3 et u_4	28
2.12	Projections des nuages de variables exploitant les COS2	30
3.1	Représentation en mosaique	47
3.2	Valeurs propres et χ^2	48
3.3	Projection des modalités	49
3.4	Projections et indicateurs de qualité	50
4.1	Choix de la dimension	57
4.2	Carré des liaisons et Projections des modalités de variables	58
4.3	Projections des modalités de variables avec ajout des relations d'ordre et projection des individus	59
5.1	Comment partitionner la distribution ?	61
5.2	Choix de la dimension	65
5.3	Analyse de l'inertie des axes et projections des individus	66
5.4	Projections des modalités et des variables	66
5.5	Analyse de l'inertie des axes 3/4 et projections des individus	67
5.6	Projections des modalités de variables et ajout des relations d'ordre	68
6.1	Données simples et données fonctionnelles	70
6.2	Exemples de méthodes simples de reconstruction de fonctions	71
6.3	Exemples de lissage	72
6.4	Exemple de courbes asynchrones	73
6.5	Recalage de courbes asynchrones	73

8.1	Schema d'intégration basé sur les trapèzes	80
8.2	Courbes et valeurs propres	81
8.3	Premier et second modes de variation	82
8.4	Projections des individus dans le plan porté par u_1 et u_2	82
8.5	Projections des stations de Thunderb et Uraniumc	83
8.6	Projections des stations de Halifax et Yarmouth	83
F.1	Illustration de probabilité de dépasser la valeur du χ^2	104

Liste des tableaux

2.1	corrélations pour chaque couple de variables	11
2.2	Indicateurs tirés des valeurs propres	25
2.3	Indicateurs de contribution et de qualité des représentations des individus	34
2.4	Indicateurs de contribution et de qualité des représentations des individus	35
3.1	Genres musicaux présents dans les playlist des étudiants d'une classe croisés avec le sexe des élèves	36
3.3	Issue à la loterie selon objet fétiche	37
3.4	Les valeurs de V et W mesurées sur une population - forme brute et tableau de contingence	37
3.5	Un tableau de contingence et table de fréquences associée	37
3.6	Cas de dépendance maximale	44
4.1	Indicateurs tirés des valeurs propres	58
5.1	Premières lignes du jeu de données	64
5.2	Décomposition de l'inertie des axes	65

Notations

Notations ensemblistes

\mathbb{R}	ensemble des réels
$\llbracket 1, p \rrbracket$	ensemble des entiers compris entre 1 et p

Notations matricielles

X'	transposée de la matrice/du vecteur X
x_i^j	élément de la i^e ligne et de la j^e colonne de la matrice X
X^{-1}	inverse de la matrice X
$Tr(X)$	trace de la matrice X
$\det(X)$	déterminant de la matrice X
$\ u\ $	la norme euclidienne du vecteur u
$\ u\ _M$	la M -norme de u
$E \oplus F$	somme directe des espaces E et F

Notations statistiques

$\text{var}(X)$	variance de la variable X
$\text{cov}(X, Y)$	covariance des variables X et Y
$\text{corr}(X, Y)$	corrélation des variables X et Y
σ_X	écart-type de X

Notations fonctionnelles

$L_2(\Omega)$	espaces des fonctions réelles de carré intégrable sur le domaine Ω
$\mathcal{C}(\Omega)$	espaces des fonctions réelles continues sur le domaine Ω
$\mathcal{C}^p(\Omega)$	espaces des fonctions réelles continues sur le domaine Ω de dérivée p -ième continue sur Ω

Notations diverses

δ_{ij}	symbole de Kronecker
$\lfloor x \rfloor$	plus grand entier inférieur ou égal à x
$\lceil x \rceil$	plus petit entier supérieur ou égal à x
C^{te}	constante réelle
$\sum_i f(i)$	sommation sur l'ensemble des valeurs de l'indice i des quantités $f(i)$
$\prod_i f(i)$	produit sur l'ensemble des valeurs de l'indice i des quantités $f(i)$
$x \perp y$	x et y orthogonaux selon le produit scalaire adopté
$\widehat{x^H}$	projection orthogonale sur l'espace H
$\min_{x:\dots} A$	valeur minimum de A par rapport à x qui satisfait les contraintes \dots
$\max_{x:\dots} A$	valeur maximum de A par rapport à x qui satisfait les contraintes \dots

Première partie

Analyses factorielles dans le cadre multivarié

Chapitre 1

Préparation des données

1.1 Contrôle et synthèse des données

En amont du travail d'analyse réalisable avec des méthodes d'analyse de données, il est important de se faire une idée des données sur lesquelles les analyses vont porter. En pratique surviennent souvent des décalages entre les données attendues et les données à disposition dont la nature peut être ambiguë (qualitatif ayant l'apparence de quantitatif) et la qualité défaillante (données mal encodées, données manquantes). Aussi, il est important de se doter de moyens de **synthèse** jouant également un rôle d'outils de **contrôle** afin d'appréhender rapidement la vraie nature des données et de déceler le plus rapidement tout problème.

Les outils classiques de la **statistique univariée** apportent des informations pertinentes dans cette perspective et peuvent également aider dans le processus de sélection des données. Ci-après, nous distinguons le cas des variables quantitatives de celui des variables qualitatives.

1.1.1 Variables quantitatives

Les résumés classiques pour une variable quantitative sont :

- la moyenne ;
- la variance - ou sa racine carrée, l'écart-type ;
- la médiane ;
- les quartiles ;
- la valeur minimale ;
- la valeur maximale.

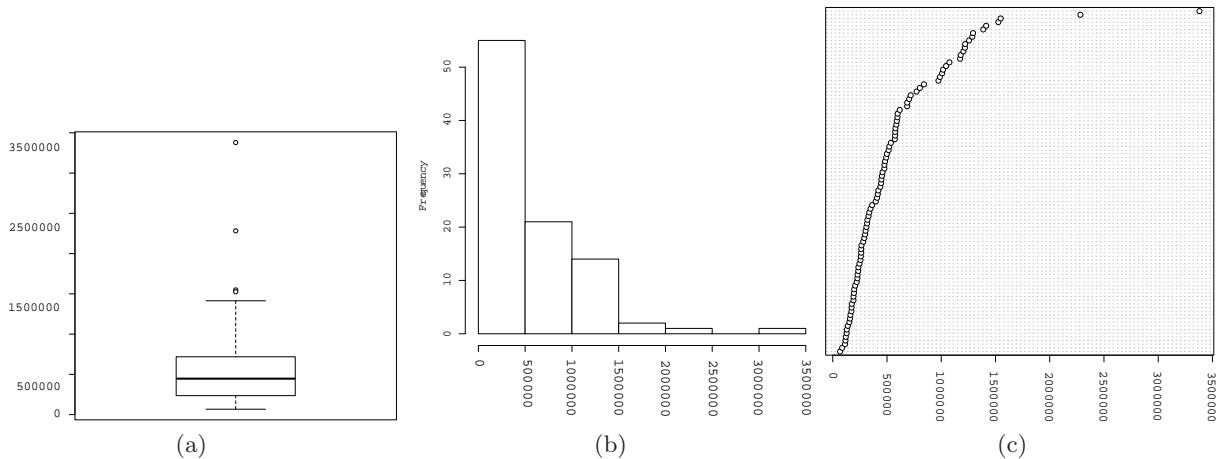


FIGURE 1.1 – Graphiques de synthèse pour variables quantitatives

Une partie de ces informations peut être représentée sur un graphique appelé boîte à moustaches (cf figure 1.1a). Un autre type de graphique classique est l'histogramme, permettant de représenter la répartition des valeurs (cf figure 1.1b). Une version moins synthétique de cette répartition est donnée par le graphe de Cleveland (cf figure 1.1c).

1.1.2 Variables qualitatives

Une variable qualitative met en relation une population avec un ensemble de modalités. Ces modalités peuvent être ordonnées ou pas. Il est classique de compter le nombre d'individus associés à une modalité. Ce comptage brut correspond aux fréquences absolues. Il est parfois ramené à l'effectif total de la population (ou de l'échantillon) et on parle alors de fréquences relatives.

Il est fréquent d'avoir recours à des graphiques en batons pour représenter ces fréquences. Dans le cas où les modalités seraient ordonnées, l'usage veut que les batons soient positionnés selon l'ordre des modalités (cf figure 1.2a). Dans le cas non ordonné, il est possible de représenter des fréquences relatives sous forme de parts de camembert (*piechart*) (cf figure 1.2b).

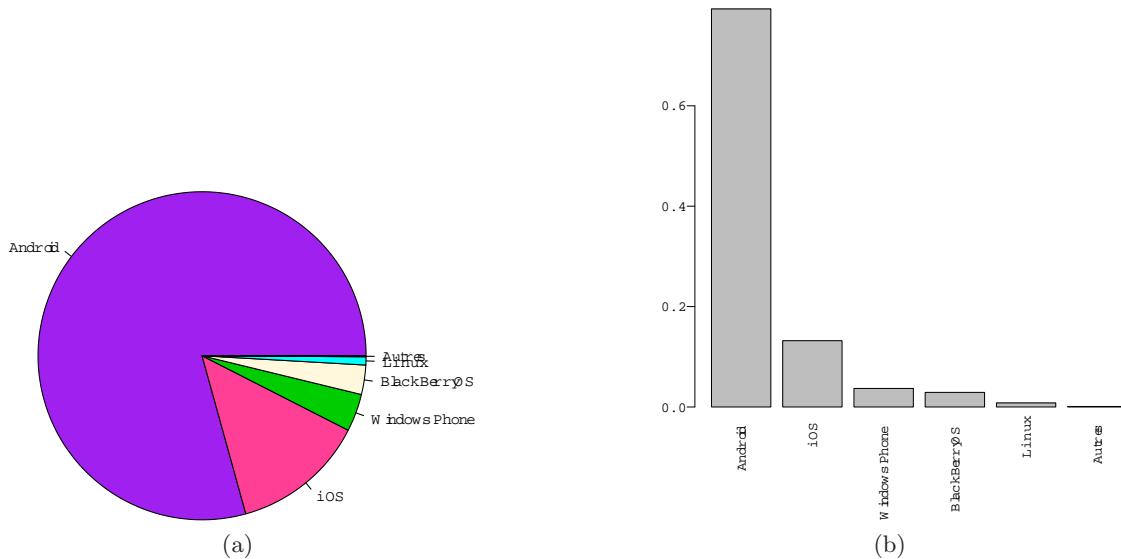


FIGURE 1.2 – Graphiques de synthèse pour variables qualitatives

1.2 Recherche de relations

Ces premiers éléments de synthèse peuvent être prolongés par des éléments d'**analyse descriptive bivariée**, i.e. portant sur des couples de variables. Les résultats produits pourront permettre l'observation de relations que les méthodes d'analyse de données permettront de poursuivre sur un nombre plus important de variables.

1.2.1 Relation entre deux variables quantitatives

Plusieurs quantités peuvent être calculées pour identifier une relation linéaire entre deux variables quantitatives X et Y :

- la **covariance** : $\text{cov}(X, Y)$,
- la **corrélation** : $\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \times \sigma_Y}$
- le **coefficient de détermination** : $\text{corr}(X, Y)^2$

Ces grandeurs devraient toujours être confrontées à une représentation graphique des données !

1.2.2 Relation entre deux variables qualitatives

Dans le cas de deux variables qualitatives, Q_1 et Q_2 , chacune dispose de modalités et la question est de savoir s'il existe un lien entre leurs modalités, c-a-d si l'on peut tirer une information sur Q_2 connaissant Q_1 , le cas le plus fort étant de pouvoir déduire la modalité de Q_2 connaissant celle de Q_1 pour un grand nombre d'individus par exemple.

Si le nombre de modalités pour chacune des variables n'est pas trop important, il est possible d'observer le tableau de contingence qui donne les effectifs pour les paires de modalités issues du croisement des deux variables. Diverses représentations graphiques découlant de ce tableau sont possibles pour faire ressortir les modalités qui se rencontrent souvent et celles qui tendent à s'éviter.

Supposons que Q_1 (resp. Q_2) ait q_1 (resp. q_2) modalités différentes et que leurs relations soient étudiées sur n individus. Pour étudier de tels liens, deux grandeurs peuvent être calculées à partir de la table de contingence croisant les deux variables :

- le χ^2 dit de **contingence** qui mesure l'écart à la situation d'indépendance,
- le **coefficient de Cramer**, basé sur le χ^2 et qui varie entre 0 et 1 :

$$\sqrt{\frac{\chi^2}{n \times \min(q_1 - 1, q_2 - 1)}}$$

1.2.3 Relation entre une variable quantitative et une variable qualitative

Une façon de s'intéresser au lien entre les valeurs d'une variable quantitative, X et les modalités d'une variable nominale, Q , est de comparer les valeurs prises par les individus associés à une même modalité. Il est très facile d'inspecter cela graphiquement lorsque le nombre de modalités de Q est petit.

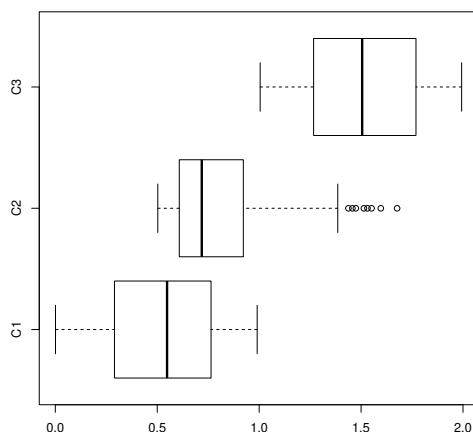


FIGURE 1.3 – Graphique pour la recherche de lien entre une variable qualitative et une variable quantitative

C'est ce que permet le **rappor de corrélation** qui compare les variances de X par sous-groupes d'individus (ceux associés à une modalité de Q) à la variance totale. défini comme le rapport de l'inertie inter-groupe ramené à l'inertie totale :

$$\eta^2 = \frac{\sum_k n_k (\bar{x}_k - \bar{X})^2}{\sum(X - \bar{X})^2} \quad \left(= \frac{\sum_k \frac{n_k}{n} (\bar{x}_k - \bar{X})^2}{\frac{1}{n} \sum(X - \bar{X})^2} = \frac{\sum_k p_k (\bar{x}_k - \bar{X})^2}{\sum_i p_i (x_i - \bar{X})^2} \right)$$

où n_k est le nombre d'individus associés à la k^e modalité de Q et \bar{x}_k la moyenne de X pour ce sous-groupe d'individus. Entre parenthèses, l'expression analogue faisant intervenir un système de poids

individuel (p_i est le poids du i^{e} individu et p_k est le poids total des individus associés à la modalité k^{e} modalité).

Nous verrons plus tard qu'il s'agit de l'inertie inter-groupes ramenée à l'inertie totale.

1.2.4 Au delà des analyses bivariées...

Considérons un groupe d'une trentaine d'individus pour lesquels nous disposons pour chacun des informations suivantes :

- durée consacrée à des activités ménagères ;
- durée consacrée à des activités professionnelles ;
- durée consacrée aux loisirs.

Nous avons la possibilité, moyennant quelques efforts, d'utiliser des outils de représentation en 3D. Avec un peu de pratique, il est possible de repérer des relations. Sur la figure 1.4, une corrélation semble se révéler entre les activités ménagères et professionnelles.

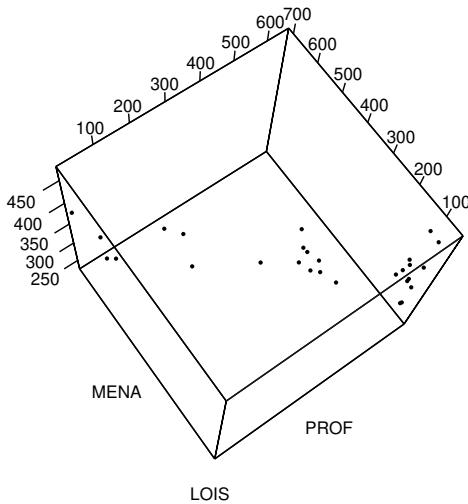


FIGURE 1.4 – Représentation en 3D du nuage de points tiré des trois variables

Ceci étant, que faire pour s'en sortir si une nouvelle variable est ajoutée ? Si une vingtaine de variables sont disponibles ? La possibilité d'examiner toutes les paires - ou tous les triplets - de variables n'étant pas raisonnable, la recherche de relation basée sur l'examen de ce type de représentation de l'information atteint les limites de son efficacité. Et que faire si des variables qualitatives s'ajoutent à cette information déjà abondante ? Les **méthodes d'analyse factorielle** permettent d'aller au delà des limites évoquées et apportent des réponses à ces différentes questions avec de nouveaux résumés d'information.

1.3 Etablir une causalité

Même si une liaison est identifiée à l'aide d'un coefficient de corrélation ou d'un χ^2 , il faut **être extrêmement prudent** dans l'interprétation de cette liaison statistique : elle peut indiquer une relation de causalité...ou pas :

- Pas dans le cas où notre indicateur statistique ne relève que d'une proximité fortuite entre les mesures portant sur deux phénomènes : Tyler Vigen s'est amusé à relever des corrélations liant par exemple le cours de l'or et la vente de disques vinyls, la production de chicorée et les maladies infectieuses ou encore le nombre de piétons tués par collision avec un train et la

pluviométrie d'un comté de l'état du Missouri aux USA¹. Un lien entre ces phénomènes semble relever de l'absurdité et toute hypothèse de causalité est directement rejetée.

- Pas lorsqu'une variable cachée est la cause réelle du lien statistique fort mesuré sur deux phénomènes. Le problème d'une variable cachée est précisément qu'il est difficile de se douter de son rôle dans l'apparente relation entre deux phénomènes. Des recherches² ont porté sur les effets que pourrait avoir l'exposition à des sources lumineuses nocturnes pendant le sommeil de très jeunes enfants. Il est apparu que la fréquence de la myopie était plus importante chez les enfants exposés à des veilleuses mais ces résultats ont été accompagnés de prudence sur l'hypothèse d'une causalité. Une seconde équipe de chercheurs a relevé³ un biais dans la constitution de l'échantillon, une sur-représentation de familles comportant des parents myopes, or les parents myopes ont plus tendance à recourir à des veilleuses la nuit et la myopie est une maladie héréditaire. L'augmentation involontaire de la part de parents myopes aurait conduit à l'augmentation de l'utilisation de veilleuses et à l'augmentation d'enfants concernés par la myopie. La liaison statistique ne révèle donc pas ici une causalité entre l'exposition de très jeunes enfants aux veilleuses et l'apparition de myopie.

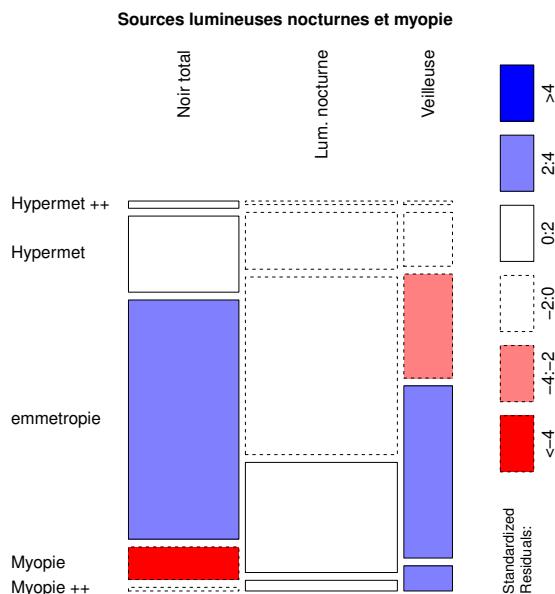


FIGURE 1.5 – Liaison entre exposition à la lumière nocturne de jeunes enfants et développement de myopie

- Il y a bien entendu de nombreux cas pour lesquels c'est bien une relation de causalité qui explique les relations entre les mesures portant sur deux phénomènes. Cependant, il n'est pas toujours facile de déterminer la variable qui agit sur l'autre...

1. Consulter <http://tylervigen.com/spurious-correlations> pour plus d'exemples.
 2. Voir <https://doi.org/10.1038/20094>
 3. Voir <https://doi.org/10.1038/35004663>

Chapitre 2

L'analyse en composantes principales

2.1 L'ACP, outil d'exploration et de synthèse

L'analyse en composantes principales est une méthode d'analyse exploratoire qui vise à sonder la **structure** d'une population observée par l'intermédiaire de différentes variables et à donner une **synthèse** globale de l'information portée par les données. Plus précisément :

- Les structures peuvent être cherchées sur les individus de la population : existe-t-il des groupes d'individus et, le cas échéant, pourrons-nous les caractériser ? Le plus simple est de les observer, mais dans le cas où il s'agit de points d'un espace de dimension arbitraire (supérieure à 2 ou 3), il faut construire des représentations restituant au mieux les positions relatives des individus dans leur espace de départ. Notre problème est de construire des synthèses dans des espaces de petite dimension (1 ou 2) qui restituent avec le moins de perte d'information ce qui se passe dans l'espace de départ, bien plus grand. Deux exemples illustreront cela après le point suivant.
- L'analyse des relations entre les variables mesurées sur la population peut également être le but d'une ACP. Cette analyse peut permettre d'identifier de la redondance et de construire des variables synthétiques permettant de résumer l'information originale. L'analyse budget/temps ci-dessous illustre cela.

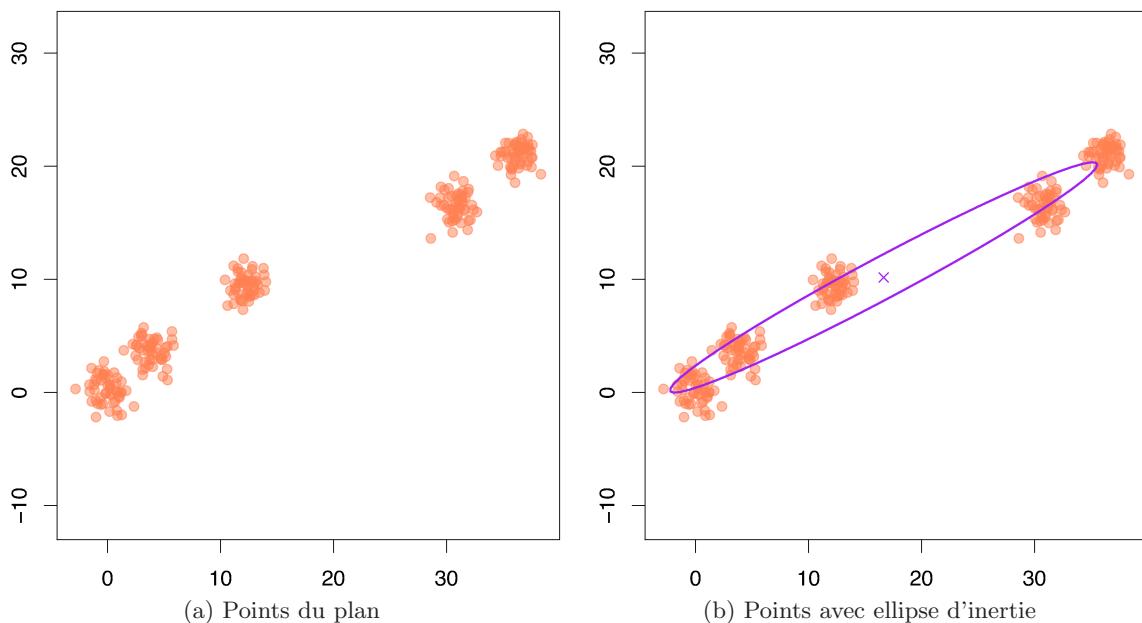


FIGURE 2.1 – Points et ellipse d'inertie marquant la direction de plus forte inertie

- **Points en 2D résumés en 1D**

Considérons ici les points représentés sur la figure 2.1a. Nous souhaitons les résumer par leurs projections sur une droite bien choisie et les résumés devront restituer au mieux l'information sur les points.

Pour choisir cette droite, nous pouvons nous aider de l'ellipse tracée sur la figure 2.1b. Il s'agit de l'ellipse d'inertie dont le grand axe coïncide avec la direction de plus grande inertie. Cet axe est la droite idéale cherchée. C'est en projetant les points orthogonalement sur elle que la dispersion des points est la mieux restituée.

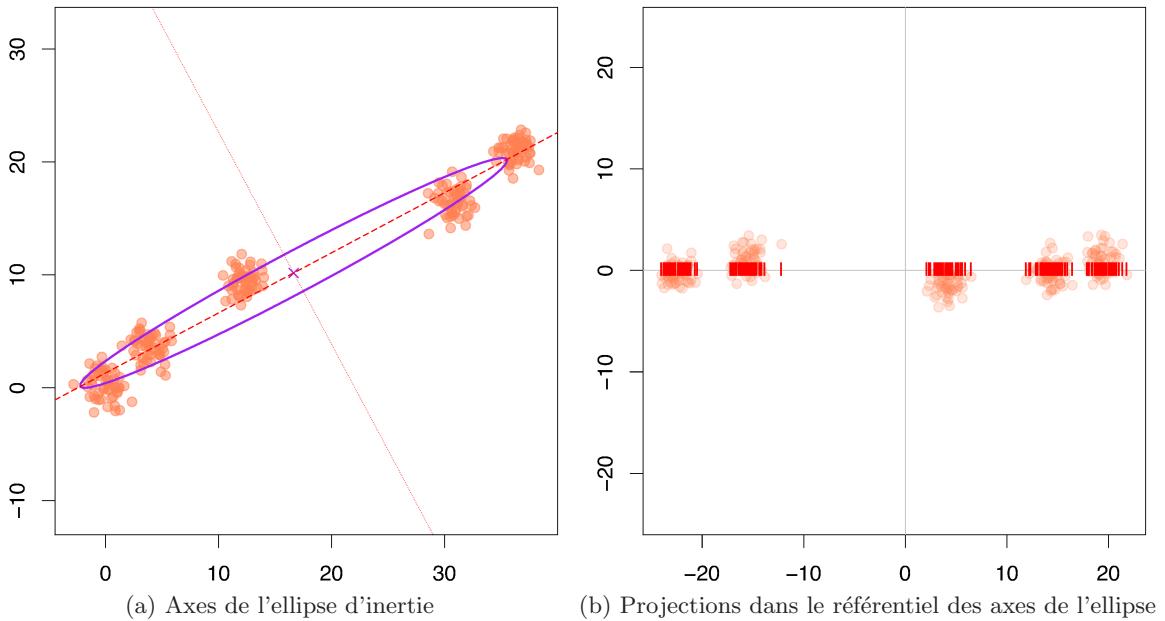


FIGURE 2.2 – Axes de projections et projections sur l'axe de plus forte inertie

Cet exemple illustre le fait qu'il est possible de restituer une grande partie de l'information fournie en 2D par une synthèse construite dans un espace à une dimension. Sur cette droite, la présence des groupes est claire. En revanche, si nous avions choisi l'autre axe pour construire cette synthèse, la présence de ces groupes n'aurait plus été révélée.

- **Points en 3D résumés en 2D**

Considérons à présent des points illustrés en 2.3 et cherchons le meilleur plan pour les représenter. Comme dans l'exemple précédent, nous pouvons avoir recours aux axes de l'ellipsoïde d'inertie pour rechercher l'axe maximisant la dispersion des points.

Les figures de 2.4 montrent qu'en plus de cette longueur d'ellipsoïde, il est possible d'exploiter sa largeur (mise en évidence sur 2.4b) et son épaisseur (voir 2.4c). C'est le plan porté par la longueur et la largeur de l'ellipsoïde qui sera conservé pour former le plan de projection.

À nouveau, ces projections restituent bien l'information initiale : les différents amas se distinguent très bien et leurs positions relatives ne semblent pas beaucoup déformées par la projection. À titre de comparaison, si le plan retenu avait été porté par la largeur et l'épaisseur, trois masses auraient été restituées, dont une correspondant à la fusion de quatre groupes, ce qui aurait constitué une grande perte d'information.

Ces deux exemples illustrent qu'il est possible de construire des **synthèses** graphiques qui permettent d'observer ce qui se passe dans un espace de dimension supérieure en limitant la perte d'information. Evidemment, ces exemples correspondent à ce que nous aimerais toujours pouvoir construire, mais il est des cas pour lesquels les pertes d'information ne sont pas évitables. Cela tient aux relations

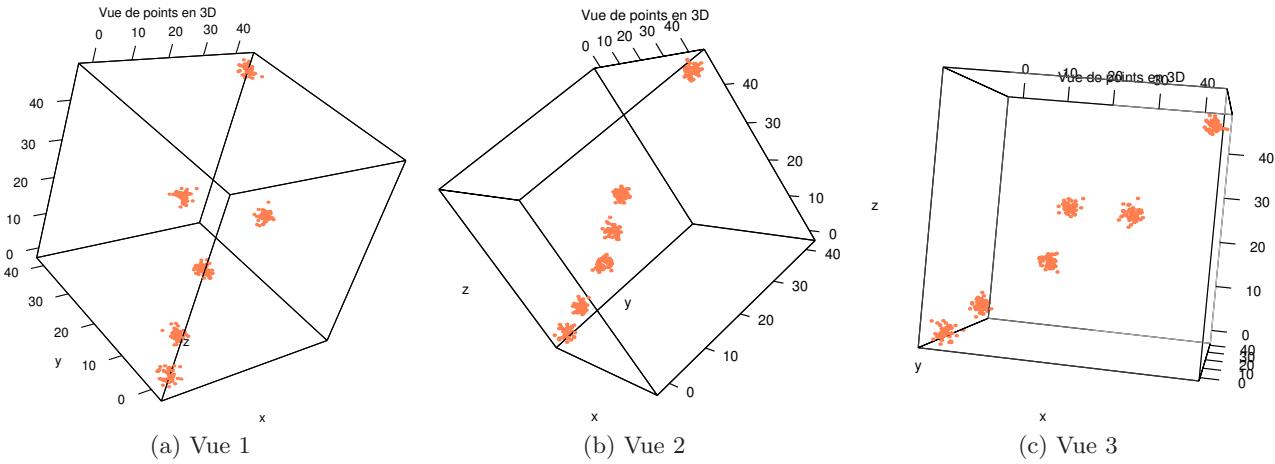


FIGURE 2.3 – Vues d’amas de points dans \mathbb{R}^3

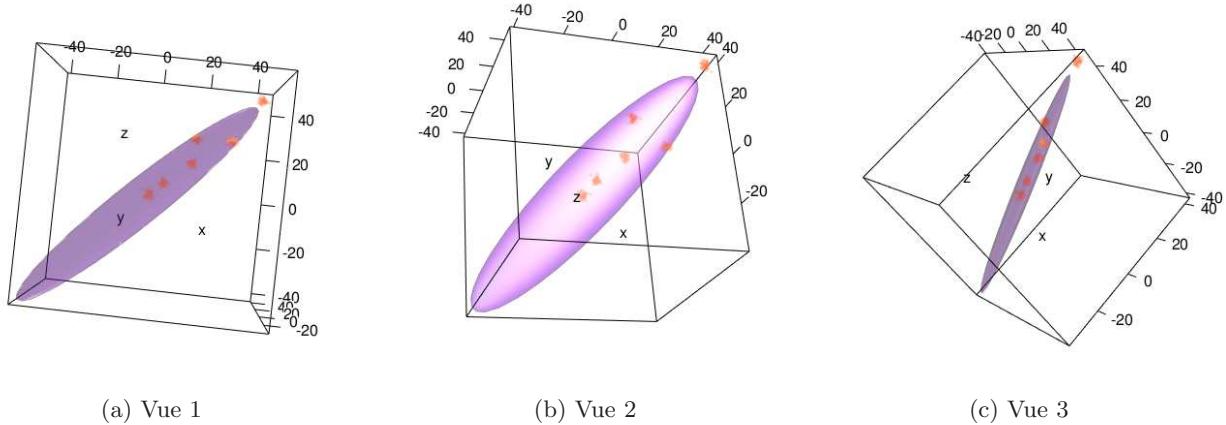


FIGURE 2.4 – Axes d’inertie de l’ellipsoïde

existant entre les variables de départ...

Ils soulignent également le caractère descriptif de l’ACP : il ne faut pas perdre de vue que quel que soit le contexte d’usage d’une ACP, ses résultats sont par nature purement descriptifs dans le sens où ils ne peuvent servir de base solide à une quelconque théorie ou à la validation d’une hypothèse.

• Budgets-temps

Nous considérons ici les données de l’analyse des budgets-temps de Jambu ([JAM76]). Il s’agit du temps consacré par des individus dans une journée à des activités génériques : temps passé dans les transports, au travail, à dormir, à manger... Il y a au total 10 variables renseignées pour une petite trentaine d’individus (les données seront considérées sous leur forme centrée et réduite) et nous savons qu’il y a de la redondance dans ces variables. Nous pouvons observer cela grâce à la matrice de corrélation suivante :

Les quatres premières variables sont assez fortement corrélées entre elles (positivement ou négativement). La dernière variable semble peu liée à toutes les autres. Considérons la variable synthétique construite comme combinaison linéaire des variables centrées (et réduites) à partir des coefficients

$$(-0.46, -0.46, 0.42, 0.40, 0.26, 0.04, 0.27, 0.31, 0.05, 0.08)$$

Les coefficients utilisés correspondent à un certain vecteur propre de la matrice de corrélation.

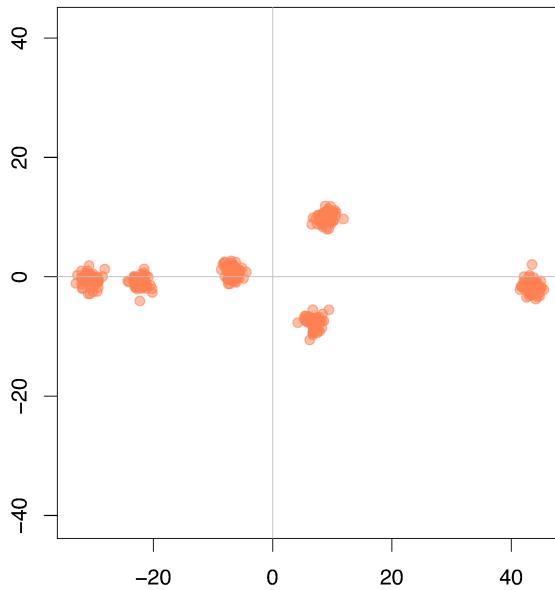


FIGURE 2.5 – Projections dans le référentiel de la longueur et la largeur de l’ellipse

	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS
PROF	1.00	0.94	-0.91	-0.86	-0.65	-0.11	-0.46	-0.56	-0.06	-0.25
TRAN		1.00	-0.87	-0.81	-0.50	-0.08	-0.61	-0.70	-0.04	-0.16
MENA			1.00	0.86	0.50	-0.04	0.36	0.44	-0.21	-0.05
ENFA				1.00	0.54	0.12	0.36	0.28	0.12	-0.06
COUR					1.00	0.59	-0.18	-0.02	0.22	0.24
TOIL						1.00	-0.35	-0.21	0.32	0.06
REPA							1.00	0.82	0.32	0.06
SOMM								1.00	0.02	0.27
TELE									1.00	-0.07
LOIS										1.00

TABLE 2.1 – corrélations pour chaque couple de variables

Soulignons que ce vecteur est normé. Le calcul, pour chaque variable, de sa projection orthogonale le long de cette variable montre que les quatre premières variables ont un cosinus, en valeur absolue, très proche de 1, ce qui signifie que cette variable peut être utilisée pour approcher ces variables. En d’autres termes, la valeur de la variable synthétique peut être utilisée pour reconstruire - avec une précision variable - la valeur de chacune de quatre variables¹. De plus, pour chacune des quatre premières variables, son approximation par sa projection orthogonale le long de l’axe porté par cette variable synthétique permet de restituer plus de 95% de sa variance pour les deux premières et plus de 74% pour les deux suivantes.

Nous pourrions mettre en avant d’autres vecteurs permettant de construire d’autres variables synthétiques, les variables ainsi formées étant deux à deux orthogonales et permettant de reconstruire l’information des variables initiales. Ces variables synthétiques permettent d’identifier des axes de variation deux à deux orthogonaux, ce qui offre la possibilité de comprendre la structure globale des variations sur l’ensemble des variables.

Cet exemple illustre qu’il est possible d’identifier des variables synthétiques, construites comme combinaison linéaire des variables initiales et deux à deux orthogonales. Elles permettent de mieux comprendre la **structure** globale des variations portées par les variables initiales. Nous verrons plus loin que ces variables synthétiques sont directement liées aux projections évoquées

1. $X_j(i) \approx \frac{1}{\lambda_1} \langle VarSynth, X_j \rangle VarSynth(i)$

dans les deux exemples de synthèse graphique.

Dans les chapitres qui suivent, nous avons choisi d'aborder l'ACP par la problématique de visualisation du nuage d'individus. Comme nous l'avons fait pressentir ci-dessus, c'est une problématique parmi de nombreuses autres qui conduisent à l'ACP et rien ne justifie particulièrement ce choix, si ce n'est une opinion personnelle sur la facilité de cette approche. Nous essaierons par la suite de présenter d'autres voies d'abord qu'on qualifie de "duales" pour des raisons exposées plus loin.

Pour l'instant, nous posons le cadre de travail usuel dans lequel nous évoluerons au fil des chapitres.

2.1.1 Cadre formel

On considère p variables quantitatives mesurées sur une population de n individus. Les mesures sont placées dans une $n \times p$ -matrice :

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^p \\ \vdots & & \vdots \\ x_n^1 & \cdots & x_n^p \end{pmatrix} \quad (2.1)$$

On identifie chaque variable au vecteur de valeurs observées sur les individus, et de même chaque individu est identifié au vecteur de ses valeurs pour chaque variable.

Adoptons les notations suivantes :

- x^j pour la j -ème variable, avec $x^j = X^j$
- x_i pour le i -ème individu, avec $x_i = X'_i$

Chaque individu, identifié à son vecteur de valeurs, est considéré comme un point d'un espace vectoriel F , appelé espace des individus et que l'on peut confondre avec \mathbb{R}^p . Chaque axe correspond à une variable. Cet espace est muni de la structure d'espace euclidien général (i.e. le produit scalaire adopté est défini par $\langle x, y \rangle = x'My$ où M est une matrice symétrique définie positive) afin de définir une distance entre ses éléments ($\forall x, y \in \mathbb{R}^p, d(x, y)^2 = \langle x - y, x - y \rangle$). Chaque individu x_i est associé à un poids $p_i \geq 0$ ², cet ensemble des poids vérifiant $\sum_{i=1}^n p_i = 1$. On désigne par $N = \{(x_i, p_i), i = 1, \dots, n\}$ le nuage de points pondérés.

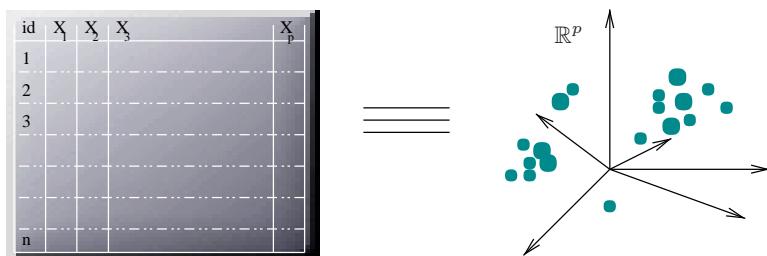


FIGURE 2.6 – Représentations des individus

De façon similaire, les variables sont considérées comme des points d'un espace vectoriel E , confondu avec \mathbb{R}^n , appelé espace des variables. Le choix du produit scalaire pour cet espace est reporté à la section 2.5.2.

2.2 L'ACP présentée comme problème de visualisation

Une façon simple de rendre compte visuellement de la forme d'un nuage est de le projeter sur des droites ou des plans, en minimisant les déformations que la projection implique. D'un point de vue

2. Il est classique de trouver dans les manuels d'ADD un système de pondération avec des $p_i > 0$. Sur un plan pratique, l'utilisation de poids éventuellement nuls est intéressante pour la gestion d'individus jugés aberrants et des individus supplémentaires. Il sera question plus loin de tels individus.

plus général, nous pouvons chercher un sous-espace affine H de \mathbb{R}^p , de dimension $d \leq p$ permettant de restituer une image de N la “meilleure possible”.

Les termes “meilleure possible” peuvent se définir à l'aide de différents critères, ce qui conduit à différentes formulations .

2.2.1 Premières formulations du problème

Ci-après, la projection orthogonale d'un point y de \mathbb{R}^n sur H est notée $\widehat{y^H}$, et éventuellement \widehat{y} en l'absence de toute ambiguïté. On note $d(y, \widehat{y})$ la distance de y à H .

Minimisation des déformations induites par la projection

Choisissons comme critère de proximité la minimisation du moment d'inertie du nuage de points par rapport à H , i.e. cherchons à minimiser $M^t(H) = \sum_{i=1}^n p_i d^2(x_i, \widehat{x}_i)$. Nous avons, d'après la relation de Huygens :

$$M^t(H) = \sum_{i=1}^n p_i d^2\left(x_i, \widehat{x}_i^H\right) = M^t(H_g) + d^2(H, H_g)$$

preuve : voir A.1 page 86.

où H_g est le sous-espace parallèle à (ou de direction) H et passant par g , le centre de gravité du nuage N . Rappelons que g a pour coordonnées les moyennes des X_k sur les individus pondérés.

Ainsi, $M^t(H) \geq M^t(H_g)$, i.e. on sait que le sous-espace cherché passe par g .

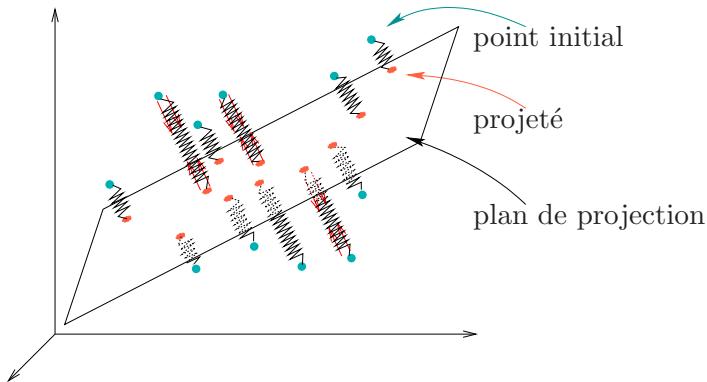


FIGURE 2.7 – Représentation physique de la première formulation

Le problème est donc $\min_{H_g} \sum_i p_i d^2\left(x_i, \widehat{x}_i^{H_g}\right)$, ce que nous pouvons écrire

$$\min_{H_g} \sum_i p_i \left\| x_i - \widehat{x}_i^{H_g} \right\|^2 \quad (2.2)$$

Maximisation des distances entre projetés pris deux à deux

Choisissons comme critère le respect des distances entre les points du nuage projeté, i.e. cherchons le sous-espace H qui maximise la quantité $Q_H = \sum_i \sum_j p_i p_j d^2(\widehat{x}_i^H, \widehat{x}_j^H)$.

Le problème est alors

$$\max_H \sum_i \sum_j p_i p_j \left\| \widehat{x}_i^H - \widehat{x}_j^H \right\|^2 \quad (2.3)$$

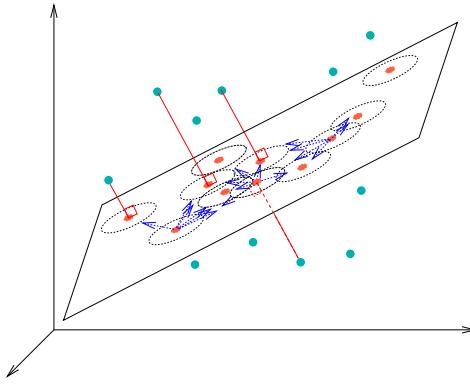


FIGURE 2.8 – Représentation physique de la seconde formulation

Equivalence des critères

Les problèmes $\min_{H_g} M^t(H_g)$ et $\max_H Q_H$ sont liés par l'égalité $Q_H = 2I_g - 2M^t(H_g)$ (cf [A.2](#) page [86](#)). Supposons que nous ayons trouvé \widetilde{H}_g solution du premier problème. Alors pour tout sous-espace affine \widetilde{H} parallèle à \widetilde{H}_g , nous avons $Q_{\widetilde{H}} = \max_H Q_H$. En effet :

$$Q_{\widetilde{H}} = 2I_g - 2\min_{H_g} M^t(H_g) = 2I_g - 2\min_H M^t(H) = \max_H (2I_g - 2M^t(H))$$

Réciproquement, si \widetilde{H} est solution du second problème, alors le sous-espace affine \widetilde{H}_g parallèle à \widetilde{H} et passant par g est solution de $\min_{H_g} M^t(H_g)$.

Ceci montre que dans un problème comme dans l'autre, c'est une direction de sous-espace que l'on cherche (autrement dit le sous-espace vectoriel sous-jacent), et qu'elle est commune aux deux problèmes.

Formulations du problème en termes d'inertie

Nous avons vu que notre recherche se limite aux sous-espaces affines passant par le centre de gravité du nuage lorsque le problème est basé sur le concept d'inertie.

Nous appelons inertie de N la quantité

$$I_g = \sum_i p_i d^2(x_i, g) = \sum_i p_i \|x_i - g\|^2 \quad (2.4)$$

Il est aisé de montrer que cette quantité égale la somme des variances des X^j .

$$\begin{aligned} I_g &= \sum_i p_i \sum_k (x_{ik} - \bar{X}^k)^2 \\ &= \sum_k \sum_i p_i (x_{ik} - \bar{X}^k)^2 \\ I_g &= \sum_k \text{var}(X^k) \end{aligned}$$

Etant donné un sous-espace vectoriel H passant par g , nous appelons inertie de N expliquée par H la quantité

$$I(H) = \sum_i p_i d^2 \left(\underbrace{x_i^H}_{g}, \underbrace{g^H}_{g} \right) = \sum_i p_i \| \widehat{x_i^H} - g \|^2 \quad (2.5)$$

Nous appelons inertie résiduelle autour de H la quantité

$$I(H^\perp) = \sum_i p_i d^2(x_i, \widehat{x_i^H}) = \sum_i p_i \|x_i - \widehat{x_i^H}\|^2 \quad (2.6)$$

i.e. la quantité expliquée par H^\perp .

Ces trois quantités sont en relations : partant de

$$\|x_i - g\|^2 = \|x_i - \widehat{x_i^H}\|^2 + \|\widehat{x_i^H} - g\|^2$$

on arrive à

$$\sum_i p_i \|x_i - g\|^2 = \sum_i p_i \|x_i - \widehat{x_i^H}\|^2 + \sum_i p_i \|\widehat{x_i^H} - g\|^2$$

i.e.

$$\underbrace{I_g}_{C^{te}} = \underbrace{I(H^\perp)}_{M^t(H)} + \underbrace{I(H)}_{M^t(H^\perp)} \quad (2.7)$$

Nous avons $M^t(H) = I(H^\perp)$, donc notre problème est de minimiser l'inertie résiduelle du nuage de points portée par le supplémentaire de H , ou, en adoptant le point de vue dual, cela revient à maximiser l'inertie du nuage expliquée par H . Comme l'inertie expliquée par H et celle de tout sous-espace de même dimension et parallèle à ce dernier sont identiques, nous nous limiterons aux sous-espaces de projection passant par le centre de gravité du nuage.

Les formulations du problème données jusqu'ici sont indépendantes de l'expression de d . Nous allons dans les chapitres suivants les particulariser selon les différents contextes : le cadre euclidien usuel (i.e. simplifié) et le cadre euclidien général.

2.3 Solution au problème dans le cadre euclidien simplifié

Nous choisissons de traiter dans un premier temps le cas le plus courant : celui où le produit scalaire est défini par :

$$\forall x, y \in \mathbb{R}^p, \langle x, y \rangle = x'y$$

i.e. le cas où la matrice définissant le produit scalaire est la matrice identité.

2.3.1 Cas où $d = 1$

Dans le cas où $d = 1$, nous cherchons un sous-espace affine de dimension 1 passant par g . Etant donné un vecteur u , avec $\|u\| = 1$, notons Δu la droite portée par u et passant par g .

Expression de l'inertie expliquée par Δu

Nous avons

$$I(\Delta u) = \sum_i p_i \|\widehat{x_i^{\Delta u}} - g\|^2$$

or $\widehat{x_i^{\Delta u}} - g = \langle x_i - g, u \rangle . u$ et donc $\|\widehat{x_i^{\Delta u}} - g\|^2 = \langle x_i - g, u \rangle^2$

$$I(\Delta u) = \sum_i p_i \langle x_i - g, u \rangle^2 = \sum_i p_i u'(x_i - g)(x_i - g)'u = u' \left(\sum_i p_i (x_i - g)(x_i - g)' \right) u$$

$$I(\Delta u) = u' \left(\sum_i (x_i - g)p_i (x_i - g)' \right) u$$

Posant $D = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$ et notant Y la matrice des données centrées ($Y = X - 1_{n \times 1} \cdot g'$), nous obtenons l'expression matricielle suivante pour l'inertie expliquée par Δu :

$$I(\Delta u) = u' Y' D Y u = u' V u$$

où $V = Y' D Y$. V est appelée matrice d'inertie du nuage. Il est aisément de remarquer que V est symétrique et positive. Notons au passage que :

$$V = \begin{pmatrix} & & \vdots & \\ \cdots & \sum_{k=1}^n p_k y_k^i y_k^j & \cdots & \\ & & \vdots & \end{pmatrix} \quad (2.8)$$

ce qui montre que V est la matrice des covariances des variables X^j .

Remarque 1 dans la plupart des cas, les données sont centrées, de sorte que $g = 0$ et $V = X' D X$.

Détermination du premier axe principal d'inertie

Nous cherchons un vecteur u de \mathbb{R}^p , normé, tel que l'inertie expliquée par la droite Δu soit maximale, cette inertie ayant pour expression :

$$I(\Delta u) = u' V u \quad (2.9)$$

La $p \times p$ -matrice V étant symétrique et positive, elle est diagonalisable, de valeurs propres positives et il existe une base orthonormée de vecteurs propres $\{u_1, \dots, u_p\}$. Nous pouvons supposer, à une permutation près, que $\lambda_1 \geq \dots \geq \lambda_p$. Ainsi u_1 correspond à la valeur propre la plus grande, etc.

Posons $u = \sum_{i=1}^p a_i u_i$. Donnons une nouvelle formulation à notre problème 2.9 :

$$\begin{aligned} I(\Delta u) &= u' V u \\ &= \sum_{i=1}^p a_i u_i' \left(V \sum_{i=1}^p a_i u_i \right) \\ &= \sum_{i=1}^p a_i u_i' \left(\sum_{i=1}^p a_i \lambda_i u_i \right) \\ I(\Delta u) &= \sum_{i=1}^p \lambda_i a_i^2 \end{aligned}$$

Notre problème s'écrit donc :

$$\max_{\sum_i a_i^2 = 1} \sum_{i=1}^p \lambda_i a_i^2 \quad (2.10)$$

auquel correspond la solution $a_1 = 1$, i.e. $u = u_1$.

preuve : voir A.6 page 88.

Ainsi, la droite Δu recherchée est portée par un vecteur propre associé à la plus grande valeur propre. Δu est appelée premier axe principal d'inertie. L'inertie portée par Δu vaut λ_1 .

2.3.2 Cas général

Lemme fondamental

Lemme 2.3.1 Soit $k < p$; si F_k est le sous-espace affine de dimension k d'inertie maximale, alors le sous-espace affine de dimension $k+1$ d'inertie expliquée maximale est $F_{k+1} = F_k \oplus \Delta u$, où Δu est la droite affine orthogonale à F_k d'inertie expliquée maximale.

preuve : voir A.7 page 88.

Ainsi, la recherche des sous-espaces F_k s'effectue de proche en proche :

— pour déterminer la droite $F_1 = \Delta u_1$, on cherche le vecteur u_1 solution du problème :

$$\max_{\|u\|=1} I(\Delta u)$$

— pour déterminer le plan F_2 , on sait que $F_2 = \Delta u_1 \oplus \Delta u_2$ où Δu_2 est solution du problème :

$$\max_{\substack{\|u\|=1 \\ u \perp u_1}} I(\Delta u)$$

— etc...

Les droites $\Delta u_1, \Delta u_2, \dots$ sont appelées premier, second, ... axe principal d'inertie du nuage. Il arrive dans la littérature que l'appellation axe principal d'inertie soit attribuée aux vecteurs u_1, u_2, \dots

Détermination du k-ième axe principal d'inertie

Par récurrence, il est facile de montrer que le k-ième axe principal d'inertie est porté par le k-ième vecteur propre de V .

En conséquence, nous avons $I(\Delta u_k) = \lambda_k$.

Solution du problème général

Finalement, la solution au problème $\max_{\substack{H \\ \dim(H)=k}} I(H)$ est donnée par :

$$F_k = \Delta u_1 \oplus \Delta u_2 \oplus \cdots \oplus \Delta u_k \quad (2.11)$$

avec

$$I(F_k) = \lambda_1 + \lambda_2 + \cdots + \lambda_k \quad (2.12)$$

Nous avons ainsi déterminé, pour une dimension donnée, la direction des sous-espaces maximisant l'inertie expliquée et l'inertie que chacun d'eux porte.

2.4 Solution au problème dans le cadre euclidien général

Dans le cadre euclidien général, le produit scalaire est défini à partir d'une matrice M symétrique, définie positive et de rang égal à la dimension de l'espace :

$$\forall x, y \in \mathbb{R}^p, \langle x, y \rangle = x' M y$$

Avant d'exposer la solution dans ce cadre, nous allons rappeler quelques définitions et des résultats classiques, transposés à notre nouveau contexte.

2.4.1 Rappels

Définition 2.4.1 Soit A une matrice carrée et M une matrice symétrique définie positive. On dit que A est M -symétrique si elle vérifie

$$MA = A'M$$

Théorème 2.4.1 Soit A une matrice M -symétrique. Alors A est diagonalisable, de valeurs propres réelles. Par ailleurs, les sous-espaces propres sont deux à deux M -orthogonaux.

2.4.2 Solution

La recherche de l'axe principal d'inertie conduit aux éléments propres de la matrice VM qui est une matrice M -symétrique. En particulier, nous retrouvons l'équivalent des résultats 2.11 et 2.12. Les détails sont renvoyés à l'annexe B.

2.5 Composantes principales : de l'espace des individus vers l'espace des variables

Le mystère des projections optimales dans l'espace des individus étant percé, nous pouvons nous intéresser aux propriétés de ces projections appelées **composantes principales**.

2.5.1 Définition et propriétés

La j -ième composante principale c_j est le vecteur contenant la coordonnée de la projection de chaque individu sur le j -ième axe principal du sous-espace affine de projection lorsque la projection de l'origine est adoptée comme référence. Comme nous supposons les données centrées, l'origine reste la référence et le sous-espace est vectoriel. Par conséquent :

$$(c_j)_i = x'_i M u_j$$

d'où

$$c_j = X M u_j$$

Relevons les propriétés suivantes pour les composantes principales :

- Les composantes principales sont des combinaisons linéaires des variables d'origine.
- Comme les variables sont centrées, les composantes principales sont des variables de moyenne nulle.

Avant de poursuivre, introduisons des éléments de formalisation nécessaires à la plongée des composantes principales dans l'espace des variables.

2.5.2 Espace des variables

Considérons les variables comme des points de l'espace \mathbb{R}^n dans lequel chaque axe est associé à un individu. Nous adoptons le produit scalaire associé à la matrice D , la matrice comportant les pondérations des individus. Nous avons donc $\langle x, y \rangle_D = x'Dy = \sum_i p_i x_i y_i$. Cette définition permet d'obtenir des mesures dont l'interprétation statistique est immédiate sous l'hypothèse de données centrées :

- $\langle x^j, x^k \rangle_D = x^{j'} D x^k = \sum_i p_i x_i^j x_i^k = \text{cov}(x^j, x^k)$
- $\|x^j\|^2 = \text{var } x^j$

Ainsi, nous considérerons que les variables sont centrées³. Notons au passage que les composantes principales satisfont les points suivants :

3. Ceci conduit à travailler sur le sous-espace orthogonal à la bissectrice de \mathbb{R}^n .

- $\langle c_j, c_j \rangle_D = \text{var } c_j = u_j M' X' D X M u_j = u_j' M V M u_j = \lambda_j$, ce que nous savions à travers la variance restituée par le j -ieme axe de projection ;
- $\forall i \neq j, \langle c_i, c_j \rangle_D = 0$

Nous adoptons également un système de pondération pour les variables, la variable X^j étant pondérée par le poids $m_j \geq 0$. Les pondérations seront placées sur la diagonale d'une matrice M . Dans le cas général, la pondération consiste à associer la valeur 1 à chaque variable - et de ce fait M correspond à la matrice identité. Ainsi, comme pour les individus, les variables sont assimilées à des points pondérés dans un espace euclidien.

L'inertie du nuage des variables par rapport à l'origine, lorsque chaque variable est pondérée à 1 est :

$$\begin{aligned} I &= \sum_j m_j d^2(x^j, 0) = \sum_j m_j \langle x^j, x^j \rangle_D = \sum_j m_j \text{var } x^j \\ I &= \sum_j \text{var } x^j \end{aligned}$$

On peut remarquer que l'inertie calculée pour le nuage d'individus par rapport à son centre de gravité correspond à cette grandeur (dans le cadre euclidien simplifié, c-à-d lorsque $M = I_d$).

Transposons à présent, dans l'espace des variables, le problème de la recherche de sous-espace de projection de dimension d permettant de maximiser l'inertie des points projetés.

2.5.3 Problème dans l'espace des variables

Cherchons un vecteur v de \mathbb{R}^n , de variance unité, tel que l'inertie⁴ des variables projetées sur la droite Δv (passant par l'origine et portée par v) soit maximale. Nous avons :

$$\begin{aligned} I(\Delta v) &= \sum_j m_j \langle x^j, v \rangle_D^2 \\ &= \sum_j m_j \text{cov}(x^j, v)^2 \\ I(\Delta v) &\stackrel{m_j=1}{=} \sum_j \text{cov}(x^j, v)^2 \end{aligned}$$

Le problème est ainsi d'identifier une variable qui maximise la somme des carrés de ses covariances avec les variables d'origine. La solution découle de ce qui suit :

$$\begin{aligned} X'Dv &= \begin{pmatrix} \text{cov}(x^1, v) \\ \vdots \\ \text{cov}(x^p, v) \end{pmatrix} \text{ d'où} \\ v'DXMX'Dv &= \sum_j m_j \text{cov}(x^j, v)^2 \end{aligned}$$

Nous pouvons observer que la matrice $XMX'D$ est une matrice D -symétrique. Elle est donc diagonalisable, de valeurs propres réelles et nous pouvons noter provisoirement ses valeurs propres $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n$ et $v_1 \dots v_n$ les vecteurs propres associés et choisis normés. La solution à notre problème devient alors évidente : il suffit de choisir v_1 pour v , c-à-d le vecteur propre associé à la plus grande valeur propre, et il en découle que l'inertie restituée sera μ_1 . Nous pouvons construire à l'aide

4. Dans l'espace des variables, nous considérerons l'inertie par rapport à l'origine.

de ces éléments propres les prochaines variables artificielles, deux à deux orthogonales et maximisant la variance des points projetés sur ces variables.

La partie qui suit montre que ces éléments propres sont en réalité déjà connus.

2.5.4 $XMX'D$ et $X'DXM$ sont de même rang

Nous pouvons observer que c_j est vecteur propre de $XMX'D$ associé à la valeur propre λ_j :

$$XMX'Dc_j = XMX'DXMu_j = XMVMu_j = \lambda_j XMu_j = \lambda_j c_j \quad (2.13)$$

Si $n = p$, alors les matrices $XMX'D$ et $X'DXM$ sont des matrices carrées de même dimension, et comme n couples d'éléments propres de $XMX'D$ sont connus, cette matrice est diagonalisable. Si $n \neq p$, les matrices $XMX'D$ et $X'DXM$ n'ont pas les mêmes dimensions, et donc, si $XMX'D$ est diagonalisable, ce qui n'est pas une évidence⁵, les p valeurs propres $\lambda_1 \dots \lambda_p$ ne constituent pas une description correcte du spectre de $XMX'D$. Nous allons montrer que les rangs des deux matrices sont identiques, puis nous montrerons que pour les λ_i non nuls la multiplicité est la même pour les deux matrices. Ceci nous permettra de conclure sur le spectre de $XMX'D$.

Partant des égalités $rg(X) = rg(X')$ et $rg(X) = rg(XX')$, et du fait que M s'écrit $OBO' = OB^{\frac{1}{2}}B^{\frac{1}{2}}O'$ avec O orthogonale et B diagonale strictement positive, nous avons :

$$\begin{aligned} rg(XMX') &= rg(XOB^{\frac{1}{2}}B^{\frac{1}{2}}O'X') \\ &= rg(XOB^{\frac{1}{2}}(XOB^{\frac{1}{2}})') \\ &= rg(XOB^{\frac{1}{2}}) \\ rg(XMX') &= rg(B^{\frac{1}{2}}O'X') \end{aligned}$$

De plus $B^{\frac{1}{2}}O'$ est inversible comme le produit de deux matrices inversibles, donc c'est la matrice d'une application bijective. Par conséquent, elle ne change pas le rang de la famille de vecteurs colonnes de la matrice X' . D'où $rg(XMX') = rg(X') = rg(X)$. De même, en raison de la bijectivité de l'application associée à $B' = B$, nous avons :

$$\begin{aligned} rg(XMX'B) &= rg((XMX'B)') \\ &= rg(B'(XMX')') \\ rg(XMX'B) &= rg(X) \end{aligned}$$

Ainsi, nous avons établi que $rg(XMX'D) = rg(X)$. De manière analogue, il est aisément de montrer que $rg(X'DXM) = rg(X)$, ce qui permet de faire le lien entre le rang des applications de matrices respectives $XMX'D$ et $X'DXM$.

Considérons les valeurs propres λ_i non nulles et les c_i associés. Comme indiqué ci-dessus, $\langle c_i, c_j \rangle = \lambda_j \delta_{ij}$, donc nous disposons d'(au moins) autant de vecteurs propres associés à λ_i pour $XMX'D$ que pour $X'DXM$. Ceci assure que la multiplicité d'un λ_i pour la matrice $XMX'D$ est au moins égale à celle pour la matrice $X'DXM$. Par conséquent, la dimension de l'image de l'application associée à la matrice $XMX'D$ est au moins égale à $rg(X'DXM)$, mais nous savons d'après ce qui précède que $rg(X'DXM)$ est précisément la dimension de cette image. Donc l'image de l'application associée à la matrice $XMX'D$ est somme directe des sous-espaces propres associés aux valeurs propres non nulles.

5. $XMX'D$ est le produit de deux matrices symétriques, donc diagonalisables. Dans le cas général, le produit de deux matrices diagonalisables n'est pas diagonalisable. Il suffit d'observer le contre-exemple suivant : $A = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ et $B = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ sont diagonalisables, mais $A \times B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ ne l'est pas !

En complétant cette somme directe par le noyau $\text{Ker}(XMX'D)$, qu'on considère comme le sous-espace propre associé à 0, il est clair que \mathbb{R}^p s'écrit comme somme directe des sous-espaces propres de $XMX'D$. Ainsi, $XMX'D$ est diagonalisable.

Nous savons que les λ_j sont des valeurs propres de cette matrice et que les valeurs propres restantes sont nulles. Ainsi, À une renumérotation près, nous posons pour $j = 1 \dots p$, $\mu_j = \lambda_j$ et $v_j = \frac{1}{\sqrt{\lambda_j}}C_j$, et pour $j > p$, $\mu_j = 0$.

2.5.5 L'ACP comme recherche de combinaisons linéaires des variables centrées et de variance maximale

D'après ce qui précède, les vecteurs propres v_j recherchés sont les composantes principales - à la norme près - c-à-d des combinaisons linéaires des variables d'origine. Nous pouvons donc reformuler le problème de l'ACP en terme de recherche de combinaisons linéaires sous certaines contraintes. Afin d'introduire ce nouveau point de vue de l'espace des variables, nous procéderons par étapes comme pour l'espace des individus.

Considérons à nouveau le problème de la recherche du premier axe principal d'inertie :

$$\max_{\|u\|=1} u'MVMu$$

i.e.

$$\max_{\|u\|=1} u'MX'DXMu$$

En posant $c = XMu$, nous obtenons

$$\max_{\|u\|=1} c'Dc$$

i.e.

$$\max_{\|u\|=1} \text{var}(XMu)$$

Le problème est donc le suivant :

1. Recherche de c_1 , la combinaison linéaire des variables centrées x^1, \dots, x^p de la forme XMu et de variance maximale sous la contrainte $u'Mu = 1$
- ⋮
- k. Recherche de c_k , la combinaison linéaire des variables centrées x^1, \dots, x^p de la forme XMu et de variance maximale sous les contraintes $u'Mu = 1, \forall i < k, \text{cov}(c_i, c_k) = 0$

Il est possible de simplifier cette écriture en posant $v = Mu$ car M est inversible. Cela nous conduit à la formulation répandue suivante :

1. Recherche de $c_1 = Xv$, la combinaison linéaire des variables centrées x^1, \dots, x^p de variance maximale sous la contrainte $v'M^{-1}v = 1$
- ⋮
- k. Recherche de $c_k = Xv$, la combinaison linéaire des variables centrées x^1, \dots, x^p de variance maximale sous les contraintes $v'M^{-1}v = 1$, et $\forall i < k, \text{cov}(c_i, c_k) = 0$

2.6 Récapitulatif des formulations du problème de l'ACP

Après avoir rappelé le cadre du problème de l'analyse en composantes principales, nous passerons en revue les différentes formulations mathématiques qui le caractérisent.

2.6.1 Cadre

On considère p variables quantitatives mesurées sur une population de n individus. Les mesures sont placées dans une $n \times p$ -matrice :

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^p \\ \vdots & & \vdots \\ x_n^1 & \cdots & x_n^p \end{pmatrix} \quad (2.14)$$

Chaque variable est identifiée au vecteur de valeurs observées sur les individus, et de même chaque individu est identifié au vecteur de ses valeurs pour chaque variable.

Les notations adoptées sont les suivantes :

- x^j pour la j -ème variable, avec $x^j = X^j$
- x_i pour le i -ème individu, avec $x_i = X'_i$

Chaque individu, identifié à son vecteur de valeurs, est considéré comme un point d'un espace vectoriel F , appelé espace des individus et que l'on peut confondre avec \mathbb{R}^p . Cet espace est muni de la structure d'espace euclidien général (i.e. le produit scalaire adopté est défini par $\langle x, y \rangle = x'My$ où M est une matrice symétrique définie positive) afin de définir une distance entre ses éléments ($\forall x, y \in \mathbb{R}^p, d(x, y)^2 = \langle x - y, x - y \rangle$). Chaque individu x_i est associé à un poids $p_i > 0$, l'ensemble des poids vérifiant $\sum_{i=1}^n p_i = 1$. On désigne par $N = \{(x_i, p_i), i = 1, \dots, n\}$ le nuage de points pondérés.

De façon similaire, les variables sont considérées comme des points d'un espace vectoriel E , confondu avec \mathbb{R}^n , appelé espace des variables. Le produit scalaire pour cet espace est défini à partir de la matrice D suivante :

$$D = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix} \quad (2.15)$$

et qui présente les particularités suivantes :

$$\langle x^i, x^j \rangle_D = x^{i'} D x^j = \text{cov}(X^i, X^j)$$

et

$$\langle x^i, x^i \rangle_D = x^{i'} D x^i = \text{var } X^i$$

2.6.2 Recherche des sous-espaces maximisant l'inertie du nuage projeté de points

La recherche porte sur les sous-espaces affines de \mathbb{R}^p de dimension $d \leq p$ maximisant l'inertie de la projection orthogonale du nuage sur le sous-espace.

$$\max_H \sum_i p_i \|\widehat{x_i^H}\|^2$$

Ce point de vue est souvent adopté dans les problématiques de visualisation des données et dans ce cas, $d = 2$ ou 3 . Les projections sur les nouveaux axes sont les composantes principales.

2.6.3 Recherche de combinaisons linéaires des variables maximisant la variance

Dans le but d'analyser et de synthétiser la variance observée sur la population, on cherche les combinaisons linéaires des variables observées X^j de variance maximale et de covariance nulle. De façon plus formelle :

1. Recherche de $c_1 = Xv$, la combinaison linéaire des variables centrées x^1, \dots, x^p de variance maximale sous la contrainte $v'M^{-1}v = 1$
- ⋮
- k. Recherche de $c_k = Xv$, la combinaison linéaire des variables centrées x^1, \dots, x^p de variance maximale sous les contraintes $v'M^{-1}v = 1$, et $\forall i < k, \text{cov}(c_i, c_k) = 0$

Ces combinaisons linéaires sont les composantes principales.

2.7 Mise en œuvre et interprétation des résultats

Nous nous appuierons, à titre d'illustration, dans cette section sur les résultats d'une analyse en composantes principales appliquée aux données centrées et réduites de l'étude des budgets-temps proposée par [JAM76]. Les données sont fournies en annexe page 95. Les individus sont pondérés de façon égale et le produit scalaire adopté dans l'espace des individus est le produit scalaire euclidien. Nous sommes dans le cas le plus standard.

2.7.1 Pré-traitements des données

Les pré-traitements présentés ci-dessous sont très courants et se justifient fortement dans la plupart des analyses, mais ils ne constituent pas une obligation.

Centrage

Le premier traitement appliqué aux données est le centrage. En effet, si nous formulons le problème comme une recherche de sous-espace affine de dimension donnée portant la plus grande part d'inertie, nous savons que ce sous-espace passe par le centre de gravité du nuage de points et que cette recherche se ramène en une recherche de sous-espace vectoriel sous l'hypothèse où le centre de gravité est confondu avec l'origine. Le centrage des données permet de se placer dans ce cas. Mais cette justification technique est peu motivante par rapport au point de vue statistique suivant : le centrage des données permet d'adopter comme point de référence le centre de gravité du nuage d'individus. Ce choix de référence semble adéquat à l'analyse des variations des variables autour de leurs moyennes et aux liaisons inter-variables.

Réduction

La réduction des données consiste à travailler avec les “variables réduites”, i.e. avec les variables divisées par leur écart-type. Ainsi toutes les variables réduites sont de variance 1 et nous sommes conduits à analyser la matrice des corrélations plutôt que celle des covariances. Ceci comporte les avantages suivants :

- les variables étant sans dimension, les distances sont indépendantes des unités de mesure. Les variations des variables sont alors comparables.
- toutes les variables ont la même importance, i.e. des variables égales à un facteur près sont considérées égales. Ceci est important, car finalement on étudie la nature des variations. Sur un plan pratique, cela permet d'éviter le phénomène d'écrasement du nuage projeté que l'on rencontre lorsqu'une minorité de variables présentent des valeurs si fortes qu'elles constituent à elles seules l'inertie du nuage d'individus.
- l'absence d'unités permet des comparaisons transversales directes de résultats (i.e. entre analyses).

On peut penser invalider le premier point dans le cas où toutes les variables sont mesurées avec la même unité, mais même dans ce cas, si les variances des différentes variables sont disproportionnées, il est préférable de réduire.

La réduction des variables constitue la deuxième étape logique (après le centrage) de l'étude des variations des variables débarrassées des éventuels effets perturbateurs d'échelle. Cela permet une analyse plus fine de la nature des variations : les projections des variables "centrées réduites" sur les composantes principales ont pour coordonnées leurs coefficients de corrélation. Ainsi, la visualisation des variables projetées donnent une idée précise de l'intensité de leurs corrélations avec les composantes principales et de la qualité de leur représentation à l'aide des composantes principales.

Signalons, pour conclure cette partie sur la réduction des données, une dernière remarque importante : contrairement à ce que l'intuition pourrait suggérer, il n'y a pas de relations simples entre le sous-espace de projection déterminé dans le cas de données réduites et celui correspondant au cas non-réduit. En effet, la réduction des variables constitue une dilatation du nuage selon les axes constitués par les variables et ceci induit une réelle modification sur les vecteurs propres cherchés.

2.7.2 Calcul des éléments propres

Les données étant centrées, les vecteurs portant les axes principaux sont donnés par les vecteurs propres de la matrice $VM = X'DXM$ et les composantes principales sont les vecteurs propres de la matrice $XMX'D$. La façon la plus rapide pour procéder au calcul de ces éléments est de calculer les éléments propres de la plus petite des matrices $VM = X'DXM$ et $XMX'D$, puis de déterminer, à partir des éléments calculés et en utilisant les formules de dualité, les éléments propres de l'autre matrice.

2.7.3 Décomposition de la matrice de covariance/corrélation

Notons de façon générale V la matrice analysée. Il s'agit de la matrice des covariances dans le cas de données centrées ou de la matrice des corrélations dans le cas de données centrées réduites. Les valeurs propres λ_i et les vecteurs propres u_i sont extraits de VM . La matrice V vérifie :

$$V = \lambda_1 u_1 u_1' + \cdots + \lambda_p u_p u_p'$$

Cette décomposition spectrale de V est intéressante car chaque matrice $\lambda_i u_i u_i'$ s'interprète comme la matrice de covariances (ou de corrélations selon la nature de V) des projections des variables X^j sur la composante principale c_i . Le terme de décomposition spectrale est justifié. Il est facile de voir qu'étant donné un ensemble I d'indice des composantes principales, la somme sur I des matrices $\lambda_i u_i u_i'$ s'interprète comme la matrice de covariances (ou corrélations) des projections des variables X^j sur le sous-espace vectoriel engendré par les composantes principales en jeu.

2.7.4 Choix du nombre d'axes

Dans un but exclusif de visualisation des données, la dimension du sous-espace de projection est 2 ou au plus 3. Mais dans un objectif de réduction de la dimensionnalité de l'information portée par le nuage, le problème est justement de déterminer la dimension de l'espace de projection. On dispose de plusieurs critères empiriques pour répondre à cette question. En fait, il existe probablement autant de critères de sélection d'axe qu'il y a de façons de définir la valeur d'un axe. Nous exposons ci-après les plus répandus, ainsi que certains d'usage rare mais dont la connaissance est importante.

Critère basé sur un seuil de qualité de représentation

Le critère de sélection le plus simple consiste à fixer un seuil minimal de qualité de représentation du sous-espace de projection. Sont alors choisis les m premiers axes principaux, où m est le plus petit entier permettant de satisfaire cette contrainte sur la part d'inertie expliquée par le sous-espace de projection.

Critère de KAISER

Dans le cas où toutes les variables d'origine sont centrées, réduites et indépendantes, il est clair que les composantes principales sont les variables d'origines et qu'elles ont pour variance 1. Si l'indépendance n'est pas vérifiée, la répartition de l'inertie selon les composantes n'est plus isotrope. Le critère de Kaiser consiste à retenir les seuls axes dont la part d'inertie expliquée est supérieure à 1, i.e. à la part d'inertie expliquée moyenne. Le seuil de rejet théorique pose problème lorsqu'il sépare des composantes dont les contributions sont très proches de 1 (par exemple 1.01 et 0.99). Afin de limiter le rejet de composantes principales pertinentes, ce seuil peut être baissé. D'après [JOL86], 0.7 est une bonne valeur.

Ce critère peut se généraliser aux variables non-réduites de la façon suivante : l'inertie du α -ième axe d'inertie étant la α -ième valeur propre λ_α et l'inertie totale du nuage de points étant $Tr(V)$, le α -ième axe ne devrait être retenu que si $\lambda_\alpha > \frac{Tr(V)}{p}$ et $\lambda_\alpha > 0.7\frac{Tr(V)}{p}$ pour la version "adoucie". Mais il ne faut pas perdre de vue que ce critère a été construit pour l'analyse de matrice de corrélations : sa justification dans le cas de matrice de covariances reste fragile... Mais le précepte selon lequel il vaut mieux un critère mal justifié que pas de critère du tout est souvent suivi.

Règle du point d'inflexion ou du coude

Sur le profil des valeurs propres triées par ordre décroissant, s'il existe un point d'inflexion, seules les valeurs propres situées avant ce point sont retenues. S'il existe plusieurs points d'inflexion, on retient les valeurs situées avant le premier.

Ce critère a été mis au point dans le cadre de l'analyse factorielle. Ainsi son usage est discutable dans le cadre d'une ACP, mais son utilisation comme les deux précédents critères est très courante.

	λ_i	% d'inertie cumulée	Kaiser	$d_j = \lambda_j - \lambda_{j+1}$ diff. premières	$d'_j = d_j - d_{j+1}$ diff. secondes
1	4.617	46.2%	4.617	2.516	1.749
2	2.100	67.2%	2.100	0.767	0.602
3	1.332	80.5%	1.332	0.164	-0.535
4	1.168	92.2%	1.168	0.700	0.436
5	0.467	96.9%	0.467	0.264	0.108
6	0.203	98.9%	0.203	0.156	0.145
7	0.047	99.4%	0.047	0.010	-0.004
8	0.037	99.8%	0.037	0.015	-0.004
9	0.022	100%	0.022	0.020	
10	0.002	100%	0.002		

TABLE 2.2 – Indicateurs tirés des valeurs propres

Critère statistique d'inégalité des valeurs propres

Ce critère qui recourt aux statistiques est basé sur l'hypothèse que les données sont issues d'une population gaussienne. Le caractère limitatif de cette contrainte va être accentué par la complexité de la procédure cherchée.

Il peut se produire, lors d'une ACP, que les dernières valeurs propres soient très proches de zéro. Il est alors intéressant de tester que ces valeurs soient statistiquement différentes de zéro afin de juger de leur éventuel caractère informatif. Plus généralement, il est possible de tester si les q dernières valeurs propres sont statistiquement différentes. En partant de $q = 2$ et en faisant croître q jusqu'à l'obtention d'une différence significative, nous pourrions espérer déterminer $p - q$ variables correspondant à une part d'information plus pertinente. Malheureusement, les seuils de significativité sont d'une expression peu commode quand ils sont connus... Il existe toutefois des cas où cette procédure est applicable, mais on observe qu'elle a tendance à sous-estimer le nombre de composantes d'intérêt. La complexité du

tissu mathématique de ce critère semble trop dense pour l'application que l'on peut faire des résultats d'une ACP de façon générale.

Remarque 2 Notre énumération de critères s'arrête là d'une part pour ne pas soulever un nouveau problème, celui du choix d'un critère, et d'autre part ceux que nous pourrions proposer comportent aussi des biais. Ce qu'il faut retenir au sujet de ces critères, c'est qu'ils s'orientent rarement vers des indications concordantes, ils peuvent pointer vers des informations contradictoires dans certains cas, et que le choix d'un critère particulier est contextuel. Par conséquent, un critère mal choisi peut conduire à une sélection d'axes éloignée de celle attendue... A ces critères, nous ajoutons les deux remarques suivantes qui constituent en quelque sorte des rappels de bon sens à utiliser conjointement à un critère.

Remarque 3 L'intérêt que l'on porte à un axe ou une composante principale doit tenir compte de la part d'inertie expliquée mais aussi du nombre de variables total.

Remarque 4 A ces critères qui fournissent une première aide pour sélectionner les axes, il faut ajouter le principe de ne retenir que les composantes principales interprétables. L'introduction de variables supplémentaires peut contribuer à éclaircir l'interprétation qu'on peut faire des composantes principales.

Choix du nombre d'axes dans l'analyse des budgets-temps

Le graphique 2.9 donne le profil des valeurs propres et fournit une première impression de la décomposition factorielle. Le premier facteur concentre à lui seul presque la moitié de la totalité de l'inertie (46.17%). La redondance semble assez importante, puisque les quatre premiers facteurs rassemblent plus de 92% de l'inertie. Ainsi les quatre premiers facteurs fournissent une bonne approximation de la structuration de la population sur les variables mesurées. Davantage de précision est apportée par le tableau 2.2.

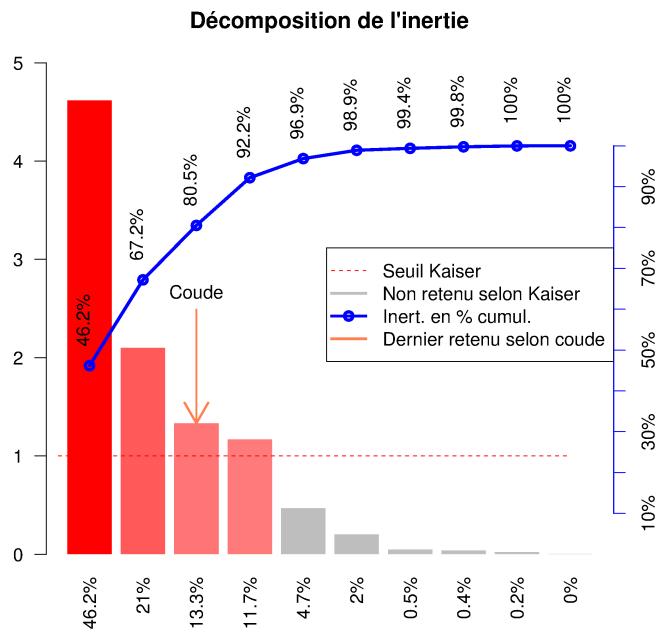


FIGURE 2.9 – Valeurs propres de V et critères

Les critères de Kaiser et du coude corroborent le pressentiment que les trois ou quatre premiers facteurs permettent une projection de bonne qualité : KAISER propose une sélection des quatre premiers facteurs et la règle du coude un de moins. Mais entre trois et quatre, que choisir ? La règle

du coude trouve son utilité dans les cas où les valeurs propres ont une décroissance trop régulière dans une région où l'on aimerait trancher. La présence de point d'infexion peut alors aider à faire son choix. Dans la situation présente, la décroissance des valeurs propres présente plusieurs cassures dont l'une d'entre elles coïncide avec la séparation de KAISER et à un niveau de restitution de l'information intéressant. La règle du coude n'est pas intéressante ici, et donc le choix des quatre premiers facteurs semble approprié.

2.7.5 Les graphiques des projections des nuages

Les représentations graphiques des projections des individus et/ou des variables peuvent constituer le résultat final ou un simple résultat intermédiaire d'une ACP. Ci-dessous, nous présentons la démarche de construction de ces graphiques et quelques aménagements visant à restituer une partie de l'information perdue par l'opération de projection.

Projections des individus

Si la dimension, déterminée à l'aide des critères précédents, du sous-espace de projection est deux ou trois, alors la projection complète du nuage pourra être visualisée à l'aide d'un seul objet graphique (puisque il s'agira d'un objet de dimension au plus trois). Dans le cas contraire, la projection complète ne peut toujours pas être appréhendée et dans ce cas, il est d'usage d'avoir recours aux projections partielles, i.e. aux projections sur les couples d'axes principaux. Il est courant de commencer par les plans portant la plus grande part d'inertie puis d'aller ainsi en décroissant.

Une façon d'observer les déformations induites par les opérations de projection est d'observer en même temps que les projections des individus la projection d'un arbre recouvrant minimal pour la distance de l'espace des individus.

Pour finir, mentionnons une astuce qui trouve son utilité dans le cas où les individus ne sont pas tous affectés du même poids. Dans ce cas, les individus qui contribuent le plus à l'inertie du nuage ne sont pas forcément ceux qui sont les plus éloignés du centre d'inertie. Il peut être intéressant de dénoter par un légendage particulier d'une part les points qui contribuent fortement à l'inertie du nuage initial (i.e. les individus de fort $CONTR(i)$... Cf. 2.7.6 page 29) et d'autre part les individus qui contribuent fortement à la détermination de l'un des axes principaux représentés sur le graphique (i.e. ceux de fort $CTR_j(i)$... Cf. 2.7.6).

Projections des variables

Les projections des variables sur les composantes principales synthétisent une information statistique sur les covariances entre les variables initiales et les variables artificielles que constituent les composantes principales. Dans le cas de variables initiales réduites, les covariances s'interprètent comme des corrélations.

Il arrive de trouver dans des analyses des représentations simultanées des projections d'individus sur un couple d'axes principaux et des projections des variables sur les composantes principales associées. Bien évidemment cette superposition d'espaces de natures différentes limite les interprétations déductibles de tels graphiques. Les projections des variables sont à interpréter comme des directions. La prise en compte de ces directions et du centre de gravité du nuage permet de situer les individus en terme de variations par rapport aux moyennes des variables associées aux directions. En général, cette démarche constitue la première étape d'une analyse que l'on affine à l'aide des indicateurs qui seront présentés plus bas.

Projections dans l'analyse des budgets-temps

Les figures 2.10 et 2.11 suivantes sont les projections du nuage d'individus (resp. des variables) dans les plans engendrés par les axes principaux (resp. composantes principales) pris deux à deux.

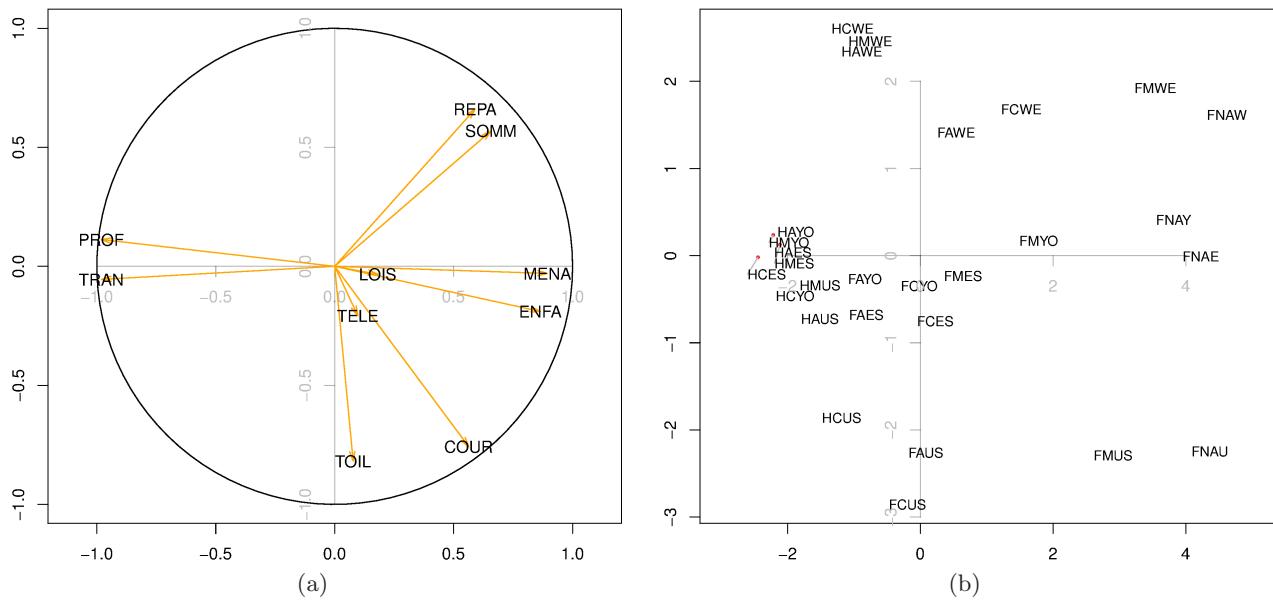


FIGURE 2.10 – Projections des nuages de variables et d’individus dans le plan porté par u_1 et u_2 (plan expliquant 67.17%)

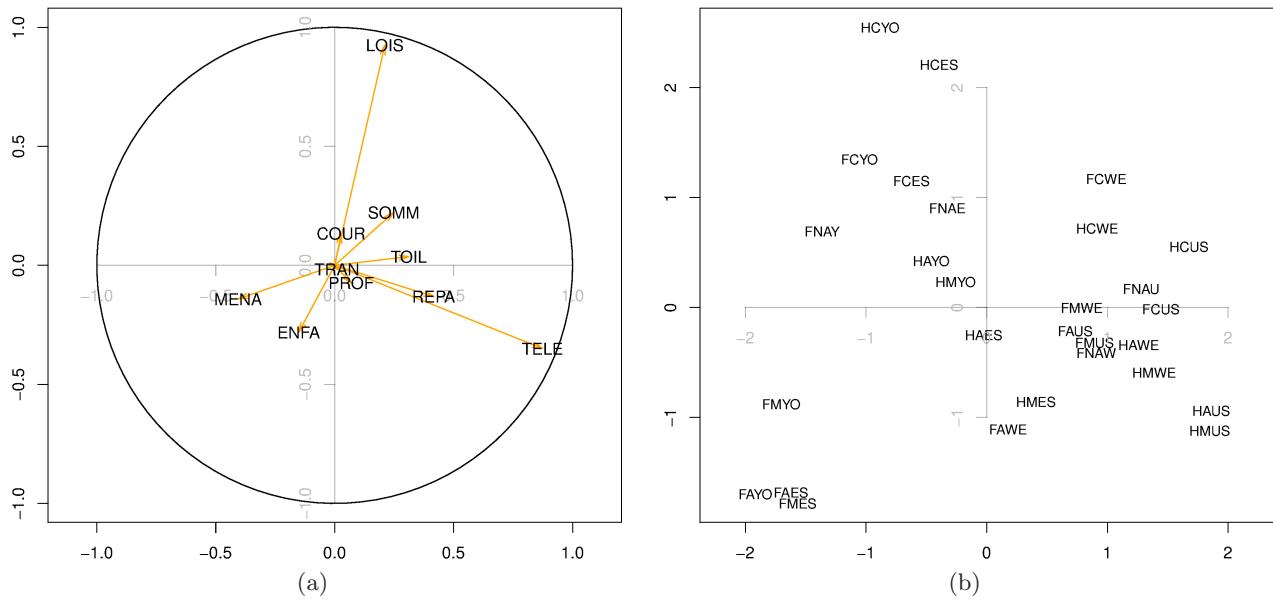


FIGURE 2.11 – Projections des nuages de variables et d’individus dans le plan porté par u_3 et u_4 (plan expliquant 25.01%)

Nous verrons en 2.7.6 les interprétations qui peuvent être proposées à partir de ces graphiques en s’aidant des indicateurs de qualité et de contribution associés aux individus. Nous pouvons toutefois faire deux remarques :

- Certaines projections des variables sont très proches du bord du disque de corrélation. Celles-ci sont donc de bonne qualité. Certaines de ces projections mettent en évidence, par leur proximité dans le disque, des corrélations très fortes tandis que d’autres soulignent des absences

de corrélation par leur apparente orthogonalité dans les plans de projection. Il y a donc des groupes de variables qui sont observables.

- Chaque facteur est fortement lié à certaines variables dont il permet une projection presque parfaite. Ceci facilitera l'interprétation des facteurs.

2.7.6 Aides à l'interprétation

Une ACP permet de faire ressortir une information sur la structuration des variables et le positionnement des individus les uns par rapport aux autres. Nous allons voir les outils que l'ACP fournit dans ce but (des indicateurs principalement) et les schémas classiques d'interprétation des graphiques, ainsi que des méthodes permettant d'utiliser une information extérieure aux données initiales afin de clarifier ou de structurer davantage les résultats de l'ACP et de faciliter leur interprétation.

Indicateurs

Les indicateurs proposés se classent en deux familles : les indicateurs globaux dont la valeur est établie par rapport à toute l'information portée par le nuage et les indicateurs locaux construits par rapport à un(e) axe(variable) particulier(e).

Indicateurs de contribution Le premier indicateur global intéressant calculé pour chaque individu est la contribution à l'inertie du nuage d'un individu :

$$CONTR(i) = \frac{p_i \|x_i\|^2}{I_0} \quad (2.16)$$

Cet indicateur permet de repérer les individus qui contribuent plus que la moyenne à l'inertie du nuage. Dans le cas où tous les individus ont le même poids (i.e. $\forall i, p_i = \frac{1}{n}$), il s'agit de ceux pour lesquels $CONTR(i) > \frac{1}{n}$ et de ceux pour lesquels $CONTR(i) > p_i$ dans le cas général. Ces individus sont souvent ceux qui orientent les résultats d'une analyse : en effet, ils s'avèrent souvent indiqués par les autres indicateurs pour participer à l'interprétation des axes principaux. Cet indicateur se décline localement à chaque axe principal comme suit :

Partant de

$$I(\Delta_{u_j}) = \lambda_j = \sum_{i=1}^n p_i (c_i^j)^2 \quad (2.17)$$

nous exprimons la contribution de l'individu i à l'inertie de l'axe Δ_{u_j} par la quantité

$$CTR_j(i) = \frac{p_i (c_i^j)^2}{\lambda_j} \quad (2.18)$$

Les individus pour lesquels $CTR_j(i) > p_i$ sont les individus qui contribuent plus que la moyenne à l'inertie portée par l'axe principal i . Les individus de fort CTR servent à l'interprétation de l'axe Δ_{u_j} . En effet, les individus extrêmes sur un axe principal sont ceux qui accentuent le mieux les caractéristiques de la tendance liée à cet axe.

Remarque 5 Les p_i interviennent dans l'expression des CTR mais pas sur les graphiques : si les p_i sont tous égaux, l'examen d'un graphique relatif à la projection des individus sur un axe particulier permet de repérer les individus de fort CTR sur cet axe, mais les p_i sont différents d'un individu à l'autre, il faut examiner les valeurs des CTRs pour se rendre compte des rôles des individus.

Ces indicateurs de contribution par individu sont déclinables aux différents sous-espaces de projection.

Indicateurs de qualité La qualité de la représentation de l'individu i sur l'axe Δ_{u_j} est exprimée par la quantité

$$CO2_j(i) = \frac{(c_i^j)^2}{\|x_i\|^2} \quad (2.19)$$

qui vérifie

$$\sum_j CO2_j(i) = 1 \quad (2.20)$$

La qualité de la représentation de l'individu i sur le sous-espace $\Delta u_1 \oplus \dots \oplus \Delta u_k$ est exprimée par la quantité

$$QLT_k(i) = \sum_{j=1}^k CO2_j(i) \quad (2.21)$$

[VOL81] préconise l'examen de l'indicateur d'écart relatif au sous-espace de projection suivant :

$$ECART_H(i) = \frac{p_i \|x_i - \widehat{x_i^H}\|^2}{\sum_i p_i \|x_i - \widehat{x_i^H}\|^2}$$

En effet, cet indicateur relativise la mauvaise qualité d'une projection par rapport à la somme des erreurs de projection commises sur la population. Il permet de se concentrer sur les points dont la projection est plus mauvaise que la moyenne. L'indicateur QLT a tendance à être systématiquement mauvais pour les individus proches du centre d'inertie, défaut que n'a pas l'indicateur $ECART_H$. Par ailleurs, $ECART_H$ permet une recherche des facteurs sur lesquels la projection s'opère mal. Toutefois, cet indicateur est peu utilisé dans la pratique, l'information apportée par les indicateurs précédemment présenté étant souvent suffisamment riche.

Les figures 2.12a et 2.12b suivantes sont les projections du nuage des variables dans les plans engendrés par les composantes principales prises deux à deux et utilisant le COS2 associé à chaque variable dans le plan considéré afin de mettre en évidence les variables les mieux représentées.

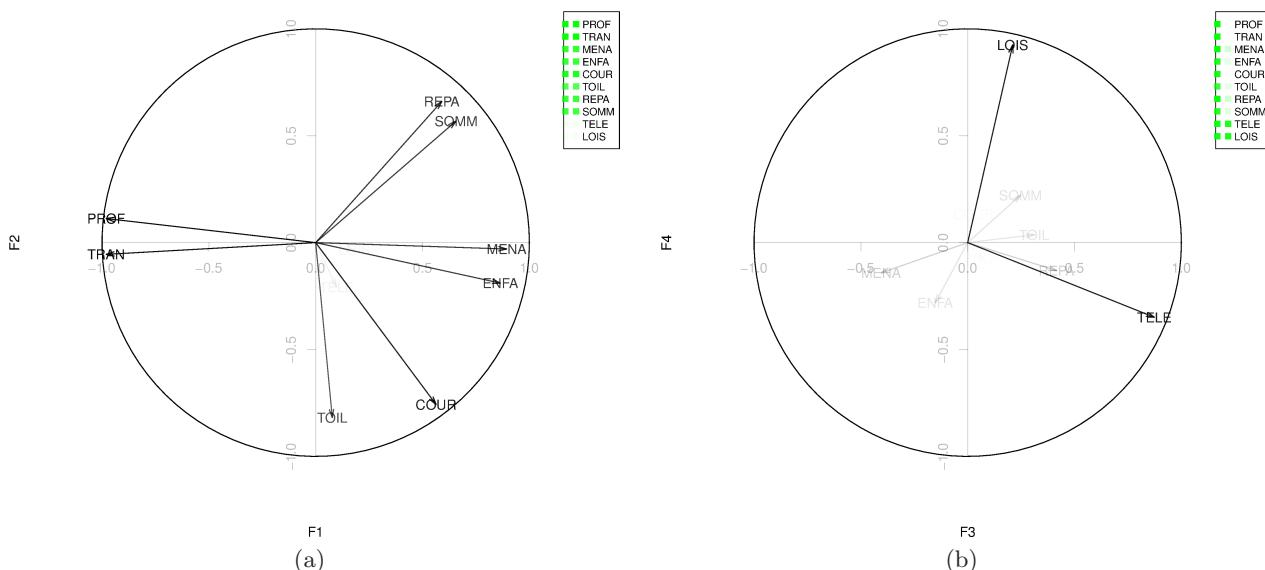


FIGURE 2.12 – Projections des nuages de variables exploitant les COS2

Le légendage tient compte du COS2 dans le plan de projection et des COS2 cumulés jusqu'à ce plan de projection.

Remarque 6 Tous ces indicateurs sont souvent rassemblés dans des tableaux afin de repérer ce qu'il y a de remarquable. Cela conduit le plus souvent à des tableaux de chiffres volumineux et difficilement lisibles. Il est alors important de se donner les moyens de lire de tels gros tableaux. Divers niveaux de simplification de ces tableaux sont envisageables dans ce but : en premier lieu, on peut limiter la précision des indicateurs à un niveau rendant la lecture moins pénible et gardant la majeure partie de l'information apportée par les indicateurs. Mais dans certains cas, il est possible d'aller jusqu'à une classification des valeurs pour synthétiser au maximum l'information d'un indicateur (corrélation positive très forte, moyennement forte, négative très forte, moyennement forte et faible corrélation par exemple...). De telles astuces peuvent réellement éclaircir des résultats.

Interprétation des axes et des composantes principaux

Dans la plupart des cas, il est tenté par l'analyste d'interpréter les axes un par un, et la fréquence de cette réalité peut faire oublier qu'il existe des cas où cette démarche n'a pas de sens. En effet, dans le cas où plusieurs axes (ou composantes) correspondent à une même valeur propre, l'ensemble de ces axes définit le sous-espace propre associé à cette valeur mais réciproquement, cette base de l'espace propre n'est pas unique (i.e. tout ensemble d'axes qui sous-tendrait le même sous-espace propre pourrait remplacer les axes fournis par la procédure de calcul de l'ACP). Ainsi, cela n'a pas de sens d'interpréter isolément un tel axe. C'est le sous-espace complet associé à une valeur propre qui, considéré comme sous-espace de projection, peut donner lieu à des interprétations.

L'examen des valeurs propres est donc la première étape de l'interprétation d'une ACP. Les résultats d'une ACP pratiquée sur les données issues d'un échantillonnage (cas de 99% des analyses) doivent être lus avec la conscience des effets propres à l'échantillonnage : les valeurs propres observées sont des estimations des valeurs propres effectives de la population et de ce fait, deux valeurs propres proches peuvent indiquer un plan de projection plutôt que deux axes de projections. Le problème soulevé ici est complexe et nous renvoyons à [JOL86] pour des références sur la distribution des valeurs propres.

Toutefois, ce cas de figure est rare et nous le supposerons écarté pour la suite (i.e. dans les sous-sections suivantes, nous supposons que les valeurs propres sont significativement distinctes). Voyons à présent les démarches les plus classiques d'interprétation dans le cas de données centrées réduites.

Etude de la k-ième composante principale du nuage de variables La projection de la variable X^j sur la composante c_k est le coefficient de corrélation de la variable X^j avec c_k . Cette propriété est utile pour tirer des graphiques des projections du nuage de variables des informations sur les dépendances et des non-corrélations et sur le sens des corrélations éventuelles ainsi que les corrélations partielles. Afin de faciliter encore la lecture de cette information sur ces graphiques de projections planaires, il est d'usage de tracer le cercle de corrélation.

Nous pouvons idéaliser les possibilités selon cinq cas :

- Un groupe de variables apparaît proche de l'intersection, située au point $(1, 0)$, du cercle de corrélation et d'un axe (l'axe horizontal par exemple). Notons ce groupe le groupe 1.
- Un (autre) groupe de variable est proche de l'autre intersection du cercle de corrélation et de l'axe horizontal située au point $(-1, 0)$. C'est le groupe 2 de variables.
- D'autres variables sont proches d'une intersection du cercle de corrélation et de l'axe vertical. Elles constituent le groupe 3.
- Les variables proches du bord du cercle qui ne semblent pas attirées par l'un des axes forment le groupe 4.
- Enfin, des variables peuvent apparaître éloignées du cercle de corrélation. Elles sont d'un intérêt moindre pour l'interprétation des axes.

Les variables des groupes 1 et 2 sont fortement corrélées entre elles, mais celles du groupe 1 sont opposées à celles du groupe 2. La composante principale correspondant à l'axe horizontal peut être interprétée comme une nouvelle variable, fonction linéaire pour laquelle les variables du groupe 1 (resp. 2) contribuent positivement (resp. négativement). Un individu qui a une valeur positive élevée pour l'une des variables du groupe 1 devrait avoir de grandes valeurs positives pour les autres variables du groupe 1 et de grandes valeurs négatives pour celles du groupe 2. Le lien entre le nuage d'individus et le nuage de variables apparaît. Si la part d'inertie portée par cette composante est très importante, le remplacement des variables du groupe 1 et 2 par cette nouvelle variable se fait moyennant un perte d'information maîtrisée.

Les variables du groupe 3 sont non corrélées avec celles des groupes 1 et 2. Les individus ont un comportement, sur les variables de ce groupe, nouveau par rapport à celles des groupes précédents. Cette composante principale traduit une nouvelle dimension de la variabilité présente dans les données. Elles s'interprète à l'aide des variables les plus fortement corrélées avec elle.

Les variables du groupe 4 sont celles qui dépendent à la fois de ces deux composantes principales.

Etude du k-ième axe principal du nuage d'individus L'interprétation d'un axe, avec pour support les graphiques des projections des individus et les tableaux de valeurs des différents indicateurs, commence par la sélection des individus dont la contribution relative à l'inertie de l'axe est importante i.e. tels que $CTR_k(i) > p_i$. Elle a pour but d'isoler des individus aux caractéristiques extrêmes soulignant les éventuels phénomènes sous-jacents à cet axe. Ces phénomènes doivent être connus en grande partie après l'analyse du nuage de variables. Il peut être intéressant de limiter la recherche précédente aux seuls individus qui contribuent plus que la moyenne à l'inertie du nuage d'individus ou au moins de les repérer afin d'observer leur rôle dans l'espace de projection. Enfin pour les points retenus par l'attention de l'analyste la qualité de la projection sur l'axe doit être examinée. Les individus dont la projection sur un axe est "bonne" doivent servir prioritairement à caractériser un axe.

Ces sélection étant effectuées, les points retenus peuvent permettre d'attacher plus de sens "commun" aux nouvelles variables que constituent les composantes principales. Des descriptions d'axe en terme de comportements d'individus peuvent être plus clair que les coefficients de contribution des variables fortement corrélées avec la composante associée.

Effet taille

L'effet taille est une manifestation particulière que l'on observe sur les graphiques de projection des variables. Dans de nombreuses analyses portant sur des mesures anatomiques effectuées sur des individus d'une même espèce, le premier facteur de variabilité apparaissant est corrélé positivement avec toutes les variables, ce qui se traduit sur les graphiques de projection des variables sur ce facteur par la répartition de tous les points projections dans le demi-plan correspondant à des valeurs positives pour cette composante principale. Ce facteur montre que toutes les variables ont tendance à augmenter en même temps. Ce facteur reflète la taille des individus : plus les individus sont grands et plus les autres mesures ont tendance à augmenter. Lorsqu'une ACP conduit à l'observation de corrélations de même signe de toutes les variables avec une composante principale, on dit qu'on observe un effet taille.

Eléments supplémentaires

Les composantes et axes principaux calculés par une ACP peuvent être utilisés pour projeter des éléments supplémentaires d'information sur les différents nuages. Ces éléments qui interviennent à postériori peuvent être des individus nouveaux ou des variables délaissées volontairement pendant l'analyse par exemple. Dans le cas des variables, il peut s'agir de variables quantitatives ou qualitatives. Ces éléments, individus ou variables, sont appelés éléments supplémentaires ou illustratifs, alors que les données initiales sont dénommées éléments actifs.

Les individus supplémentaires apporteront une information qui n'apparaîtra pas dans les projections des nuages de variables. Les variables quantitatives pourront être projetées sur les différentes composantes principales mais n'apporteront rien de nouveau sur les projections du nuage d'individus. En revanche, des variables qualitatives (ou des variables considérées comme telles), qui ne peuvent être projetées sur l'espace des variables en raison de leur nature, peuvent apporter une information dans l'espace des individus. Voyons comment ces différents éléments d'information s'insèrent dans les résultats de l'ACP :

- Individu supplémentaire : avant de projeter un individu supplémentaire, il est nécessaire d'appliquer aux données qui le caractérisent les mêmes pré-traitements qu'aux données actives (centrage, réduction,...). La projection se fait alors alors naturellement sur les axes principaux comme pour les autres individus. Quand rencontre-t-on de telles projections ? Dans le cas où des individus ont été mis de côté à cause d'un doute portant sur la qualité de leurs valeurs (individus jugés aberrant...) ou encore dans le cas où les composantes ont été calculées à partir d'une population référence et que les individus supplémentaires constituent des groupes que l'on cherche à positionner par rapport à la population référence. Ce second cas peut constituer une démarche intéressante pour vérifier la robustesse des interprétations émises à partir des données initiales.
- Variable quantitative supplémentaire : comme précédemment, il faut appliquer aux variables quantitatives supplémentaires les mêmes pré-traitements qu'aux données actives. Les variables ainsi préparées peuvent se projeter sur les différentes composantes principales.
- Variable qualitative supplémentaire : une variable qualitative permet d'apporter une information sur les projections du nuage d'individus, et ce de différentes façons. La plus simple consiste à légendrer chaque projection d'individu par la modalité prise pour la variable qualitative. Ceci permet d'observer si une répartition particulière des modalités de cette variable apparaît. Un aperçu plus global de la même information peut être donné en affichant la projection du centre de gravité de la sous-population associée à une modalité pour chaque modalité. Si cette variable est de plus ordinaire, le tracé de la jonction de ces points selon l'ordre existant sur les modalités peut mettre en évidence une structuration intéressante pour l'interprétation des composantes principales.

Interprétation dans l'analyse budget temps

Le tableau suivant regroupe pour chaque individu les indicateurs de contribution à l'inertie totale, les indicateurs de contribution et de qualité de représentation pour les deux premiers axes principaux.

L'analyse du tableau 2.3 nous pousse à porter notre attention sur FNAU, FNAE, FNAW et FMWE qui ont contribué largement à la détermination du premier axe principal avec des coordonnées positives sur ce dernier et sur HCES, FCES, HMYO et HAES avec des coordonnées négatives. De même, pour le deuxième facteur, FAUS, FNAU, FMUS et FCUS s'opposent par leurs coordonnées à HAWE, HMWE et HCWE avec d'importantes contributions pour tous.

D'après le tableau 2.4 les individus FAYO, FMYO, FAES, HAUS, HMUS et HCUS contribuent de façon importante à la formation du troisième axe, avec des coordonnées positives pour les trois premiers et des coordonnées négatives pour les autres. Enfin, HCYO, HCES, FCES, FAYO, FAES et FMES jouent les mêmes rôles pour le dernier axe principal considéré.

Partant des graphiques des projections 2.10 et 2.11 et du tableau précédent, nous pouvons proposer les interprétations suivantes :

- D'après la figure 2.10a, le premier facteur oppose les projections des activités liées au travail professionnel (transport et travail) et les activités du travail au foyer (ménage et enfants). Ces activités sont très bien représentées par ce premier facteur. Nous pouvons en conclure que ces oppositions concernent bien les variables (non leurs seules projections...) sur la population étudiée. La signification qui peut être attachée au premier facteur est le temps de travail

ind	CONTRIB en %	$\langle x_i, u_1 \rangle$	100xCtr1	CO2 sur u_1	$\langle x_i, u_2 \rangle$	100xCtr2	CO2 sur u_2
"HAUS"	3.106859	1.794638	2.491228	0.370232	0.723763	0.890650	0.060216
"FAUS"	2.457043	0.178964	0.024774	0.004655	2.258784	>>8.674884	0.741615
"FNAU"	>>8.865843	-4.082490	>>12.89169	0.67138	2.246546	>>8.581136	0.203307
"HMUS"	3.309804	1.815789	2.550297	0.355771	0.340057	0.196615	0.012478
"FMUS"	>>4.708323	-2.610169	>>5.269838	0.516788	2.290990	>>8.924019	0.398127
"HCUS"	3.264790	1.477940	1.689560	0.238946	1.86019	>>5.883459	0.378534
"FCUS"	>>4.554039	0.477390	0.176281	0.017872	2.854486	>>13.8538	0.639000
"Hawe"	2.976570	1.183639	1.083674	0.168098	-2.345527	>>9.353946	0.660095
"FAWE"	1.781394	-0.252436	0.049290	0.012775	-1.41547	3.406573	0.401685
"FNAW"	>>8.013288	-4.316941	>>14.41491	0.830583	-1.615072	>>4.435045	0.116256
"HMWE"	3.2778311	1.140709	1.006490	0.141756	-2.431353	>>10.05102	0.644003
"FMWE"	>>5.40271	-3.22354	>>8.037596	0.686904	-1.924236	>>6.295507	0.244763
"HCWE"	3.505462	1.336222	1.381075	0.181909	-2.608805	>>11.57170	0.693394
"FCWE"	3.060838	-1.214924	1.141716	0.172226	-1.680630	>>4.802401	0.329568
"HAYO"	1.986300	2.149457	>>3.573694	0.830720	-0.273969	0.127620	0.013495
"FAYO"	3.026985	1.091545	0.921602	0.140577	0.267899	0.122027	0.008467
"FNAY"	>>5.89359	-3.550900	>>9.75296	0.764079	-0.410628	0.286689	0.010217
"HMYO"	1.967113	2.213862	>>3.791064	0.889844	-0.236673	0.095238	0.010169
"FMYO"	2.430255	-1.491689	1.721140	0.326999	-0.172835	0.050790	0.004389
"HCYO"	>>4.444169	2.046479	3.239475	0.336562	0.439535	0.328474	0.015525
"FCYO"	1.302245	0.296704	0.068093	0.024143	0.345669	0.203159	0.032769
"HAES"	1.84872	2.155009	>>3.592181	0.897156	-0.068665	0.008016	0.000910
"FAES"	2.742356	1.076874	0.896994	0.151024	0.678160	0.781949	0.059893
"FNAE"	>>6.256073	-3.948083	>>12.05680	0.889841	0.007241	0.000089	0.000003
"HMES"	2.115181	2.113334	3.45458	0.754103	-0.127808	0.027773	0.002758
"FMES"	2.350938	-0.353024	0.096398	0.018932	0.230935	0.090676	0.008101
"HCES"	>>4.132679	2.445154	>>4.624581	0.516681	0.017493	0.000520	0.000026
"FCES"	>>4.132679	2.445154	>>4.624581	0.516681	0.017493	0.000520	0.000026

TABLE 2.3 – Indicateurs de contribution et de qualité des représentations des individus sur les deux premiers facteurs

ménager. Il ne faut pas oublier que le facteur est attaché à un vecteur propre qui donne une direction privilégiée mais pas de sens. Ainsi, il est tout à fait justifié de présenter le premier facteur comme lié au temps accordé à ce qui touche l'activité professionnelle. Ces faits observés dans l'espace des variables se manifestent également dans l'espace des individus. La figure 2.10b met en regard les travailleurs à gauche et les inactifs à droite. Ainsi, ceux qui travaillent professionnellement consacrent peu de temps au ménage et à leur enfants et réciproquement.

- Le second facteur oppose les temps passés aux courses et à la toilette à ceux utilisés pour les repas et le sommeil (Cf. figure 2.10a). L'activité de toilette est particulièrement bien représentée par ce facteur. En revanche les durées pour les courses, le sommeil et les repas dépendent également du premier facteur, et ce avec une corrélation positive. Les projections des individus montrent que les USA préfèrent passer du temps dans la salle de bain et les centres commerciaux tandis que les pays de l'ouest favorisent la cuisine et le lit. Les individus des pays de l'est ont de faibles composantes sur ce facteur, i.e. ont des valeurs proches de la moyenne mondiale pour ce facteur. Ce second facteur peut être compris comme une mesure des attitudes casanières.
- Le troisième facteur est fortement lié à la variable TELE, bien que le temps passé devant la télé ne dépend pas seulement de ce facteur (Cf. figure 2.11a). Ce facteur dépend également dans une moindre mesure des variables liées aux activités domestiques. Il structure la population avec d'un côté les pays occidentaux qui sont friands de télévision et de l'autre les pays de l'est et la Yougoslavie qui la regardent peu.
- Le quatrième facteur oppose la télévision et les loisirs. La télévision a toutefois une corrélation inférieure à 0.4. Les célibataires sont les plus enclins à avoir des loisirs variés et différents de celui proposé par la télévision. Cette tendance est inversée chez les couples mariés et qui ont des enfants : ils s'accordent peu de moments de loisirs autres que la télévision.

ind	CONTRIB en %	$\langle x_i, u_3 \rangle$	100xCtr	CO2 sur u	$\langle x_i, u_4 \rangle$	100xCtr	CO2 sur u
"HAUS"	3.106859	-1.704220	>>7.781844	0.333865	-0.937423	2.686345	0.101016
"FAUS"	2.457043	-0.589614	0.931467	0.050531	-0.212729	0.138339	0.006577
"FNAU"	>>8.865843	-1.127682	3.407254	0.051226	0.168674	0.086974	0.001146
"HMUS"	3.309804	-1.679779	>>7.560233	0.304469	-1.119829	>>3.833490	0.135314
"FMUS"	>>4.708323	-0.731780	1.434804	0.040619	-0.318161	0.309446	0.007678
"HCUS"	3.264790	-1.51492	>>6.149104	0.251054	0.554235	0.939029	0.033602
"FCUS"	>>4.554039	-1.287956	>>4.444604	0.130091	-0.014454	0.000638	0.000016
"Hawe"	2.976570	-1.09109	3.189738	0.142840	-0.339244	0.351816	0.013808
"FAWE"	1.781394	-0.018495	0.000916	0.000068	-1.107415	>>3.748965	0.245868
"FNAW"	>>8.013288	-0.852289	1.946279	0.032374	-0.374472	0.428677	0.006249
"HMWE"	3.278311	-1.207246	>>3.905014	0.158775	-0.590454	1.06576	0.037980
"FMWE"	>>5.40271	-0.612715	1.005884	0.024816	-0.002841	0.000024	5.338E-0
"HCWE"	3.505462	-0.741447	1.472963	0.056009	0.719939	1.584463	0.052806
"FCWE"	3.060838	-0.821435	1.807915	0.078731	1.170865	>>4.190869	0.159961
"HAYO"	1.986300	0.610875	0.999852	0.067096	0.422438	0.545528	0.032086
"FAYO"	3.026985	2.056502	>>11.33155	0.498988	-1.698519	>>8.819246	0.340386
"FNAY"	>>5.89359	1.508887	>>6.100205	0.137967	0.692377	1.465468	0.029050
"HMYO"	1.967113	0.423213	0.479898	0.032518	0.233534	0.166721	0.009901
"FMYO"	2.430255	1.862032	>>9.289781	0.509524	-0.876879	2.350552	0.112997
"HCYO"	>>4.444169	1.042643	2.91274	0.087362	2.548693	>>19.85754	0.522019
"FCYO"	1.302245	1.208476	>>3.912973	0.400520	1.348387	>>5.558020	0.498630
"HAES"	1.84872	0.180028	0.086838	0.006261	-0.250965	0.192539	0.012167
"FAES"	2.742356	1.798010	>>8.66193	0.421019	-1.644242	>>8.264611	0.352087
"FNAE"	>>6.256073	0.478212	0.612733	0.013055	0.905125	2.504427	0.046769
"HMES"	2.115181	-0.243721	0.159153	0.010029	-0.855892	2.239386	0.123689
"FMES"	2.350938	1.725615	>>7.978459	0.452365	-1.781476	>>9.701765	0.482126
"HCES"	>>4.132679	0.553618	0.821205	0.026486	2.208581	>>14.91136	0.421538
"FCES"	>>4.132679	0.553618	0.821205	0.026486	2.208581	>>14.91136	0.421538

TABLE 2.4 – Indicateurs de contribution et de qualité des représentations des individus sur les deux derniers facteurs

2.8 Davantage d'exemples

A titre d'exemple, les rapports en ligne indiqués ci-dessous pourront être étudiés :

- l'analyse budget/temps ;
- l'analyse portant sur la consommation de drogues (données non-réduites, données réduites, données exprimées pour 10000 habitants) ;
- ...

Chapitre 3

L'analyse factorielle des correspondances

3.1 L'AFC pour rechercher des liens dans une paire de variables qualitatives

À partir de la playlist d'un inconnu, pouvez-vous deviner son sexe ? Sa CSP ? Ses inclinations politiques ? Quelle offre de restauration serait adaptée au public d'un projet de pôle d'activités ? Existe-t-il une influence des pratiques sportives sur la productivité au travail ? La possession d'un trèfle à quatre feuilles et les chances de gain à la loterie sont-elles indépendantes ?

	Classique	Pop	Hip Hop	Rock	Altern.	Jazz/Blues	Electr.	Country	Autre
homme	1	0	15	9	2	1	5	5	5
femme	1	7	11	9	6	0	0	6	3

TABLE 3.1 – Genres musicaux présents dans les playlist des étudiants d'une classe croisés avec le sexe des élèves

CSP	Aucun	Un	Deux	Plus de deux
Agriculteurs	23,1	44,8	26,1	6,0
Patrons de l'industrie et du commerce	12,7	44,4	24,4	18,5
Cadres supérieurs	2,4	43,0	31,6	23,1
Professions intermédiaires	2,2	45,2	33,8	18,9
Employés	6,1	53,6	26,5	13,8
Ouvriers	6,9	55,1	26,4	11,6
Étudiants	0,4	40,8	30,9	28,0
Retraités	32,4	38,4	19,1	10,1
Autres inactifs	17,4	45,0	26,0	11,6

TABLE 3.2 – Profils lignes des genres musicaux cités selon les CSPs¹

Les réponses à ces questions peuvent être approchées à travers la recherche de liens dans des paires de variables qualitatives. Ces liens s'expriment au niveau des couples de modalités de ces variables : certains couples de modalités peuvent être extrêmement et étonnamment rares, tandis que d'autres sont observés à un niveau d'abondance dépassant les attentes. L'AFC définit un cadre dans lequel ce type d'analyse est réalisé. En reprenant les éléments de l'ACP, elle fait intervenir la distance du χ^2 ², parfaitement adaptée pour évaluer l'hypothèse d'indépendance entre variables qualitatives et entrer dans le détail des liaisons le cas échéant.

1. Données partielles tirées de «La stratification sociale des goûts musicaux», P. Coulangeon, 2003 (<https://www.cairn.info/revue-francaise-de-sociologie-1-2003-1-page-3.htm>).

2. Se reporter à la présentation de Julien Barnier pour une présentation du χ^2 (<http://alea.fr.eu.org/pages/khi2>).

	Perdant	Gagnant
Trèfle à 4 feuilles	220	7
Fer à cheval	200	1
Rien	200	1

TABLE 3.3 – Issue à la loterie selon objet fétiche

3.2 L'analyse de tableaux de contingence

Considérons deux variables qualitatives notées V et W . Nous nous intéressons à l'existence de relations entre les modalités de ces deux variables. La variable V (resp. W) a n (resp. p) modalités notées v_1, \dots, v_n (resp. w_1, \dots, w_p). Ces deux variables sont mesurées sur un échantillon de N individus. Nous pouvons construire à partir des données les effectifs obtenus pour chaque modalité de V , et de même pour W .

Il est également intéressant de connaître les effectifs obtenus pour chaque couple de modalités (v_i, w_j) . Cette information permet de construire le tableau suivant, dit tableau de contingence croisant en ligne les modalités de V avec les modalités de W en colonnes :

		w_1	\cdots	w_j	\cdots	w_p
V	v_1					
	v_i			\cdots	n_{ij}	\cdots
	v_n				\vdots	

 TABLE 3.4 – Les valeurs de V et W mesurées sur une population - forme brute et tableau de contingence

Les marges de ce tableau sont calculées, c'est à dire les effectifs des modalités données en lignes et en colonnes.

		w_1	\cdots	w_j	\cdots	w_p
V	v_1					
	v_i		\cdots	n_{ij}	\cdots	$n_{i.}$
	v_n			\vdots		
		\cdots	$n_{.j}$	\cdots		N

TABLE 3.5 – Un tableau de contingence et table de fréquences associée

Les notations utilisées ci-dessus sont les suivantes : $\sum_{i=1}^n n_{ij} = n_{.j}$ et $\sum_{j=1}^p n_{ij} = n_{i.}$. Nous avons évidemment

$$\sum_{i=1}^n \sum_{j=1}^p n_{ij} = \sum_{i=1}^n n_{i.} = \sum_{j=1}^p n_{.j} = N$$

À partir du tableau de contingence, le tableau des fréquences T de couples de modalités peut être construit par simple division des cases par l'effectif total de l'échantillon.

Comme précédemment, nous posons $f_{\cdot i} = \sum_j f_{ij}$ et $f_{\cdot j} = \sum_i f_{ij}$. Observons que les marges de T ainsi calculées correspondent aux distributions en fréquences respectives des modalités de V et W .

Deux autres tableaux présentent un intérêt tout particulier : le tableau des profils lignes et le tableau des profils colonnes. Le premier permet de considérer pour chaque modalité de V la répartition des modalités de W , cette répartition étant considérée comme une distribution. L'élément de base du tableau des profils lignes est $\frac{f_{ij}}{f_{\cdot i}}$ (ou 0 si $f_{\cdot i} = 0$). Ainsi, la somme des fréquences d'une ligne égale l'unité (sous l'hypothèse que toute modalité est observée au moins une fois). De même, l'élément de base du tableau des profils colonnes est $\frac{f_{ij}}{f_{\cdot j}}$ (ou 0 si $f_{\cdot j} = 0$). La somme en colonne est égale à un pour chaque modalité de W . Pour simplifier nos notations, nous supposerons à partir de maintenant que chaque modalité de chaque variable a été observée au moins une fois.

Notons Z la matrice des profils lignes :

$$Z = \begin{pmatrix} & \vdots & \\ \dots & \frac{f_{ij}}{f_{\cdot i}} & \dots \\ & \vdots & \end{pmatrix}$$

Notre but ici est d'appliquer l'ACP au nuage de profils lignes (nous verrons plus loin que l'ACP du nuage de profils colonnes est le problème dual du premier). Pour cela, il nous faut préciser la distance adoptée dans l'espace des individus et les pondérations attachées aux profils lignes.

3.2.1 Pondération et centre de gravité

Chaque profil ligne est associé à une modalité de V et chaque modalité v_i apparaît à la fréquence $f_{\cdot i}$. Dans la suite, nous associerons à la i^e ligne le poids $f_{\cdot i}$ et qui ce qui suit justifie ce choix.

Notons D la matrice de pondération associée aux profils lignes :

$$D = \begin{pmatrix} \ddots & & \\ & f_{\cdot i} & \\ & & \ddots \end{pmatrix}$$

Notons au passage que $Z = D^{-1}T$. Le système de pondération adopté conduit au centre de gravité suivant pour les profils lignes :

$$\left(\dots \ f_{\cdot i} \ \dots \right) \times \begin{pmatrix} & \vdots & \\ \dots & \frac{f_{ij}}{f_{\cdot i}} & \dots \\ & \vdots & \end{pmatrix} = \left(\dots \ f_{\cdot j} \ \dots \right)$$

Ainsi, le profil ligne moyen est la distribution des modalités de W . Notons Y la matrice des profils lignes centrés :

$$Y = \begin{pmatrix} & \vdots & \\ \dots & \frac{f_{ij}}{f_{\cdot i}} - f_{\cdot j} & \dots \\ & \vdots & \end{pmatrix} = \begin{pmatrix} & \vdots & \\ \dots & \frac{f_{ij} - f_{\cdot i}f_{\cdot j}}{f_{\cdot i}} & \dots \\ & \vdots & \end{pmatrix}$$

3.2.2 Distance

Etant donné que les individus sont des distributions, il est assez naturel de travailler avec la distance du χ^2 . Prenant en compte que le centre de gravité des profils lignes est la distribution des modalités de W , la distance du χ^2 centrée sur cette distribution semble toute indiquée³.

3. Nous renvoyons à l'annexe E pour quelques rappels sur la distance du χ^2 .

Ainsi, le carré de la distance entre les i^e et k^e profils lignes est

$$\sum_{j=1}^p \frac{\left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{kj}}{f_{k\cdot}}\right)^2}{f_{\cdot j}}$$

Il est ais  d'observer que cette distance est d riv e d'un produit scalaire d fini par la matrice M suivante :

$$M = \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{\cdot j}} & \\ & & \ddots \end{pmatrix}$$

Equivalence distributionnelle

Remarquons que si deux modalit s v_i et $v_{i'}$ de V ont des distributions en fr quences identiques (i.e. des profils lignes identiques), alors elles sont repr sent es par des points dans \mathbb{R}^p qui sont confondus et donc sont   m me distance des autres points :

$$\forall k, d(i, k) = d(i', k)$$

La modalit  artificielle " v_i ou $v_{i'}$ " est aussi   la m me distance des autres modalit s et l'analyse peut se faire en rempla ant les deux modalit s v_i et $v_{i'}$ par la modalit  artificielle pond r e par la somme $f_{i\cdot} + f_{i'\cdot}$. Cette agr gation est remarquable par le fait qu'elle ne modifie ni les distances entre profils-lignes, ni les distances entre profils colonnes dans le probl me dual ! Cette propri t  est d nomm e  l' quivalence distributionnelle.

Cette remarque est d'un grand int  et sur le plan pratique :

- elle permet d'op rer des regroupements permettant de simplifier les interpr tations ;
- ces regroupements sont aussi r alisables pour des modalit s de profils tr s proches, car l'agr gation ne modifie que peu les r sultats sans en changer la nature ;
- lorsque des modalit s sont ordonn es, il semble naturel de fusionner ensemble des modalit s qui sont proches par rapport   leur relation d'ordre.

Nous montrerons en annexe E que la fusion de deux modalit s conduit toujours   une perte d'inertie positive.

Conditions sur les effectifs th oriques

L'utilisation des r sultats du χ^2 est d licate en pr sence d'effectifs th oriques faibles. Un crit re commun ement admis est de v rifier que toutes les cellules du tableau des effectifs th oriques ont une valeur sup rieure   1 et que 80% des cellules ont des valeurs sup rieures   5 (crit re de Cochran).

Dans le cas o  cette condition n'est pas satisfaite, il est possible de rechercher une fusion de modalit s qui permettrait de lever cette «infraction». Plusieurs fa ons de proc der sont possibles : fusion de modalit s rares avec des modalit s dont le profil est proche (avec l'inconv nient de la dissolution des individus atypiques), fusion de modalit s rares ensemble . . . le choix peut  tre compliqu  car une fusion peut modifier sensiblement les r sultats de l'analyse. D'autres voies existent pour faire face au probl me des petits effectifs.

Expression de l'inertie

Nous disposons de tous les ´l ments pour exprimer l'inertie en fonction des particularit s de notre cadre. Reprenons l'expression 2.4 de la page 14 :

$$\begin{aligned}
 I_g &= \sum_i p_i \|x_i - g\|^2 \\
 &= \sum_i f_i \left(\cdots \frac{f_{ij} - f_{i \cdot} f_{\cdot j}}{f_{i \cdot}} \cdots \right) \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{\cdot j}} & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ \frac{f_{ij} - f_{i \cdot} f_{\cdot j}}{f_{i \cdot}} \\ \vdots \end{pmatrix} \\
 &= \sum_i \left(\cdots \frac{f_{ij} - f_{i \cdot} f_{\cdot j}}{f_{i \cdot} f_{\cdot j}} \cdots \right) \begin{pmatrix} \vdots \\ \frac{f_{ij} - f_{i \cdot} f_{\cdot j}}{f_{i \cdot} f_{\cdot j}} \\ \vdots \end{pmatrix} \\
 I_g &= \sum_i \sum_j \frac{(f_{ij} - f_{i \cdot} f_{\cdot j})^2}{f_{i \cdot} f_{\cdot j}}
 \end{aligned} \tag{3.1}$$

Il est important de remarquer pour la suite que cette expression est directement liée à la mesure de l'écart à l'indépendance entre les deux variables V et W. L'expression de cet écart, détaillé en annexe E, est la suivante :

$$d^2 = N \sum_i \sum_j \frac{(f_{ij} - f_{i \cdot} f_{\cdot j})^2}{f_{i \cdot} f_{\cdot j}}$$

Ainsi, nous avons $I_g = \frac{d^2}{N}$, i.e. l'inertie du nuage des individus par rapport au centre de gravité est le N^e de la distance du χ^2 à $(p-1)(n-1)$ degrés de liberté entre W et V . Nous verrons plus loin les conséquences de ce lien.

3.3 ACP des profils lignes

Tous les éléments nécessaires à la formulation du problème de l'ACP étant identifiés, nous pouvons poser le problème. Exprimons le problème de la recherche du sous-espace de dimension 1 maximisant l'inertie du nuage projeté :

$$\max_{\|u\|=1} u' M Y' D Y M u \tag{3.2}$$

où $\|u\|^2 = u' M u$

Nous savons que la solution à ce problème est donnée par un vecteur propre normé de la matrice $Y' D Y M$ et associé à la plus grande valeur propre. Notons u_1 un tel vecteur et notons λ_1 la valeur propre associée. u_1 vérifie :

$$\begin{aligned}
 Y' D Y M u_1 &= \lambda_1 u_1 \\
 u_1' M u_1 &= 1
 \end{aligned}$$

La composante principale associée à ce vecteur est notée φ_1 :

$$\varphi_1 = Y M u_1$$

D'après les liens existant entre ce problème et le problème dual associé, nous savons que :

$$YMY'D\varphi_1 = YMY'DYM u_1 = YM(\lambda_1 u_1) = \lambda_1 YM u_1 = \lambda_1 \varphi_1$$

Comme $\|\varphi_1\|^2 = \varphi_1'D\varphi_1 = \lambda_1$, nous en déduisons que le vecteur $\frac{1}{\sqrt{\lambda_1}} YM u_1$ est solution du problème dual

$$\max_{\|v\|=1} v'DYMY'Dv \quad \text{i.e.} \quad \max_{v'Dv=1} v'DYMY'Dv \quad (3.3)$$

Quel lien existe-t-il entre le problème dual défini dans l'espace des «variables» et l'espace des profils colonnes ? L'expression du problème de l'ACP appliquée aux profils colonnes va nous permettre d'identifier ce lien.

3.4 ACP des profils colonnes

Nous nous plaçons ici dans l'espace des profils colonnes. Dans ce nouveau cadre de travail, les matrices importantes sont les suivantes :

$$\tilde{Z} = \begin{pmatrix} & \vdots & \\ \dots & \frac{f_{ij}}{f_{.j}} & \dots \\ & \vdots & \end{pmatrix} \quad \tilde{D} = \begin{pmatrix} \ddots & & \\ & f_{.j} & \\ & & \ddots \end{pmatrix} \quad \tilde{M} = \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{i.}} & \\ & & \ddots \end{pmatrix}$$

$$\tilde{Y} = \begin{pmatrix} & \vdots & \\ \dots & \frac{f_{ij}}{f_{.j}} - f_{i.} & \dots \\ & \vdots & \end{pmatrix} = \begin{pmatrix} & \vdots & \\ \dots & \frac{f_{ij} - f_{i.}f_{.j}}{f_{.j}} & \dots \\ & \vdots & \end{pmatrix}$$

où, à présent, j indexe les lignes et i indexe les colonnes (\otimes) ! Dans l'espace des profils colonnes, le système de pondération est donné par la distribution des modalités de W . Le centre de gravité des profils colonnes est donné par la distribution des modalités de V . Enfin, la distance utilisée est la distance du χ^2 centrée sur le centre de gravité du nuage des profils colonnes.

Notons que

$$\tilde{Z} = \tilde{D}^{-1}T'$$

Le problème de la recherche du sous-espace de dimension 1 maximisant l'inertie du nuage projeté s'exprime comme suit :

$$\max_{\|v\|=1} v'\tilde{M}\tilde{Y}'\tilde{D}\tilde{Y}\tilde{M}v$$

où $\|v\|^2 = v'\tilde{M}v$.

Nous pouvons exprimer les matrices \tilde{M} , \tilde{Y} et \tilde{D} en fonction des matrices intervenant dans l'espace des profils lignes :

$$\tilde{D} = M^{-1} \quad \tilde{M} = D^{-1} \quad \tilde{Y} = (DYM)' = MY'D$$

Ceci nous permet de donner une expression au problème dans l'espace des profils colonnes en utilisant les objets de l'espace des profils lignes.

$$\begin{aligned} \max_{v'\tilde{M}v=1} v'\tilde{M}\tilde{Y}'\tilde{D}\tilde{Y}\tilde{M}v &= \max_{v'D^{-1}v=1} v'D^{-1}DYMM^{-1}MY'DD^{-1}v \\ &= \max_{v'D^{-1}v=1} v'D^{-1}DYMY'DD^{-1}v \end{aligned}$$

Posons $w = D^{-1}v$, i.e. $v = Dw$. L'équation $v'D^{-1}v = 1$ s'écrit $w'DD^{-1}Dw = w'Dw = 1$, le problème s'écrit

$$\max_{w'Dw=1} w'DYMY'Dw$$

Nous retrouvons l'expression du problème dual 3.3. Par conséquent les problèmes de l'ACP des profils lignes et des profils colonnes sont duals. Connaissant un vecteur propre normé w_1 associé à la plus grande valeur propre λ_1 de la matrice $YMY'D$, nous en déduisons un vecteur propre solution de l'ACP dans l'espace des profils colonnes. En effet, $v_1 = Dw_1$ est un tel vecteur. La composante principale associée à v_1 , notée ψ_1 , est donc

$$\begin{aligned}\psi_1 &= \tilde{Y}\tilde{M}v_1 \\ &= MY'DD^{-1}Dw_1 \\ \psi_1 &= MY'D\frac{1}{\sqrt{\lambda_1}}\varphi_1\end{aligned}$$

soit, en poursuivant le développement

$$\begin{aligned}\psi_1 &= \frac{1}{\sqrt{\lambda_1}}MY'DYMu_1 \\ &= \frac{1}{\sqrt{\lambda_1}}M\lambda_1 u_1 \\ \psi_1 &= \sqrt{\lambda_1}Mu_1\end{aligned}$$

Ainsi les relations entre ACP des profils lignes et ACP des profils colonnes sont précisées. Avant de passer à la présentation d'un modèle d'interprétation, nous devons souligner l'existence de relations justifiant la simultanéité des représentations des projections des profils lignes et des profils colonnes qui sont une particularité de l'AFC.

3.5 Exploitation des relations de dualité

3.5.1 Relations quasi-barycentriques

À partir des relations ci-dessus, nous obtenons les relations dites quasi-barycentriques suivantes :

$$\psi_{1j} = \frac{1}{\sqrt{\lambda_1}} \sum_{i=1}^n \frac{f_{ij}}{f_{.j}} \varphi_{1i} \quad \text{i.e.} \quad \psi_1 = \frac{1}{\sqrt{\lambda_1}} \left(\frac{f_{ij}}{f_{.j}} \right)_{ji} \varphi_1 \quad \text{ou} \quad \psi_1 = \frac{1}{\sqrt{\lambda_1}} Z' \varphi_1 \quad (3.4)$$

$$\varphi_{1i} = \frac{1}{\sqrt{\lambda_1}} \sum_{j=1}^p \frac{f_{ij}}{f_{.i}} \psi_{1j} \quad \text{i.e.} \quad \varphi_1 = \frac{1}{\sqrt{\lambda_1}} \left(\frac{f_{ij}}{f_{.i}} \right)_{ji} \psi_1 \quad \text{ou} \quad \varphi_1 = \frac{1}{\sqrt{\lambda_1}} Z \psi_1 \quad (3.5)$$

Considérons 3.4 par exemple. La somme sur i des $\frac{f_{ij}}{f_{.i}}$ égalant l'unité, la quantité ψ_{1j} peut être considérée comme le barycentre des φ_{1i} , au coefficient $\frac{1}{\sqrt{\lambda_1}}$ près⁴. Ce lien justifie une représentation simultanée des projections des profils lignes et des profils colonnes : en effet cela permet d'observer les positionnements relatifs des projections, sans pour autant permettre d'utiliser l'apparente distance

4. d'où le quasi avant le qualificatif barycentrique

entre un point ligne et un point colonne. Nous verrons plus loin à quel type d'interprétation cela conduit.

Démontrons rapidement la première relation. Tout d'abord rappelons que prémultiplier une matrice par une matrice diagonale revient à multiplier la i^e ligne de la première par le i^e coefficient de la diagonale, et que postmultiplier une matrice par une matrice diagonale revient à multiplier la i^e colonne de la première par le i^e coefficient de la diagonale. Partant des éléments suivants :

- $\psi_1 = MY'Dw_1$ où w_1 est solution de 3.3
 - $\frac{1}{\sqrt{\lambda_1}}\varphi_1$ est solution de 3.3
- nous pouvons écrire :

$$\begin{aligned}
 \psi_1 &= \frac{1}{\sqrt{\lambda_1}} MY'D\varphi_1 \\
 &= \frac{1}{\sqrt{\lambda_1}} \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{.j}} & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \dots & \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}} & \dots \\ \vdots & & \vdots \end{pmatrix}' \begin{pmatrix} \ddots & & \\ & f_{i.} & \\ & & \ddots \end{pmatrix} \varphi_1 \\
 &= \frac{1}{\sqrt{\lambda_1}} \begin{pmatrix} \dots & \frac{f_{ij} - f_{i.}f_{.j}}{f_{.j}} & \dots \\ \vdots & & \vdots \end{pmatrix}' \varphi_1 \\
 \psi_1 &= \frac{1}{\sqrt{\lambda_1}} \begin{pmatrix} \vdots & & \\ \dots & \frac{f_{ij}}{f_{.j}} & \dots \\ \vdots & & \vdots \end{pmatrix}' \varphi_1 - \underbrace{\begin{pmatrix} \dots & \vdots & \dots \\ \dots & f_{i.} & \dots \\ \vdots & & \vdots \end{pmatrix}'}_{=0 \text{ dans } \mathbb{R}^p} \varphi_1
 \end{aligned}$$

En effet, φ_1 est centrée comme combinaison linéaire de colonnes centrées par rapport à la distribution moyenne⁵.

D'où

$$\psi_1 = \frac{1}{\sqrt{\lambda_1}} \begin{pmatrix} \vdots & & \\ \dots & \frac{f_{ij}}{f_{.j}} & \dots \\ \vdots & & \vdots \end{pmatrix}' \varphi_1 \text{ i.e. } \psi_{1j} = \frac{1}{\sqrt{\lambda_1}} \sum_{i=1}^n \frac{f_{ij}}{f_{.j}} \varphi_{1i}$$

La relation 3.5 se démontre de façon similaire.

3.5.2 Les valeurs propres sont bornées par 1

Les relations barycentriques permettent également de prouver que toute valeur propre cherchée est inférieure à l'unité. Vérifions cela pour λ_1

5. formulé ainsi, la justification peut sembler à la fois triviale et floue. Pour lever l'éventuel flou, il suffit de se rappeler qu'ici l'opération de centrage et le calcul d'élément moyen font intervenir les poids fournis par les coefficients de la distribution des modalités de V .

$$\sqrt{\lambda_1} \psi_{1j} = \sum_{i=1}^n \frac{f_{ij}}{f_{\cdot j}} \varphi_{1i} \leq \max \varphi_{1i}$$

d'où $\sqrt{\lambda_1} \max \psi_{1j} \leq \max \varphi_{1i}$

et

$$\begin{aligned} \sqrt{\lambda_1} \varphi_{1i} &= \sum_{j=1}^p \frac{f_{ij}}{f_{i \cdot}} \psi_{1j} \leq \max \psi_{1j} \\ \sqrt{\lambda_1} \max \varphi_{1i} &\leq \max \psi_{1j} \\ \text{i.e. } \lambda_1 \max \varphi_{1i} &\leq \sqrt{\lambda_1} \max \psi_{1j} \end{aligned}$$

d'où

$$\lambda_1 \max \varphi_{1i} \leq \max \varphi_{1i}$$

ce qui implique, en tenant compte du fait que $\lambda_1 \geq 0$ et que $\max \varphi_{1i} > 0$ (car $\sum_i f_{i \cdot} \varphi_{1i} = 0$), que $\lambda_1 \leq 1$.

Une valeur propre égale à l'unité correspond à une dépendance fonctionnelle, ce qui signifie qu'il est possible d'organiser le tableau de contingence avec une structure diagonale par blocs, avec des blocs diagonaux de taille variable (et éventuellement rectangulaires).

3.5.3 Une borne pour la mesure du χ^2

Pour finir, soulignons que le rang de la table de contingence vaut au plus $\min(n - 1, p - 1)$, ce qui implique qu'il y a au plus $\min(n - 1, p - 1)$ valeurs propres non nulles et d'après la borne sur les valeurs propres, on dispose de la borne suivante :

$$\chi^2 \leq N \times \min(n - 1, p - 1)$$

Cette borne est réalisable (Cf. annexe E) comme illustré ci-après. Supposons que $n \leq p$. La table suivante permet de réaliser la valeur de la borne :

$$\begin{matrix} 1 & \left(\begin{array}{ccccccccc} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots & \vdots & & \vdots \\ \vdots & 0 & \ddots & 0 & \vdots & & \vdots \\ \vdots & & \ddots & 1 & 0 & \cdots & 0 \\ n-1 & & & 0 & 1 & \cdots & 1 \\ n & \left(\begin{array}{ccccccccc} 0 & \cdots & \cdots & 0 & 1 & \cdots & 1 \end{array} \right) \end{array} \right) \end{matrix}$$

TABLE 3.6 – Cas de dépendance maximale

Les $(n - 1)$ premières modalités de V sont en bijection avec les $(n - 1)$ premières modalités de W , et la n^e modalité de V peut être associée à n'importe laquelle des $p - (n - 1)$ dernières modalités de W . Par exemple, pour $n = 5$ et $p = 10$, $\chi^2 = 40$, ce qui correspond bien à $10 \times \min(4, 9)$.

3.6 L'analyse des profils lignes non-centrés

L'AFC se distingue de l'ACP par une propriété particulière : pour une AFC, l'analyse des profils lignes centrés est équivalente à l'analyse des profils lignes non centrés si dans cette dernière nous retirons l'axe factoriel donné par le centre de gravité des profils lignes !

Pour cela, il suffit de considérer le problème défini sur les données non-centrées, de vérifier que le centre de gravité est bien un vecteur propre normé et de vérifier que les autres vecteurs propres sont vecteurs propres dans le premier problème :

- Dans le problème 3.2, nous remplaçons la matrice des profils centrés par la matrice des profils lignes initiaux : plus précisément, Z se substitue à Y . Rappelons que nous avons $Z = D^{-1}T$ et

$$Y = Z - \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} (\cdots \ f_{.j} \ \cdots). \text{ Le problème caractérisant l'analyse non centrée est :}$$

$$\max_{\|u\|=1} u' M Z' D Z M u \quad (3.6)$$

Ainsi, les éléments propres cherchés sont ceux de la matrice $Z' D Z M$ i.e. $T' D^{-1} T M$.

- Vérifions que le centre de gravité des profils lignes est un vecteur propre de $T' D^{-1} T M$.

$$\begin{aligned} T' D^{-1} T M \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix} &= \begin{pmatrix} \vdots & \vdots & \cdots \\ \cdots & f_{ij} & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix}' \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{i.}} & \\ & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \vdots & \vdots & \cdots \\ \cdots & f_{ij} & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix}' \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{.j}} & \\ & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix} \\ &= \begin{pmatrix} \vdots & \vdots & \cdots \\ \cdots & f_{ij} & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix}' \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{i.}} & \\ & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \vdots & \vdots & \cdots \\ \cdots & f_{ij} & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} \\ &= \begin{pmatrix} \vdots & \vdots & \cdots \\ \cdots & f_{ij} & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix}' \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{i.}} & \\ & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ f_{i.} \\ \vdots \end{pmatrix} \\ &= \begin{pmatrix} \vdots & \vdots & \cdots \\ \cdots & f_{ij} & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix}' \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} \\ T' D^{-1} T M \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix} &= \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix} \end{aligned}$$

Ainsi, le centre de gravité est bien un vecteur propre de la matrice $Z' D Z M$. Remarquons que ce vecteur est de norme 1 :

$$\begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix}' M \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix}' \begin{pmatrix} \ddots & & \\ & \frac{1}{f_{.j}} & \\ & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix}' \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} = 1$$

Par conséquent, d'après ce qui précède, la valeur propre associée est 1.

- Considérons à présent u un vecteur propre de la matrice $Z' D Z M$, distinct du centre de gravité, M -normé et associé à la valeur propre λ .

$$Y' D Y M u = \left[Z - \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} (\cdots \ f_{.j} \ \cdots) \right]' D \left[Z - \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} (\cdots \ f_{.j} \ \cdots) \right] M u$$

$$\begin{aligned}
 &= \left[Z - \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} (\cdots \ f_{.j} \ \cdots) \right]' DZMu = 0 \text{ car } (\cdots \ f_{.j} \ \cdots) Mu = 0 \\
 &= Z'DZMu - \left[\begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} (\cdots \ f_{.j} \ \cdots) \right]' DZMu \\
 &= \lambda u - \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix} \underbrace{(\cdots \ 1 \ \cdots)}_T \underbrace{DZMu}_{(\cdots \ f_{.j} \ \cdots)} \\
 &\quad 0 \\
 Y'DYMu &= \lambda u
 \end{aligned}$$

Ainsi, (λ, u) est un couple d'éléments propres de la matrice $Y'DYM$.

Nous venons de prouver que les couples d'éléments propres de la matrice $Z'DZM$ sont des couples d'éléments propres de la matrice $Y'DYM$, pour tout u distinct du centre de gravité du nuage de profils lignes. Ceci nous renseigne sur $p - 1$ couples d'éléments propres. Par ailleurs, il est remarquable que les colonnes de la matrice des profils lignes sont liées par une relation du type

$$\sum_j^p Y_j = Cte$$

Par suite, il ne peut y avoir au plus que $p - 1$ facteurs permettant de décomposer l'information du nuage de profils lignes. Il est facile de vérifier que le centre de gravité est un vecteur propre associé à la valeur propre nulle pour la matrice $Y'DYM$.

3.7 Interprétation des résultats

3.7.1 Analyse des écarts à la situation d'indépendance entre V et W

D'après l'expression 3.1, l'inertie totale est liée au test d'indépendance du χ^2 entre V et W par la relation :

$$d^2 = NI_g$$

Ainsi, NI_g permet de savoir si les écarts à l'indépendance sont significatifs. NI_g est comparé à la valeur du χ^2 à $(n - 1)(p - 1)$ ddl pour un seuil d'erreur fixé. Si la valeur de NI_g excède cette valeur, l'hypothèse d'indépendance peut être rejetée (avec le risque de se tromper fixé).

Il est évident que lorsqu'il n'est pas possible de mettre en doute l'indépendance, l'intérêt de la poursuite de l'analyse des relations entre les modalités est remise en question puisque dépourvue de significativité statistique.

Dans le cas où l'écart mesuré est significatif, on peut considérer les **résidus standardisés**, qui sont l'expression de la différence entre effectif observé et effectif théorique, ramenée à l'effectif théorique :

$$d_{ij} = \frac{n_{ij} - \frac{n_{i.}n_{.j}}{N}}{\sqrt{\frac{n_{i.}n_{.j}}{N}}} = \sqrt{N} \times \frac{f_{ij} - f_{i.}f_{.j}}{\sqrt{f_{i.}f_{.j}}} \quad (3.7)$$

Le signe du résidu d_{ij} indique si l'association est positive (les modalités «s'attirent» i.e. le couple de modalités est sur-représenté par rapport à la situation d'indépendance) ou négative (les modalités «s'évitent», i.e. le couple est sous-représenté par rapport à la situation d'indépendance). Un résidu supérieur à deux en valeur absolue tend à indiquer un écart notable à la valeur théorique pour un couple de modalités.

Les graphiques en mosaique constituent une bonne synthèse des écarts intéressants et du sens de ces écarts. Ci-contre, sur le graphique 3.1, les couleurs sont utilisées à la fois pour distinguer les couples rares de ceux très fréquents et pour définir des gradations pour les valeurs des résidus (écarts compris entre 2 et 4 ou supérieurs à 4). Ici, nous observons que les hommes écoutent beaucoup de Hip Hop, alors que les femmes semblent fuir ce genre musical. C'est le contraire, mais avec un écart moins prononcé, pour la musique country. Les couples ainsi mis en évidence pourront être recherchés dans les projections de l'AFC.

Soulignons que la somme des carrés des d_{ij} , pour tous les couples (i, i) , correspond au χ^2 : $\chi^2 = \sum_{i,j} d_{ij}^2$. Il est donc intéressant de considérer le tableau dont l'élément de base est :

$$\frac{\left(n_{ij} - \frac{n_{i \cdot} n_{\cdot j}}{N}\right)^2}{\frac{n_{i \cdot} n_{\cdot j}}{N}} \times \frac{1}{NI_g} = \frac{(f_{ij} - f_{i \cdot} f_{\cdot j})^2}{f_{i \cdot} f_{\cdot j}} \times \frac{1}{I_g}$$

L'élément correspondant au couple (i, j) donne la contribution du couple de la i^{e} modalité de V et de la j^{e} modalité de W au χ^2 .

3.7.2 Choix du nombre d'axe

Poursuivons avec l'inertie. Lors d'une ACP, la composition de l'inertie est analysée à travers les parts relatives des valeurs propres, la part moyenne et divers autres critères. Dans le cadre de l'AFC, il est possible de tirer profit de l'information apportée par le χ^2 .

Critère dérivé du χ^2

Si l'indépendance peut être mise en doute, il est possible d'utiliser conjointement le lien entre le χ^2 et l'inertie et sa décomposition en valeurs propres. Elles sont liées au χ^2 comme suit :

$$\chi^2 = N \times \sum_i \lambda_i$$

Ainsi, $N \times \lambda_i$ correspond à la partie de l'écart à l'indépendance restitué par projection sur le i^{e} axe, elle mesure la significativité de la dépendance des profils projetés sur cet axe.

Pour l'analyse des liens à travers les projections, il semble pertinent de procéder à la sélection des axes permettant de restituer un écart permettant de rejeter l'indépendance.

Nous pouvons de même considérer $N \sum_{i \in K} \lambda_i$ pour connaître le niveau de significativité d'un sous-espace de projection porté par les axes indexé par K .

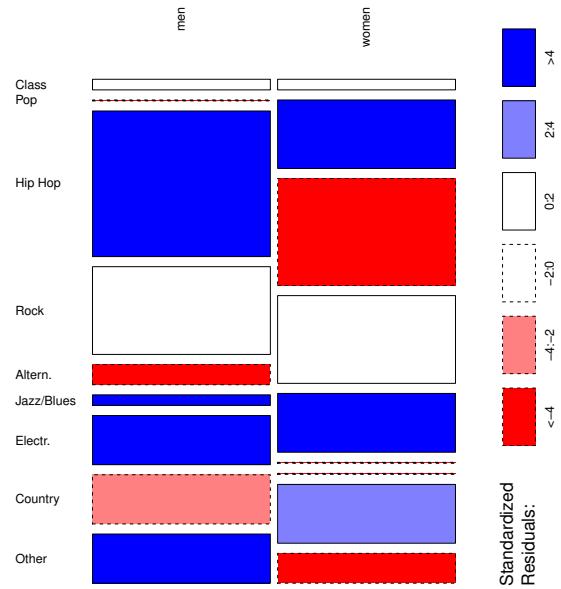


FIGURE 3.1 – Représentation en mosaique

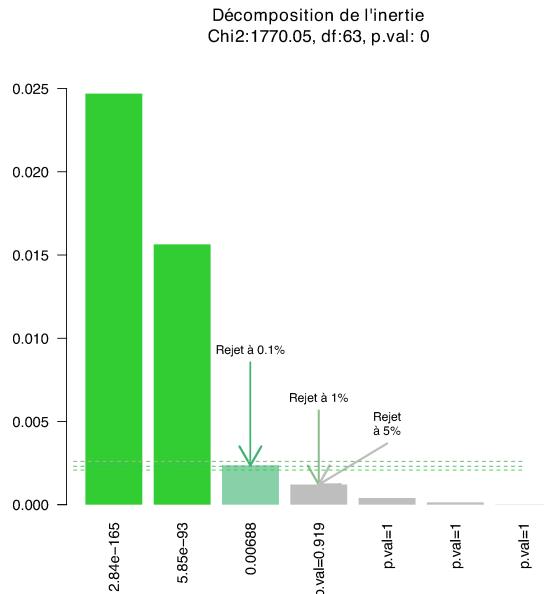


FIGURE 3.2 – Valeurs propres et χ^2

3.7.3 Interprétation des axes

Comme pour l'ACP, nous disposons pour chaque projection de divers indicateurs : contributions à l'inertie, qualité de la représentation calculables globablement et pour chaque espace de projection considérés... mais avant d'illustrer leur utilisation, présentons la superposition des représentations des profils lignes et des profils colonnes.

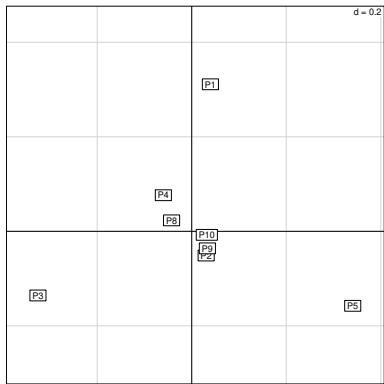
Double interprétation

Pour chaque dimension, il convient d'observer conjointement (i.e. par le biais d'une représentation graphique simultanée) les projections des profils-lignes et des profils colonnes par rapport au i^e axe factoriel⁶. La figure 3.3 montre d'une part les projections des profils lignes et des profils colonnes dans leurs espaces de représentation respectifs, et d'autre part comment les projections des profils lignes et des profils colonnes sont rassemblées sur une même représentation en superposant leurs espaces de représentation.

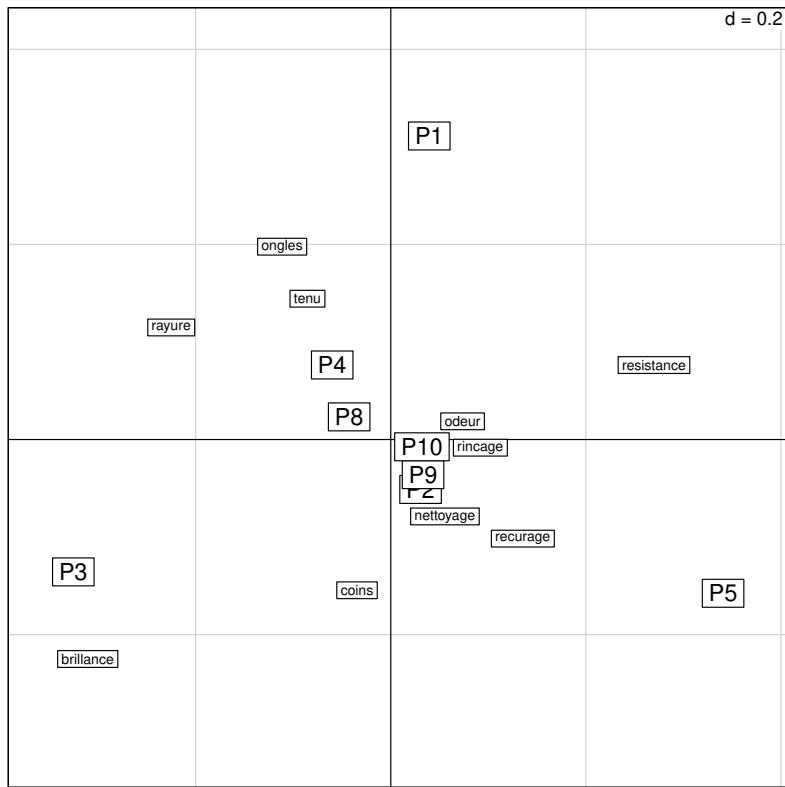
Sur la base de cette représentation, il importe de repérer des modalités de V (resp. W) qui sont bien représentées, qui contribuent fortement à la formation de l'axe et qui s'opposent (les indicateurs présentés ci-dessous aident à cela et la figure 3.4 propose une façon d'exploiter ces indicateurs liés aux projections). Une modalité de V et une modalité de W qui se trouvent du même côté de l'axe par rapport à l'origine tendent à s'associer fréquemment. Selon le même principe, celles opposées par rapport à l'origine tendent à s'éviter.

Si ces positions informent sur les liaisons entre les modalités d'une variable à l'autre, les distances qu'elles montrent ne traduisent rien d'interprétable à priori : la superposition des représentations masque le fait que les espaces de projections sont distincts. Ainsi, l'apparente proximité d'une modalité de V et d'une modalité de W n'est qu'un artefact de la superposition et n'indique rien.

6. Il est évident que cette phrase repose sur un énorme abus de langage. Les projections des profils-lignes sur le j^e axe factoriel et les projections des profils-colonnes sur le j^e axe doivent être reportées selon un axe commun pour le support des deux axes.



(a) Projections séparées des profils lignes et des profils colonnes



(b) Projections simultanées

FIGURE 3.3 – Projections des modalités avec ou sans superposition des espaces de représentation

Contribution à l'inertie d'un axe

Comme dans le cadre de l'ACP, il est intéressant d'observer la contribution du i^e individu à l'inertie du j^e axe factoriel. Cette contribution a pour expression

$$CTR_j(i) = \frac{f_i \varphi_{j_i}^2}{\lambda_j}$$

Il convient d'observer les individus contribuant le plus.

Indicateur de qualité de représentation

De même, le cosinus carré du i^e individu qui quantifie la qualité de sa projection sur le j^e axe est

$$CO2_j(i) = \frac{\varphi_{j_i}^2}{\sum_j \frac{1}{f_{j,j}} \left(\frac{f_{ij} - f_i \cdot f_{j,j}}{f_i} \right)^2}$$

À des fins d'interprétation, il convient d'observer les individus les mieux représentés par leur projection. La figure 3.4 utilise la contribution dans le plan représenté d'une modalité comme paramètre pour ajuster sa taille - une modalité qui contribue relativement beaucoup est représentée avec une taille plus grande qu'une modalité qui contribue peu. De même, le cosinus carré d'une modalité est utilisé pour jouer sur la transparence - une modalité mal représentée dans le plan et qui sera donc associée à un faible cosinus carré sera presque invisible sur la représentation.

CP1: Chi2: 983.21 p.val: 2.84e-165
 CP2: Chi2: 622.37 p.val: 5.85e-93

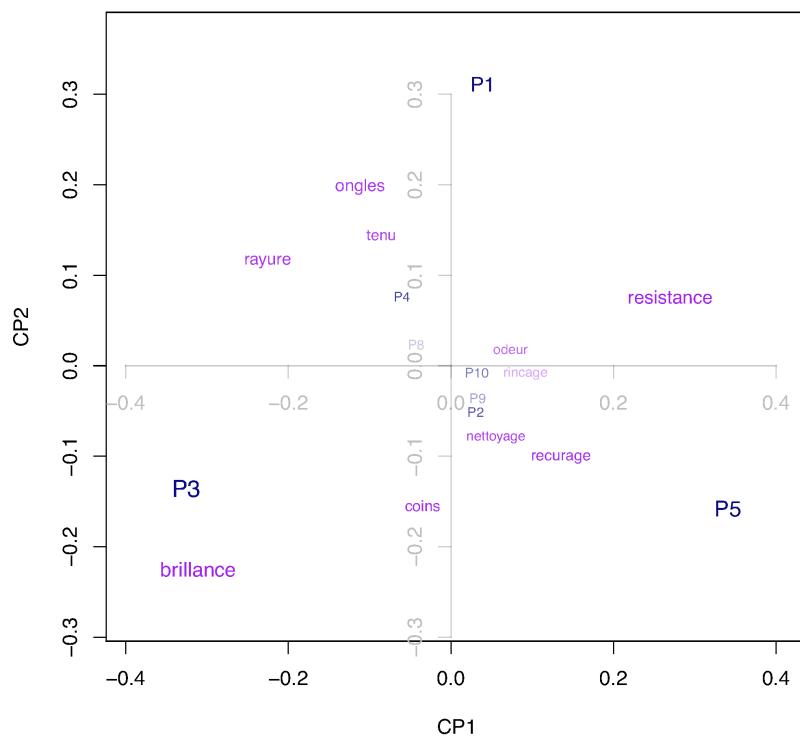


FIGURE 3.4 – Projections et indicateurs de qualité

Interprétations dans le cas de valeurs propres particulières

Nous savons que toutes les valeurs propres sont inférieures ou égales à un : $\forall i, \lambda_i \leq 1$. D'après des résultats fournis en annexe E, nous savons que $I_g \leq \min(n - 1, p - 1)$.

Une valeur propre égale à l'unité correspond au cas où l'axe permet de réaliser une dichotomie des données.

Chapitre 4

L'analyse factorielle des correspondances multiples

4.1 Cadre

L'analyse factorielle des correspondances multiples (AFCM) est une méthode d'analyse factorielle adaptée à l'étude des relations d'un ensemble de variables qualitatives. La présentation qui en suit est fortement inspirée de [PAG13]. Considérons s variables qualitatives notées $Q_1 \dots Q_s$ observées sur une population de n individus. Notons q_j le nombre de modalités observées de la variable Q_j et posons $q = \sum_j q_j$. Les données sont rassemblées dans la matrice $Q = [Q_1 \dots Q_s]$. De façon classique, deux transformations de cette matrice seront considérées : le tableau disjonctif complet et le tableau de Burt. Ces transformations reposent sur un codage de l'information intéressant : un sens peut être donné aux sommes des éléments des lignes (resp. colonnes)...

4.1.1 Tableau disjonctif complet

Les données de la matrice $Q = [Q_1 \dots Q_s]$ peuvent être développées selon le codage binaire suivant :

- Posons $Q_j = \begin{pmatrix} \vdots \\ q_i^j \\ \vdots \end{pmatrix}$ où $\forall i, q_i^j \in \{m_1^j, \dots, m_{q_j}^j\}$
- À ce vecteur correspond la $n \times q_j$ -matrice, notée B_j , dont la k^{e} colonne code la modalité m_k^j :

$$B_j = \begin{pmatrix} & \vdots & & \\ \cdots & b_{ik}^j & \cdots & \\ & \vdots & & \end{pmatrix} \text{ avec } \forall k : 1 \leq k \leq q_j \text{ et } b_{ik}^j = \begin{cases} 1 & \text{si } q_i^j = m_k^j \\ 0 & \text{sinon} \end{cases}$$

- En remplaçant chaque colonne de Q par les colonnes correspondant à son codage binaire, nous obtenons la $n \times q$ -matrice B dite **tableau disjonctif complet** :

$$B = \left(\begin{array}{ccccccccc} & & & & & & & & \\ & \overbrace{\hspace{10em}}^{q \text{ colonnes}} & & & & & & & \\ & \vdots & \\ b_{i1}^1 & \dots & b_{iq_1}^1 & \dots & b_{i1}^s & \dots & b_{iq_s}^s & & \\ \vdots & & \vdots & & \vdots & & \vdots & & \\ & & & & & & & & \end{array} \right) = \begin{pmatrix} b_{11} & \dots & b_{1q} \\ \vdots & & \vdots \\ b_{n1} & \dots & b_{nq} \end{pmatrix}$$

q_1 colonnes définissant la ss-matrice B_1 q_s colonnes définissant la ss-matrice B_s

La matrice B est formée à partir des sous-matrices B_j encodant les Q_j : les q_1 premières colonnes correspondent au codage de Q_1 , et ainsi de suite. Ceci permet l'écriture par blocs $B = [B_1 \cdots B_s]$.

Cette matrice est intéressante pour ses propriétés :

$$b_{\cdot k} = \sum_i b_{ik} = \text{nb d'individus possédant la modalité codée par la } k^{\text{e}} \text{ colonne} \quad (4.1)$$

$$b_{i \cdot} = \sum_j b_{ij} = s \quad (4.2)$$

$$\sum_{k \in I_j} \sum_i b_{ik} = n \quad (4.3)$$

où I_j est l'ensemble des indices associés aux modalités de Q_j , ce qui revient à sommer les éléments de B_j .

$$b_{\cdot \cdot} = \sum_j \sum_i b_{ij} = ns \quad (4.4)$$

4.1.2 Tableau de contingence de Burt

Revenons sur la matrice B et ses sous-matrices B_j . Remarquons que le produit $B'_j * B_{j^*}$, où B'_j est la transposée de B_j , est le tableau de contingence des variables Q_j et Q_{j^*} avec les modalités de Q_j données en lignes.

$$B'_j * B_{j^*} = \begin{pmatrix} & & \vdots & \\ \cdots & \sum_i b_{ik}^j b_{il}^{j^*} & \cdots & \\ & & \vdots & \end{pmatrix}$$

où $b_{ik}^j b_{il}^{j^*} = \begin{cases} 1 & \text{si le } i^{\text{e}} \text{ individu a les modalités associées aux indices } k \text{ et } l \\ 0 & \text{sinon} \end{cases}$

En particulier, pour $j = j^*$, le produit $B'_j * B_j$ est la matrice carrée diagonale dont le k^{e} élément de la diagonale correspond au nombre d'individus possédant la modalité associée à l'indice k .

Le **tableau de Burt** est la juxtaposition de tous les tableaux de contingence réalisables à partir des B_j . Notons C cette $q \times q$ -matrice :

$$C = B' * B \quad (4.5)$$

En notant $B = [B_1 \cdots B_s]$, nous avons :

$$C = \left(\begin{array}{c|c|c|c} B'_1 B_1 & B'_1 B_2 & \cdots & B'_1 B_s \\ \hline B'_2 B_1 & B'_2 B_2 & & B'_2 B_s \\ \hline \vdots & & \ddots & \vdots \\ \hline B'_s B_1 & B'_s B_2 & \cdots & B'_s B_s \end{array} \right)$$

Nous reviendrons sur l'analyse d'un tableau de Burt plus tard.

4.1.3 Exploitation du tableau disjonctif complet

Revenons sur le tableau disjonctif complet. Dans la matrice B , chaque colonne correspond à une **indicatrice** qui encode pour chaque individu s'il possède la modalité en correspondance avec la colonne. Pour faire simple, nous dirons que la k^{e} colonne encode la modalité k . Une modalité caractérise un individu d'autant plus qu'elle est rare. Notons f_k la fréquence de la modalité k et b_{ik} les éléments de l'indicatrice correspondante. Afin de donner plus d'importance aux modalités rares, les b_{ik} seront remplacés par $\frac{b_{ik}}{f_k}$.

Afin de nous rapprocher du cadre formel de l'ACP dans lequel les variables étaient centrées, nous allons centrer ces indicatrices transformées. En général, les individus sont équipondérés i.e pondérés par $1/n$, mais nous pouvons en toute généralité supposer que chaque individu est associé à un poids $p_i \geq 0$ et que $\sum_i p_i = 1$.

Nous pouvons dès lors établir que l'indicatrice transformée a pour moyenne 1 :

$$\sum_i p_i \frac{b_{ik}}{f_k} = \frac{1}{f_k} \underbrace{\sum_i p_i b_{ik}}_{f_k} = 1$$

Le terme générique pour la matrice des indicatrices transformées et centrées est donc $\frac{b_{ik}}{f_k} - 1$.
Posons

$$X = \begin{pmatrix} & & \vdots & \\ \cdots & \frac{b_{ik}}{f_k} - 1 & \cdots & \\ & \vdots & & \end{pmatrix} \quad (4.6)$$

C'est à partir de la matrice X que nous allons étudier l'espace des individus et celui des modalités.

Distances et inerties dans l'espace des individus

Dans l'espace des individus, le produit scalaire sera basé sur la matrice diagonale de terme f_k/s .

$$M = \begin{pmatrix} \ddots & & \\ & \frac{f_k}{s} & \\ & & \ddots \end{pmatrix} \quad (4.7)$$

Ainsi, chaque modalité est prise en compte proportionnellement à sa fréquence d'apparition. Le carré de la distance entre les individus i et j est donc

$$d^2(i, j) = \sum_k \frac{f_k}{s} \left(\frac{b_{ik}}{f_k} - \frac{b_{jk}}{f_k} \right)^2 = \frac{1}{s} \sum_k \frac{1}{f_k} (b_{ik} - b_{jk})^2$$

Le carré de la distance de l'individu i au centre de gravité g , confondu à présent avec l'origine, est

$$d^2(i, g) = \sum_k \frac{f_k}{s} \left(\frac{b_{ik}}{f_k} - 1 \right)^2 = \frac{1}{s} \sum_k \left(\frac{b_{ik}^2}{f_k} - 2b_{ik} + f_k \right) = \frac{1}{s} \left(\sum_k \frac{b_{ik}^2}{f_k} - 2 \underbrace{\sum_k b_{ik}}_s + \underbrace{\sum_k f_k}_s \right)$$

$$d^2(i, g) = \frac{1}{s} \sum_k \frac{b_{ik}}{f_k} - 1 \quad (\text{avec } b_{ik}^2 = b_{ik})$$

Un individu est d'autant plus éloigné du centre de gravité qu'il a des modalités rares. De cette dernière expression découle l'inertie du nuage d'individus :

$$I_g = \sum_i p_i d^2(i, g) = \sum_i p_i \left(\frac{1}{s} \sum_k \frac{b_{ik}}{f_k} - 1 \right) = \sum_i p_i \frac{1}{s} \sum_k \frac{b_{ik}}{f_k} - \underbrace{\sum_i p_i}_1$$

$$I_g = \frac{1}{s} \sum_k \frac{1}{f_k} \underbrace{\sum_i p_i b_{ik}}_{f_k} - 1 = \frac{1}{s} \underbrace{\sum_k 1}_q - 1 = \frac{q}{s} - 1$$

$$\text{d'où } I_g = \frac{q}{s} - 1 \quad (4.8)$$

Cette expression conduit aux remarques suivantes :

- sous l'hypothèse qu'au moins deux modalités sont observées pour chaque variable, l'inertie est minorée par 1 ;
- l'inertie n'est fonction que du nombre de variables et du nombre total de modalités. Autrement dit, tous les tableaux concernant s variables, la j^e variable comportant q_j modalités et le nombre d'individus étant arbitraire, ont la même inertie ! Rien ne peut donc être dit par rapport à la significativité des relations observées d'une façon globale (contrairement à l'AFC...). Aucun sens statistique ne peut être donné à I_g .

Distances et inerties dans l'espace des modalités

Considérons à présent l'inertie du côté de l'espace des modalités. Le produit scalaire découle ici des poids des individus et est basé sur la matrice D :

$$D = \begin{pmatrix} & & \\ \ddots & p_i & \\ & & \ddots \end{pmatrix} \quad (4.9)$$

Dans le tableau disjonctif transformé, chaque colonne correspond à une modalité. Ainsi, les projections qui vont être possibles concerteront les modalités des variables et non les variables elles-mêmes. Soit k une modalité. Le carré de la distance de la modalité au centre de gravité, l'origine, est :

$$d^2(k, 0) = \sum_i p_i \left(\frac{b_{ik}}{f_k} - 1 \right)^2 = \sum_i p_i \left(\frac{b_{ik}}{f_k^2} - 2 \frac{b_{ik}}{f_k} + 1 \right) = \frac{\sum_i p_i b_{ik}}{f_k^2} - 2 \frac{\sum_i p_i b_{ik}}{f_k} + \sum_i p_i = \frac{1}{f_k} - 1$$

Par suite, l'inertie apportée par cette modalité est

$$I(k) = \frac{f_k}{s} d^2(k, 0) = \frac{f_k}{s} \left(\frac{1}{f_k} - 1 \right) = \frac{1 - f_k}{s} \quad (4.10)$$

Remarquons ainsi que l'inertie d'une modalité est d'autant plus importante qu'elle est peu présente dans la population observée. Il est important de s'assurer que l'existence de telles modalités ne perturbe pas l'analyse en conduisant à la formation d'axes mettant l'accent sur les modalités les moins représentées...

En cumulant les inerties des modalités de Q_j , on obtient, avec l qui parcourt les q_j modalités de Q_j :

$$I(Q_j) = \sum_l \frac{1 - f_l}{s} = \frac{q_j - \sum_l f_l}{s} = \frac{q_j - 1}{s} \quad (4.11)$$

L'inertie apportée par une variable est d'autant plus importante que cette variable comporte beaucoup de modalités. Et en cumulant sur l'ensemble des variables Q_1 à Q_s , l'inertie totale du nuage de modalités trouve son expression :

$$\sum_j \frac{q_j - 1}{s} = \frac{\sum_j (q_j - 1)}{s} = \frac{q - s}{s} = \frac{q}{s} - 1$$

L'inertie du nuage de modalités est la même que l'inertie du nuage d'individus, ce qui est une manifestation de la dualité des problèmes posés dans les deux espaces.

Le carré de la distance entre deux modalités k et h est

$$\begin{aligned}
 d^2(k, h) &= \sum_i p_i \left(\frac{b_{ik}}{f_k} - \frac{b_{ih}}{f_h} \right)^2 = \sum_i p_i \left(\frac{b_{ik}^2}{f_k^2} - 2 \frac{b_{ik} b_{ih}}{f_k f_h} + \frac{b_{ih}^2}{f_h^2} \right) \\
 d^2(k, h) &= \sum_i p_i \left(\frac{b_{ik}}{f_k^2} - 2 \frac{b_{ik} b_{ih}}{f_k f_h} + \frac{b_{ih}}{f_h^2} \right) = \overbrace{\sum_i p_i b_{ik}}^{f_k} - 2 \frac{\sum_i p_i b_{ik} b_{ih}}{f_k f_h} + \overbrace{\sum_i p_i b_{ih}}^{f_h} \\
 d^2(k, h) &= \frac{1}{f_k} - 2 \frac{f_{kh}}{f_k f_h} + \frac{1}{f_h} = \frac{f_k + f_h - 2 f_{kh}}{f_k f_h} \tag{4.12}
 \end{aligned}$$

où f_{kh} représente la fréquence des individus ayant les modalités k et h . Ainsi, deux modalités sont d'autant plus proches qu'elles sont partagées par les mêmes individus et d'autant plus éloignées qu'elles sont rares et peu partagées.

Considérons le centre de gravité des modalités de Q_j . En supposant que l'indice l parcourt ces modalités, sa coordonnée sur le i^e axe est

$$\sum_l \frac{f_l}{s} \left(\frac{b_{il}}{f_l} - 1 \right) = \underbrace{\frac{1}{s} \sum_l b_{il}}_1 - \underbrace{\frac{1}{s} \sum_l f_l}_1 = 0$$

Ainsi, les modalités de la variable Q_j sont centrées autour de l'origine, et en étendant cette remarque à l'ensemble des variables, le nuage de modalités est centré autour de l'origine.

Revenons sur les transformations apportées au tableau disjonctif initial. Les colonnes liées aux modalités de Q_j sont des indicatrices codant ses modalités. Il est facile d'observer que ces indicatrices sont deux à deux orthogonales, de même que les indicatrices transformées avant centrage, notée ci-dessous k' et h' :

$$\langle k', h' \rangle = \sum_i p_i \frac{b_{ik}}{f_k} \frac{b_{ih}}{f_h} = \sum_i p_i \overbrace{\frac{b_{ik} b_{ih}}{f_k f_h}}^0 = 0$$

Ainsi, ces indicatrices transformées avant centrage engendrent un sous-espace de dimension q_j . Il est aisément vérifiable que le vecteur ne comportant que des 1 est compris dans ce sous-espace. La somme pondérée par les f_l de ces indicatrices le montre :

$$\sum_l f_l \frac{b_{il}}{f_l} = \sum_l b_{il} = 1$$

Le centrage revient à remplacer les indicatrices par leurs projections dans le sous-espace complémentaire au sous-espace engendré par le vecteur unitaire. Le rang des indicatrices transformées et centrées est donc $q_j - 1$ ¹ ! Ce point est important car il signifie qu'une variable qui comporte q_j modalités aura besoin d'un sous-espace de projection de dimension $q_j - 1$ pour être parfaitement restituée.

Considérons un vecteur v de l'espace des modalités, normé et centré (i.e. de moyenne nulle). La

1. Et au passage, l'orthogonalité des indicatrices transformées est perdue !

projection de la modalité k sur v vaut

$$\begin{aligned}\langle k, v \rangle &= \sum_i p_i \left(\frac{b_{ik}}{f_k} - 1 \right) v_i \\ &= \sum_{i/b_{ik}=1} p_i \left(\frac{v_i}{f_k} - v_i \right) + \sum_{i/b_{ik}=0} p_i (-v_i) = \sum_{i/b_{ik}=1} p_i \frac{v_i}{f_k} - \underbrace{\sum_i p_i v_i}_0 \\ \langle k, v \rangle &= \frac{1}{f_k} \sum_{i/b_{ik}=1} p_i v_i\end{aligned}\tag{4.13}$$

Ainsi, cette projection correspond à la moyenne de v sur les individus possédant la modalité k , que l'on notera \bar{v}_k . L'inertie des modalités de Q_j projetées sur v est exprimée ci-dessous, avec k qui décrit les modalités de Q_j :

$$\sum_{k \text{ modal. de } Q_j} \frac{f_k}{s} \langle k, v \rangle^2 = \sum_k \frac{f_k}{s} \bar{v}_k^2 = \frac{1}{s} \sum_k f_k \bar{v}_k^2 = \frac{1}{s} \eta^2(Q_j, v)\tag{4.14}$$

En effet, v est centrée et normée. Ainsi, l'inertie de la projection d'une variable sur un axe porté par v correspond à son **rapport de corrélation** avec v ramené au nombre de variables. C'est donc un indicateur de liaison bien connu entre la variable Q_j et la direction v considérée. On pourra construire à l'aide de cette information un analogue du cercle des corrélations dans nos résultats.

4.2 Projections des nuages d'individus et de modalités

Récapitulons les éléments de notre cadre de travail :

- À partir de Q_1 à Q_s , l'information a été encodée à l'aide de fonctions indicatrices associées aux modalités (tableau disjonctif complet). Ces variables quantitatives ont été transformées et centrées dans la matrice X .
- Les individus ont été associés à un système de poids (les p_i), rassemblés sur la diagonale de D , matrice définissant le produit scalaire dans l'espace des modalités.
- La matrice M , basée sur les fréquences relatives des modalités (les f_k/s), a été introduite pour définir le produit scalaire dans l'espace des individus. Sa diagonale sert également de poids aux modalités dans leur espace.

Nous disposons ainsi de tous les éléments utiles à la position du problème de la projection des individus, représentés par ces indicatrices transformées, dans un sous-espaces qui restitue au mieux l'information dans un sens précis dans notre présentation de l'ACP (cf. chap. 2) : celui de la maximisation de l'inertie des projections des individus.

Nous savons que les solutions sont basées sur les éléments propres de la matrice $X'DXM$. De plus, d'après la nature des Q_j , la matrice X est de rang au plus $q-s$. Ainsi, il y a au plus $q-s$ valeurs propres non nulles, notées par ordre décroissant $\lambda_1, \dots, \lambda_{q-s}$, et u_1, \dots, u_{q-s} les vecteurs propres associés, orthogonaux deux à deux et M -normés. Les u_j donnent les directions des axes principaux sur lesquels les projections des individus se font.

Ces projections sont contenues dans les composantes principales : la j^e composante principale c_j est XMu_j , et a pour inertie λ_j .

Comme cette inertie est également l'inertie des projections des modalités sur le j^e axe dans l'espace des modalités, nous pouvons interpréter celle-ci. Notons v_j le vecteur D -normé portant le j^e axe de projection dans l'espace des modalités. Nous avons alors d'après 4.14 :

$$\lambda_j = \frac{1}{s} \sum_l \eta^2(Q_l, v_j)\tag{4.15}$$

L'inertie projetée s'interprète comme la moyenne des rapports de corrélations des variables avec l'axe de projection. Ainsi, dans l'espace des modalités, les v_j permettent de maximiser la moyenne des rapports de corrélation des Q_l avec les directions de projection. Au passage, il est remarquable qu'une valeur propre aura une valeur inférieure à l'unité, étant moyenne de valeurs comprises entre 0 et 1, la valeur 1 correspondant à une situation de partition du tableau disjonctif.

Pour terminer, rappelons que le nuage de modalités est centré, comme établi précédemment. Plus précisément, pour chaque variable, les modalités sont centrées autour de l'origine.

Exploitation des relations de dualité

Comme pour l'AFC, il existe des relations quasi-barycentriques, mais dans le cadre de l'AFCM, les interprétations diffèrent :

- la projection d'un individu sur le $j^{\text{e}}\text{axe}$ est, à un facteur $\frac{1}{\sqrt{\lambda_j}}$ près, le barycentre des projections des modalités qu'il a choisi ;
- de même, la projection d'une modalité sur le $j^{\text{e}}\text{axe}$ est, à un facteur $\frac{1}{\sqrt{\lambda_j}}$ près, le barycentre des projections des individus qui l'ont choisie.

En général, les projections respectives sont observées séparément, puis, si le nombre d'individus n'est pas trop important, les projections sont représentées simultanément.

4.3 Mise en œuvre et interprétation des résultats

Nous nous appuierons, à titre d'illustration, dans cette section sur les résultats d'une AFCM appliquée à des données tirées de [PAG13] et portant sur l'expression par 25 étudiants de l'utilité de divers outils pédagogiques : du texte, des films, des animations portant sur l'utilisation d'un logiciel, ainsi qu'un livre de cours et un d'exercices. Pour chaque outil, l'utilité est encodée de 1 (pas utile) à 5 (très utile). Il y aura donc au plus 20 ($5 \times (5 - 1)$) valeurs propres non nulles.

4.3.1 Choix du nombre d'axes de projection

Le profil des valeurs propres décroissantes est bien souvent plus doux pour une AFCM que ce que l'on peut obtenir pour une ACP. Rappelons qu'une variable à q_j modalités aura besoin de $q_j - 1$ axes pour être restituée sans perte. Ce besoin de place pour reconstruire l'information explique en partie l'absence de concentration forte d'inertie sur un petit nombre d'axes.

Les critères utilisables sont les critères classiques de l'ACP : seuil d'inertie cumulée, critères du coude et de Kaiser... pondérés par la remarque qui précède.

Sur la figure 4.1, on peut observer la décroissance régulière des valeurs propres.

4.3.2 Interprétation des résultats

Comme pour les analyses factorielles précédentes, les projections doivent être considérées en tenant compte des habituels indicateurs : indicateurs de contribution à l'inertie et indicateurs de qualité (CO2).

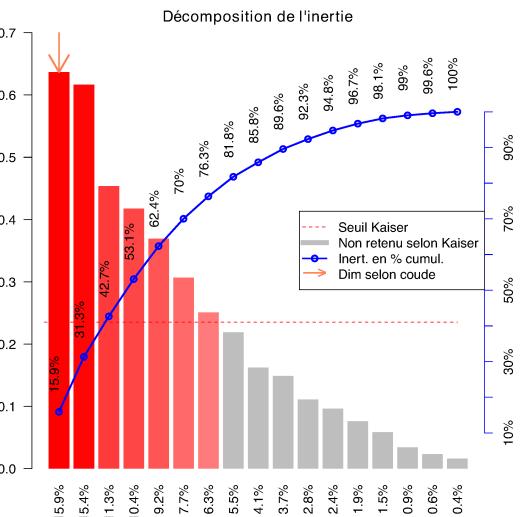


FIGURE 4.1 – Choix de la dimension

Les graphiques de projection des individus et des modalités peuvent être superposés. Le plus intéressant est a priori le second. Il est souvent étudié en complément d'un autre graphique présenté ci-après : le carré des liaisons.

Interprétation des axes

Chaque valeur propre est la moyenne des rapports de corrélation entre l'axe de projection et les variables Q_j . La décomposition de la valeur propre en ces rapports indique :

- la force du lien entre l'axe et chaque Q_j ;
- la part relative d'inertie apportée par Q_j .

Il est possible de représenter cette décomposition graphiquement dans un intervalle $[0, 1]$ - ou sur un carré de côté 1 pour représenter les décompositions de deux axes simultanément. La représentation graphique obtenue est dite carré des liaisons. Le calcul pour chaque axe représente son rapport de corrélation avec chaque variable est nécessaire pour connaître la décomposition de la valeur propre associée et construire le carré des liaisons.

La table 4.1 donne les deux premières valeurs propres et les rapports de corrélation des variables avec les axes de projection leur correspondant. Rappelons que chaque valeur propre est la moyenne des rapports de corrélation. Ici, on voit que les 5 variables sont utiles à l'interprétation du premier axe, mais que seules les 3 premières interviennent sur le second axe. En particulier, les pourcentages indiqués aident à cette lecture.

La figure 4.2a illustre le carré des liaisons qui synthétise ces informations. Le fait que les variables Cours et Exercices n'interviennent que de façon négligeable pour le 2^eaxe apparaît clairement.

valeur propre	0.6366	0.6166
Texte	0.80 (25.21%)	0.93 (30.19%)
Anim.	0.67 (21.18%)	0.95 (30.91%)
Films	0.71 (22.32%)	0.96 (31.18%)
Cours	0.46 (14.68%)	0.09 (2.97%)
Exer.	0.52 (16.61%)	0.14 (4.75%)

TABLE 4.1 – Indicateurs tirés des valeurs propres

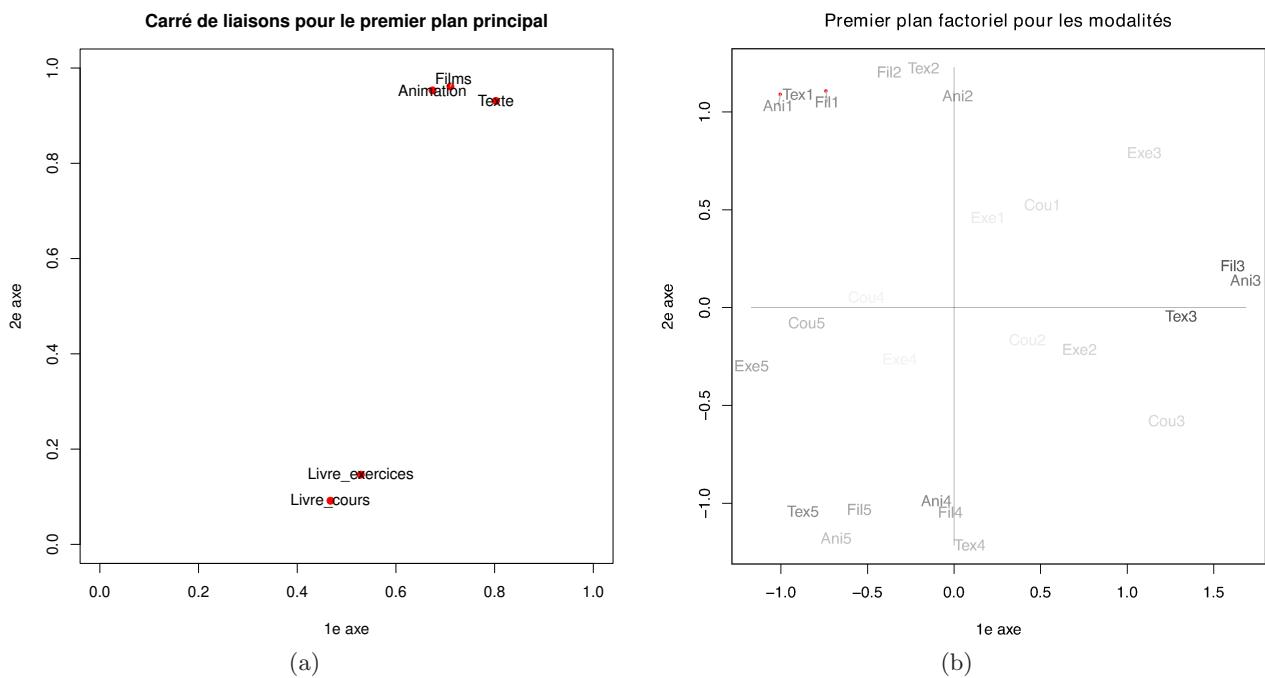


FIGURE 4.2 – Carré des liaisons et Projections des modalités de variables

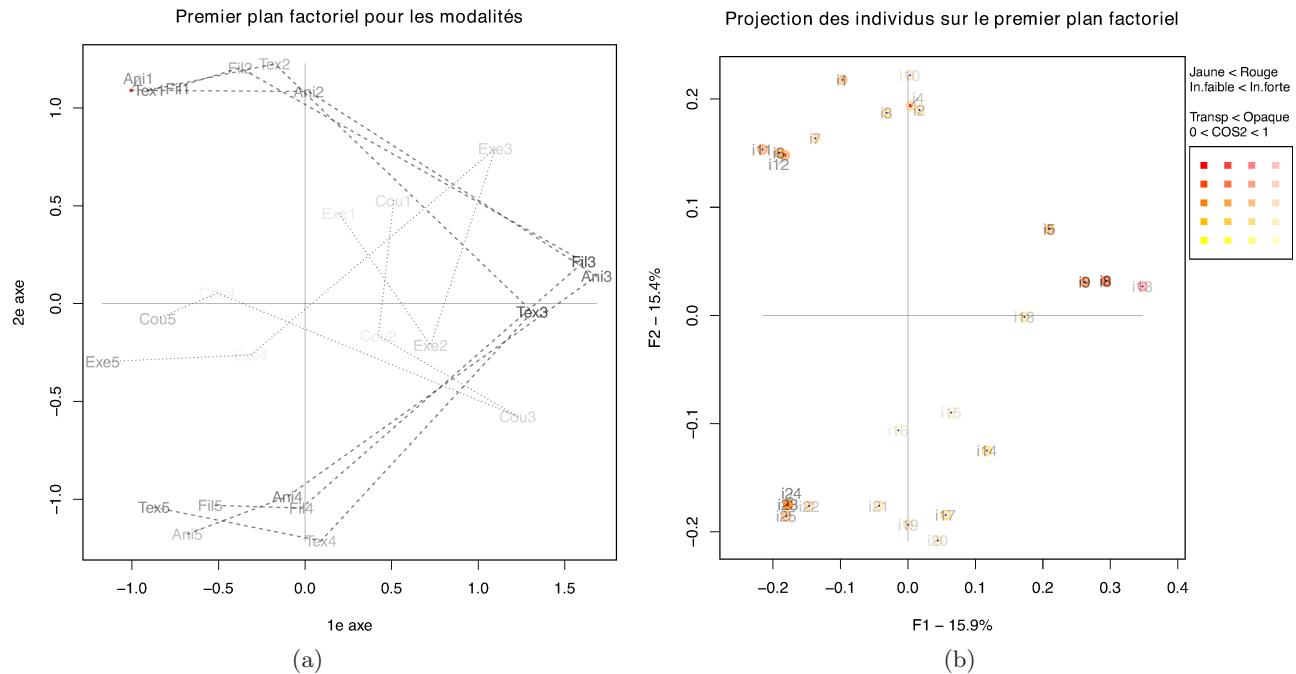


FIGURE 4.3 – Projections des modalités de variables et des individus

Sur la figure 4.2b, les modalités ont été représentées en tenant compte de la somme de leurs COS2. On observe que certaines modalités de Cours et Exercices n'apparaissent presque pas en raison d'un COS2 faible. Les autres modalités sont plus visibles, mais le détails des COS2 montrent qu'aucune n'est très bien représentée.

Le premier axe oppose les individus qui ont trouvé les supports pour le logiciel d'une utilité moyenne aux autres, les autres ayant en particulier trouvé une très grande utilité aux livres. Le second axe oppose les individus qui ont trouvé très utile les supports pour le logiciel à ceux qui ne les ont pas appréciés.

La figure 4.3a est construite comme 4.2b mais matérialise la relation d'ordre existante entre les modalités. Enfin, la figure 4.3b montre les individus projetés dans le premier plan factoriel.

Interprétation des proximités sur les graphiques des projections

Comment s'interprètent les proximités entre projections dans les graphiques produits ?

- Des (projections d') individus sont d'autant plus proches qu'ils sont associés globalement aux mêmes modalités.
- Deux (projections de) modalités de deux variables distinctes sont d'autant plus proches que les sous-groupes d'individus associés sont proches.
- Deux (projections de) modalités d'une même variable sont proches si les groupes d'individus associés sont proches sur les autres variables.

4.4 Équivalence de l'analyse d'un tableau de Burt

L'analyse du nuage de profils-lignes centrés obtenus à partir du tableau de Burt conduit à des résultats fortement liés à ceux obtenus par l'analyse des profils-lignes centrés obtenus d'après le tableau disjonctif complet : notant (λ, ϕ) un couple formé d'une valeur propre et d'une composante principale du second problème, $(\lambda^2, \sqrt{\lambda}\phi)$ est le couple correspondant pour l'analyse du tableau de Burt.

4.5 Trois méthodes pour l'analyse d'un couple de variables

Supposons que deux variables qualitatives X_1 et X_2 soient à analyser, du point de vue des relations de leurs modalités. Nous pouvons proposer trois analyses différentes à partir de nos outils :

1. l'analyse du tableau de contingence obtenu par le croisement des modalités de X_1 et X_2 ;
2. l'analyse du tableau disjonctif complet des modalités de X_1 et X_2 ;
3. l'analyse du tableau de Burt des modalités de X_1 et X_2 .

Intuitivement, les résultats formés par ces trois analyses devraient être équivalents. Le tableau ci-dessous fait le point sur ce sentiment :

Tableau analysé	Facteur	Valeur propre
$Z'_1 Z_2$	ϕ et ψ	μ
$[Z_1, Z_2]$	$\begin{pmatrix} \phi \\ \psi \end{pmatrix}$	$\lambda = \frac{1+\sqrt{\mu}}{2}$
$[Z_1, Z_2]'[Z_1, Z_2]$	$\sqrt{\lambda} \begin{pmatrix} \phi \\ \psi \end{pmatrix}$	λ^2

où Z_1 (resp. Z_2) est le tableau recodant X_1 selon un codage binaire. Ainsi, les projections obtenues sont similaires, ce qui doit conduire à des interprétations de même nature...

Chapitre 5

L'analyse factorielle des données mixtes

Il est fréquent de disposer, pour un ensemble d'individus, à la fois de variables quantitatives et de variables qualitatives. Une approche classique consiste à remplacer l'information apportée par chaque variable quantitative par une nouvelle variable qualitative dont les modalités correspondent aux intervalles de valeurs résultant d'un découpage fondé sur la distribution des valeurs quantitatives.

Ce ré-encodage de l'information soulève quelques questions : combien de classes former pour une variable ? Comment découper les valeurs autour des modes s'il en existe et s'ils s'observent facilement ? La distribution des valeurs reflète-t-elle la distribution de la population générale afin de garantir une stabilité des classes formées ? Cela étant, cette transformation permet de se replacer dans le cadre de l'AFCM assez simplement.

Ce procédé peut sembler comporter une part d'arbitraire, que l'on peut souhaiter éviter. D'autres approches existent, dont l'analyse factorielle de données mixtes qui permet d'éviter ce ré-encodage. La présentation qui suit est adaptée de [PAG13].

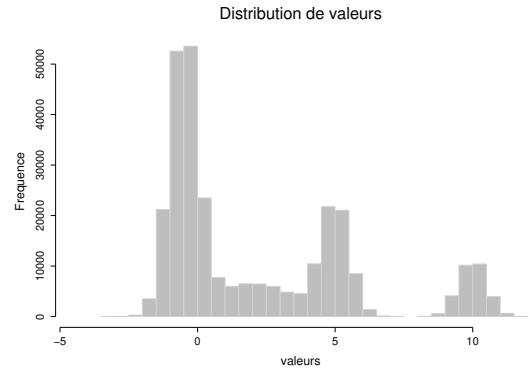


FIGURE 5.1 – Comment partitionner la distribution ?

5.1 Cadre

Nous tâcherons de conserver au mieux les notations adoptées jusqu'ici. Considérons n individus pour lesquels ont été observées :

- p variables quantitatives notées $X_1 \dots X_p$,
- s variables qualitatives notées $Q_1 \dots Q_s$. Notons q_j le nombre de modalités de Q_j et nous posons $q = \sum_j q_j$.

Le tableau de données $[X_1 \dots X_p Q_1 \dots Q_s]$ est transformé, dans un premier temps, en utilisant le codage binaire pour les variables qualitatives, comme présenté dans le chapitre précédent :

$$\left(\begin{array}{ccccccccc}
 & \overbrace{\quad \quad \quad}^{p \text{ colonnes}} & & \overbrace{\quad \quad \quad}^{q \text{ colonnes}} & & \\
 \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots \\
 x_i^1 & \cdots & x_i^p & b_{i1}^1 & \cdots & b_{iq_1}^1 & \cdots & b_{i1}^s & \cdots & b_{iq_s}^s \\
 \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
 & \overbrace{\quad \quad \quad}^{X_1 \dots X_p} & & \overbrace{\quad \quad \quad}^{q_1 \text{ indicatrices encodant } Q_1} & & \overbrace{\quad \quad \quad}^{q_s \text{ indicatrices encodant } Q_s} & & & &
 \end{array} \right)$$

Le tableau précédent va être transformé afin de mieux maîtriser l'inertie apportée par chaque variable.

Chaque individu est associé à un poids $p_i \geq 0$ et l'ensemble des poids satisfait $\sum_i p_i = 1$. Notons f_k^j la fréquence d'apparition de la k^e modalité de Q_j . Nous avons, avec j indiquant Q_j et k correspondant à la k^e modalité de Q_j

$$\sum_i p_i b_{ik}^j = \sum_{i/b_{ik}^j=1} p_i = f_k^j \text{ et } \sum_k f_k^j = 1$$

Considérons le tableau précédent transformé par les deux opérations suivantes :

- centrage et réduction des variables X_j
- transformation de l'indicatrice associée à la k^e modalité de Q_j par division par sa fréquence f_k^j
puis centrage

Ces transformations donnent le tableau suivant :

$$X = \left(\begin{array}{c|ccccccccc} & \overbrace{\quad \quad \quad}^{p \text{ colonnes}} & & \overbrace{\quad \quad \quad}^{q \text{ indicatrices, transformées et centrées}} \\ \vdots & \vdots \\ y_i^1 & \cdots & y_i^p & \frac{b_{i1}^1 - 1}{f_1^1} & \cdots & \frac{b_{iq_1}^1 - 1}{f_{q_1}^1} & \cdots & \frac{b_{i1}^s - 1}{f_1^s} & \cdots & \frac{b_{iq_1}^s - 1}{f_{q_1}^s} \\ \vdots & \vdots \\ \hline Y_1 & \cdots & Y_p & & & & & & & \\ & & & \text{centrées et} & & & & & & \\ & & & \text{réduites} & & & & & & \end{array} \right)$$

Nous pouvons observer - ou nous souvenir - que le bloc d'indicatrices transformées encodant Q_j forme une famille de vecteurs de rang $q_j - 1$. Ainsi la matrice est de rang au plus $p + q - s$.

5.1.1 Distance et inertie dans l'espace des individus

La matrice de produit scalaire dans l'espace des individus est construite par morceaux : les colonnes correspondant à $X_1 \dots X_p$ sont pondérées à 1 tandis que les indicatrices sont pondérées par la fréquence de la modalité associée. Ainsi, nous avons :

$$M = \left(\begin{array}{c|c} 1 & \\ \vdots & \\ \hline & 1 \\ & f_1^1 \\ & \ddots & f_{q_1}^1 \\ & & \ddots & f_1^s \\ & & & \ddots & f_{q_s}^s \end{array} \right)$$

De là, nous tirons la distance d'un individu au centre de gravité du nuage d'individus, confondu avec l'origine en raison du centrage de l'ensemble des variables :

$$\begin{aligned}
 d^2(i, 0) &= \sum_{j=1}^p y_i^j{}^2 + \sum_{j=1}^s \sum_{k=1}^{q_j} f_k^j \left(\frac{b_{ik}^j}{f_k^j} - 1 \right)^2 = \sum_{j=1}^p y_i^j{}^2 + \sum_{j=1}^s \sum_{k=1}^{q_j} f_k^j \left(\frac{b_{ik}^j}{f_k^j} - 2 \frac{b_{ik}^j}{f_k^j} + 1 \right) \\
 d^2(i, 0) &= \sum_{j=1}^p y_i^j{}^2 + \sum_{j=1}^s \sum_{k=1}^{q_j} \left(\frac{b_{ik}^j}{f_k^j} - 2b_{ik}^j + f_k^j \right) = \sum_{j=1}^p y_i^j{}^2 + \sum_{j=1}^s \left(\underbrace{\sum_{k=1}^{q_j} \frac{b_{ik}^j}{f_k^j}}_1 - 2 \underbrace{\sum_{k=1}^{q_j} b_{ik}^j}_{\text{1}} + \underbrace{\sum_{k=1}^{q_j} f_k^j}_{\text{1}} \right) \\
 d^2(i, 0) &= \sum_{j=1}^p y_i^j{}^2 + \sum_{j=1}^s \left(\sum_{k=1}^{q_j} \frac{b_{ik}^j}{f_k^j} - 2 + 1 \right) = \sum_{j=1}^p y_i^j{}^2 + \sum_{j=1}^s \sum_{k=1}^{q_j} \frac{b_{ik}^j}{f_k^j} - s = \sum_{j=1}^p y_i^j{}^2 + \sum_{j=1}^s \frac{1}{f_k^j} - s
 \end{aligned}$$

où $f_{k^i}^j$ est la fréquence de la modalité de Q_j prise par l'individu. Ainsi, il est clair que sa distance au centre de gravité est d'autant plus grande qu'il prend des modalités peu fréquentes et/ou des valeurs éloignées des valeurs moyennes. L'expression de l'inertie du nuage d'individus découle facilement de l'avant dernière écriture de $d^2(i, 0)$:

$$\begin{aligned}
 I_g &= \sum_i p_i d^2(i, g) = \sum_i p_i \sum_{j=1}^p y_i^j{}^2 + \sum_i p_i \left(\sum_{j=1}^s \sum_{k=1}^{q_j} \frac{b_{ik}^j}{f_k^j} - s \right) \\
 I_g &= \sum_{j=1}^p \text{var } Y_j + \sum_{j=1}^s \sum_{k=1}^{q_j} \sum_i \frac{p_i b_{ik}^j}{f_k^j} - s \sum_i p_i = p + \sum_{j=1}^s \sum_{k=1}^{q_j} 1 - s \\
 I_g &= p + \sum_{j=1}^s q_j - s = p + q - s
 \end{aligned}$$

i.e.

$$I_g = p + q - s \quad (5.1)$$

Comme pour l'AFCM, l'expression dépend des paramètres généraux sur les variables : nombres de variables quantitatives et qualitatives, et nombre de modalités distinctes.

5.1.2 Espace des variables

Nous adoptons pour matrice de produit scalaire la matrice diagonale notée D et portant les poids :

$$D = \begin{pmatrix} \ddots & & \\ & p_i & \\ & & \ddots \end{pmatrix}$$

Soit v un vecteur centré et réduit de l'espace des variables. Déterminons l'expression de l'inertie du nuage de variables projetées dans le sous-espace qu'il engendre.

Nous avons d'une part, pour les Y_j :

$$\langle k, Y_j \rangle = \sum_i p_i (v_i \times y_i^j) = \text{corr}(v, Y_j)$$

et d'autre part, pour les indicatrices :

$$\left\langle k, B_k^j \right\rangle = \sum_i p_i v_i \times \left(\frac{b_{ik}^j}{f_k^j} - 1 \right) = \sum_i \underbrace{\frac{p_i v_i b_{ik}^j}{f_k^j}}_0 - \sum_i p_i v_i$$

$$\left\langle k, B_k^j \right\rangle = \sum_{i/b_{ik}^j=1} \frac{p_i v_i}{f_k^j} = \frac{1}{f_k^j} \sum_{i/b_{ik}^j=1} p_i v_i = \bar{v}_k^j$$

où \bar{v}_k^j correspond à la moyenne de v sur les individus possédant la k^e modalité de Q_j .

Calculons à présent l'inertie. Nous conservons les pondérations définies précédemment : chaque Y_j est pondéré à 1 et chaque indicatrice est pondérée par la fréquence d'apparition de sa modalité.

$$I(\Delta_v) = \sum_{j=1}^p \text{corr}^2(v, Y_j) + \sum_{j=1}^s \sum_{k=1}^{q_j} f_k^j \bar{v}_k^j - \sum_{j=1}^p \text{corr}^2(v, Y_j) + \sum_{j=1}^s \eta^2(Q_j, v) \quad (5.2)$$

En effet, $\sum_{k=1}^{q_j} f_k^j \bar{v}_k^j$ correspond à la variance intergroupe relativement aux modalités de Q_j de la variable centrée v . Et comme v est réduite, cette grandeur correspond aussi au rapport de corrélation de v avec Q_j , noté $\eta^2(Q_j, v)$.

5.2 Projections des nuages d'individus et de variables et modalités

Les solutions sont basées sur les éléments propres de la matrice $X'DX^{-1}$. De plus, d'après la nature des Q_j , la matrice X est de rang au plus $p+q-s$. Ainsi, il y a au plus $p+q-s$ valeurs propres non nulles, notées par ordre décroissant $\lambda_1, \dots, \lambda_{p+q-s}$, avec u_1, \dots, u_{p+q-s} les vecteurs propres associés, orthogonaux deux à deux et M -normés. Les u_j donnent les directions des axes principaux sur lesquels les projections des individus se font.

5.3 Mise en œuvre et interprétation des résultats

Nous nous appuierons, à titre d'illustration, sur l'analyse d'un jeu de données portant sur les caractéristiques d'un ensemble d'engins électriques (skate, gyropode, trottinette...) utilisés pour les déplacements urbains¹. Aux données quantitatives telles que le prix, l'autonomie, le temps de charge, la vitesse maximale et le poids s'ajoutent trois variables qualitatives : le type d'engin, le contrôle à distance et la note attribuée par un jury d'experts intergalactiques.

Plus précisément, le type d'engin a quatre modalités (gyropode, skate, solo wheel, trottinette), le contrôle à distance deux (oui ou non) et la note est un entier compris entre 2 et 5.

Nom	Autonomie (km)	Poids (kg)	Tmp charge (h)	Vit max (km/h)	Pilotage à distance	Type	Prix	Note
InMotion V5F	35	12	3.5	25	Non	solo wheel	690	5
InMotion V8	45	13.8	4.5	25	Non	solo wheel	825	5
Trottix TX1	25	13.9	3	25	Non	trottinette	1580	5
Kingsong KS-14C 680 Wh	50	14.7	5	25	Non	solo wheel	1600	5
InMotion V5+	35	11.5	3	18	Non	solo wheel	690	5
...

TABLE 5.1 – Premières lignes du jeu de données

Après transformation des données (centrage et réduction pour les variables quantitatives et encodage via des indicatrices transformées et centrées pour les variables qualitatives), nous pouvons réaliser

1. Les données sont tirées d'un comparatif trouvé sur <https://www.lesnumeriques.com/>.

une ACP non réduite et basée sur la pondération des variables correspondant à la diagonale de M . Ceci conduit dans un premier temps à l'exploitation des éléments propres.

5.3.1 Choix du nombre d'axes de projection

D'après les variables du jeu de données, nous savons que nous allons considérer un tableau X qui aura au plus un rang égal à $5+(2+4+4)-3=12$. Il y aura donc au plus 12 axes à considérer. La figure 5.2 illustre la décomposition de l'inertie.

À partir des vecteurs propres, il est possible de déterminer les composantes principales, ce qui permet dans un second temps de calculer, pour chaque variable et selon sa nature, son coefficient de détermination (carré de la corrélation) ou son rapport de corrélation avec la composante principale. Ceci permet d'observer la part des différentes variables dans l'inertie d'un axe puisque la somme de ces valeurs, données en colonnes dans la table 5.2, correspond à l'inertie de la composante principale.

Par ailleurs, la somme des éléments en ligne vaut 1 pour les variables quantitatives et $q_j - 1$ pour les variables qualitatives (lorsque les colonnes pour toutes les composantes principales sont données).

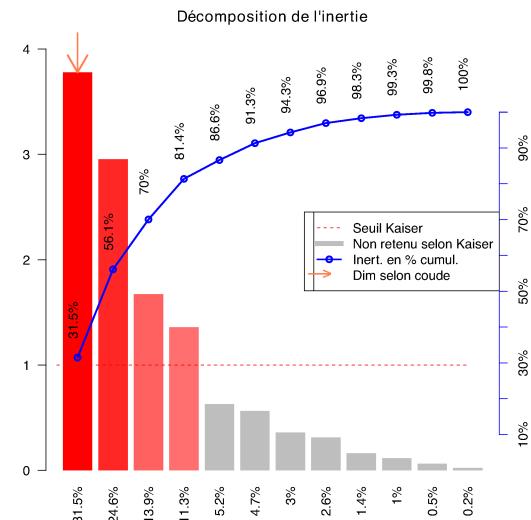


FIGURE 5.2 – Choix de la dimension

corr ² ou η^2	F1	F2	F3	F4
λ_i	3.7788	2.9540	1.6731	1.3597
Autonomie	0.86	0.00	0.00	0.02
Poids	0.61	0.30	0.01	0.01
Tps charge	0.53	0.01	0.00	0.19
Vit max	0.26	0.07	0.48	0.04
Prix	0.06	0.75	0.00	0.01
Pilot. dist	0.42	0.23	0.15	0.03
Type	0.70	0.86	0.29	0.67
Note	0.34	0.72	0.74	0.38

TABLE 5.2 – Décomposition de l'inertie des axes

Cette décomposition permettra de construire les carrés des liaisons. Celui correspondant aux deux premières colonnes est représenté sur la figure 5.3a et le suivant sur la figure 5.5a.

5.3.2 Interprétation des axes et des projections

L'interprétation va se baser sur différents éléments déjà rencontrés dans l'ACP et l'AFCM :

- le carré des liaisons avec l'apparition des variables quantitatives grâce à l'exploitation de leur coefficient de détermination ;
 - le cercle de corrélations qui ne fait figurer que les variables quantitatives ;
 - les projections des modalités - avec lesquelles nous pourrions faire figurer les projections des variables quantitatives car ces projections sont bien dans le même espace ;
 - les projections des individus ;
 - et les indicateurs habituels de contribution , de qualité, utilisés ici à travers les représentations construites à l'aide de couleurs et de transparence...

Ci-après, nous pouvons utiliser le carré des liaisons pour identifier les variables qui sont les plus liées avec le premier axe. Nous pouvons voir que l'autonomie, le poids et le temps de charge, mais également le type (en réalité, une seule modalité, le type skate) semblent liés au premier axe. Le cercle de corrélations montre qu'il y a un effet taille sur les variables quantitatives, tandis que les projections des modalités montrent que le type skate et le pilotage à distance se projettent à l'opposé des variables quantitatives.

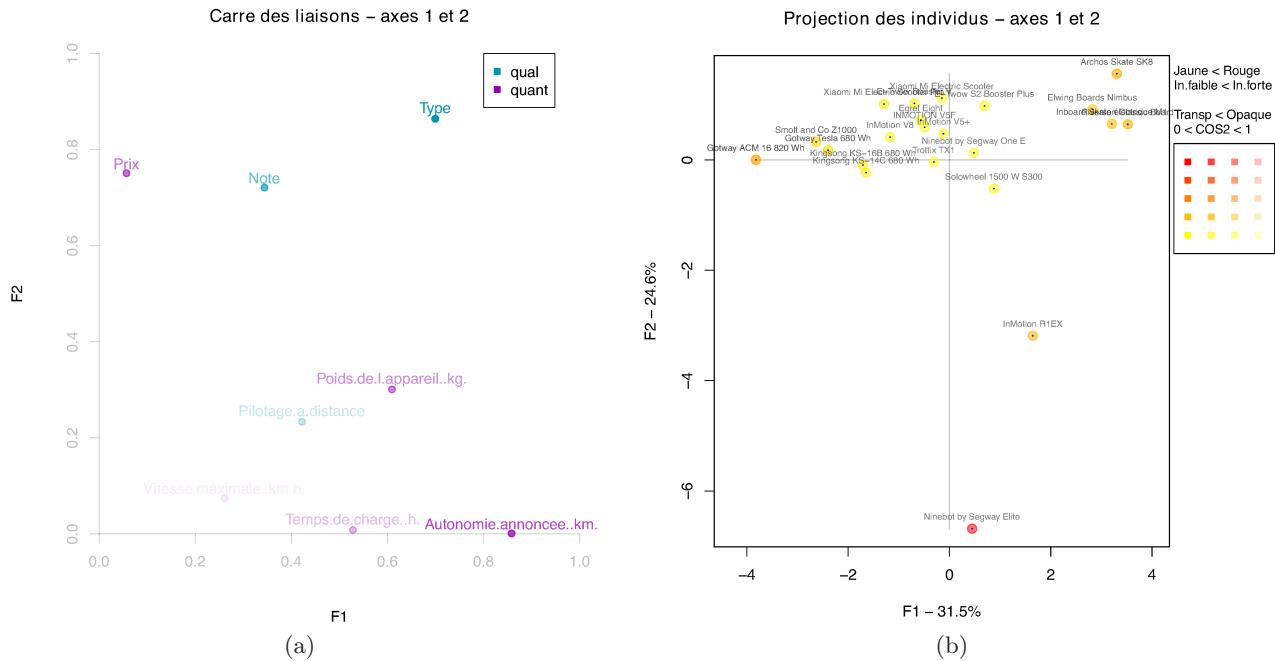


FIGURE 5.3 – Analyse de l'inertie et projections des individus

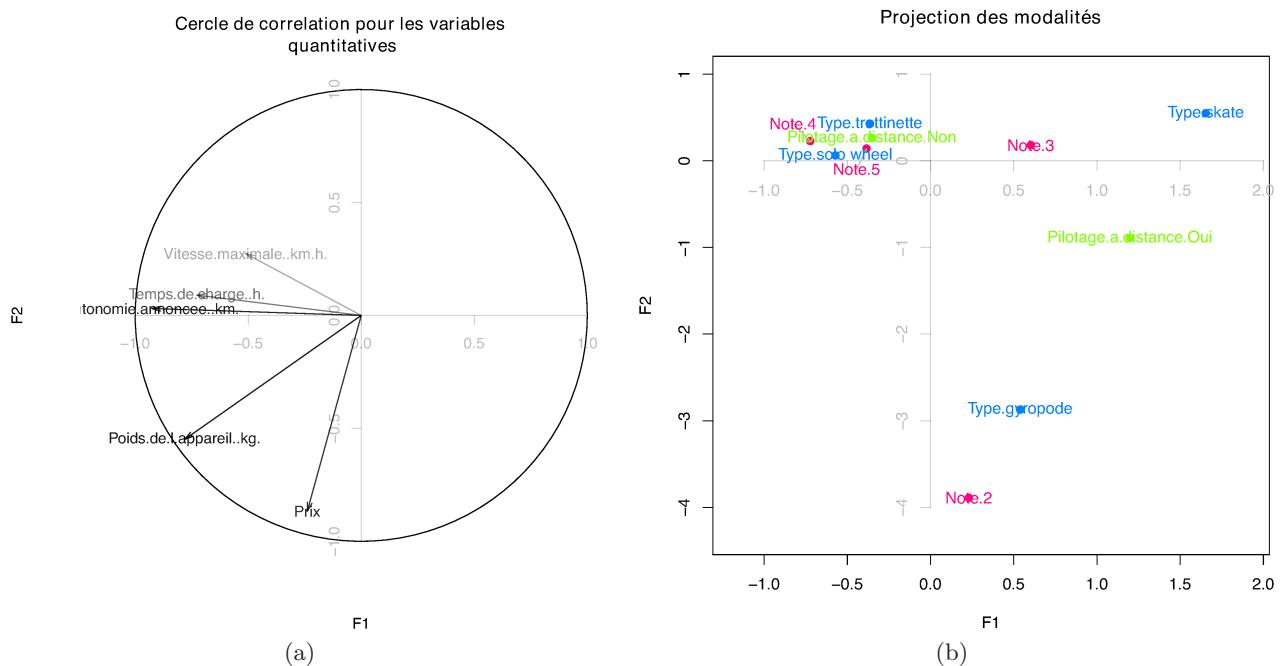


FIGURE 5.4 – Projections des modalités de variables

Le fait que les skates électriques aient des valeurs sous les niveaux moyens et également soient les seuls à être pilotables à distance, est mis en évidence ici. Les projections des individus permettent d'observer que les skates sont opposés à la plupart des autres le long du premier axe.

Le second axe est construit sur les liaisons entre le prix, la note et le type. Les projections des modalités montrent que le type gyropode est très bas sur le deuxième axe, ainsi que la note 2. Le prix est également très dépendant. Les projections des engins montrent que les 2 seuls gyropodes sont sous le 2^e axe, distants des autres, ce qui correspond au fait qu'ils sont très chers, et mal notés.

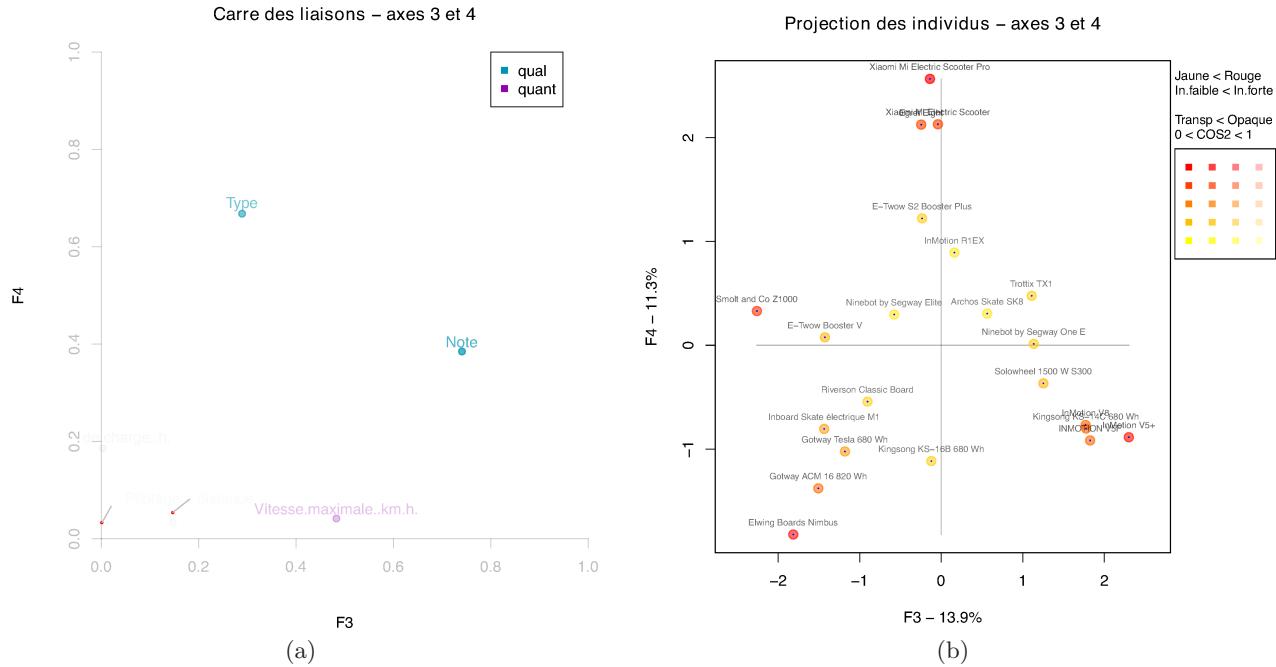


FIGURE 5.5 – Analyse de l'inertie et projections des individus

Les graphiques des figures 5.5 et 5.6 donnent les projections sur les axes 3 et 4.

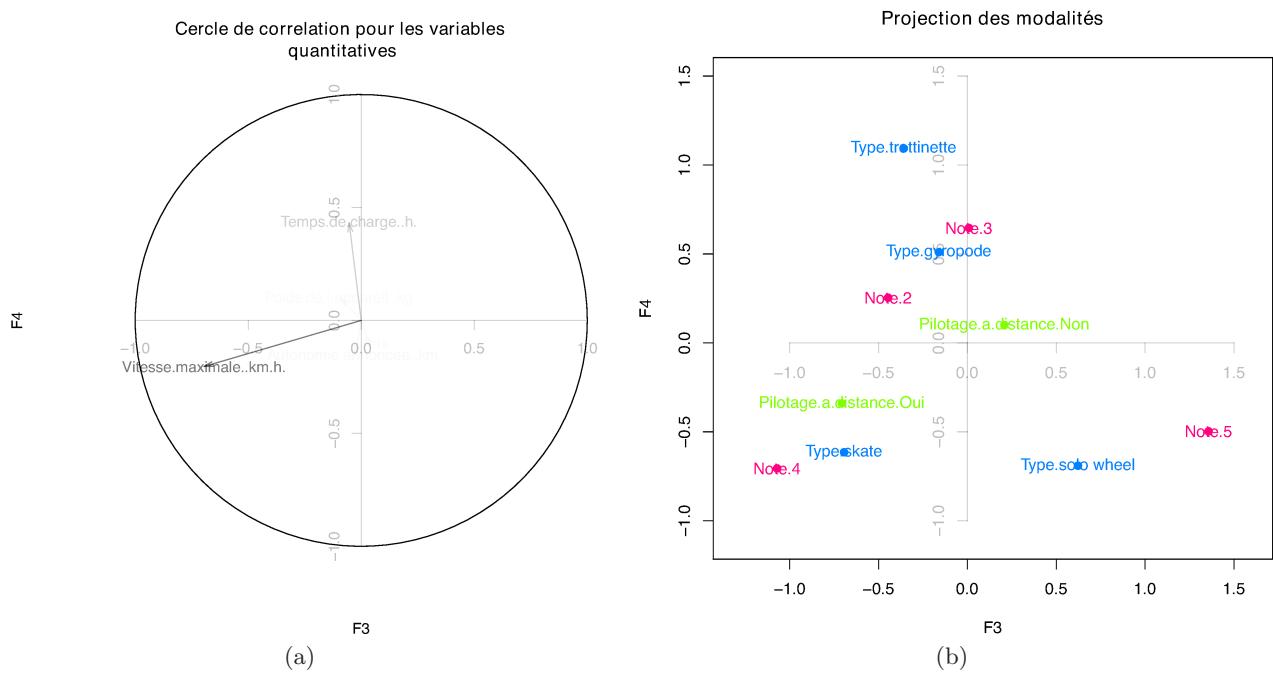


FIGURE 5.6 – Projections des modalités et des variables

Deuxième partie

Analyses factorielles dans le cadre fonctionnel

Chapitre 6

Introduction au cadre fonctionnel

Le type de données qui va être étudié à présent est d'une grande généralité. Une fois sa nature précisée, nous identifierons, dans le cadre auquel il est attaché, les structures qui permettent de définir le problème de l'analyse en composantes principales.

6.1 Données fonctionnelles

Qu'entendons nous par "données fonctionnelles" ? Il est courant en analyse de données d'avoir un ensemble de variables aléatoires (réelles le plus souvent, mais pouvant être qualitatives) mesurées pour chaque individu d'une population. Chaque individu est alors associé à un vecteur fini d'éléments de natures diverses (qualitatif, quantitatif...). Dans le contexte fonctionnel, les variables aléatoires réelles sont remplacées par des variables aléatoires fonctionnelles, i.e. dont les valeurs sont des fonctions : ainsi les valeurs ponctuelles sont remplacées par des fonctions ; pour chaque individu, une ou plusieurs fonctions sont mesurées.

i	X_1	X_2		X_p
1	1.2	0.2		5.2
2	1.0	-0.3		5.3
n	2.3	0.9		6.3

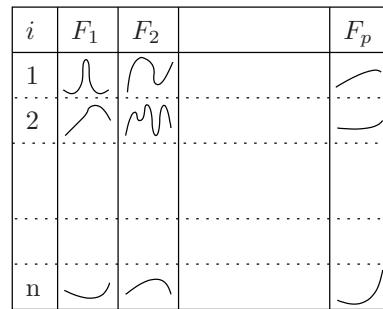


FIGURE 6.1 – Données simples et données fonctionnelles

Pour l'emploi que nous réservons aux données, nous devons préciser la nature des fonctions étudiées. [KLE73] a montré l'applicabilité de la recherche de composantes principales à des variables aléatoires à valeurs dans des espaces de Hilbert séparables et réels. Ce cadre est très vaste. Dans les applications, nous sommes la plupart du temps confrontés à des fonctions appartenant à des espaces bien plus restreints mais aussi bien plus riches. Afin de rester dans un cadre de travail raisonnable et familier, nous supposerons vérifié certaines hypothèses concernant les fonctions manipulées :

- Pour une variable aléatoire fonctionnelle donnée, les fonctions considérées sont à support compact. Ce support est le même pour toutes les fonctions. Pour assurer une certaine simplicité de l'exposé, nous supposerons ici qu'il s'agit d'un intervalle réel fermé borné noté T . Dans toute cette partie, la notation T peut être remplacée par celle d'un ensemble Ω plus général, mais mesurable. Cela peut être \mathbb{R} , $\cup_i T_i$ une réunion d'intervalles bornés, un domaine de \mathbb{R}^4 ...
- Les fonctions sont supposées de carré intégrable pour la mesure de Lebesgue.

Nous nous plaçons ainsi dans l'espace de fonctions $L_2(T)$. C'est un espace de Hilbert séparable réel muni du produit scalaire suivant :

$$\forall f, g \in L_2(T), \langle f, g \rangle = \int_T fg \quad (6.1)$$

6.1.1 Considérations pratiques

Les fonctions que nous nous proposons d'étudier sont rarement connues de façon exacte. Sous l'hypothèse qu'un phénomène continu dans le temps est l'objet d'étude, il n'est pas possible d'obtenir par des mesures physiques la totalité des points qui caractérisent une courbe. Dans un tel cas, il est nécessaire de disposer d'un procédé de reconstruction des données à partir des mesures. Le problème de la reconstruction des données dépend de plusieurs paramètres parmi lesquels :

- les hypothèses sur la nature des fonctions mesurées
- les paramètres du processus de mesure (fréquence d'échantillonage, caractéristiques de erreurs de mesure...)
- le modèle de représentation pour les données à reconstruire
- la/le coïncidence/décalage entre le temps de mesure et le référentiel commun de comparaison

Les hypothèses sur la nature des fonctions mesurées sont primordiales, puisqu'elles guident le processus de reconstruction des fonctions "originales". Nous avons déjà posé ci-dessus quelques hypothèses. Nous voyons que cela influe directement sur le modèle de représentation. Les paramètres du processus de mesure sont importants notamment pour pouvoir estimer la précision des résultats que l'on peut attendre d'une analyse.

Modèles de représentation des données

Supposons que les données fonctionnelles soient recueillies par un procédé de discréétisation et qu'elles correspondent à un phénomène continu.

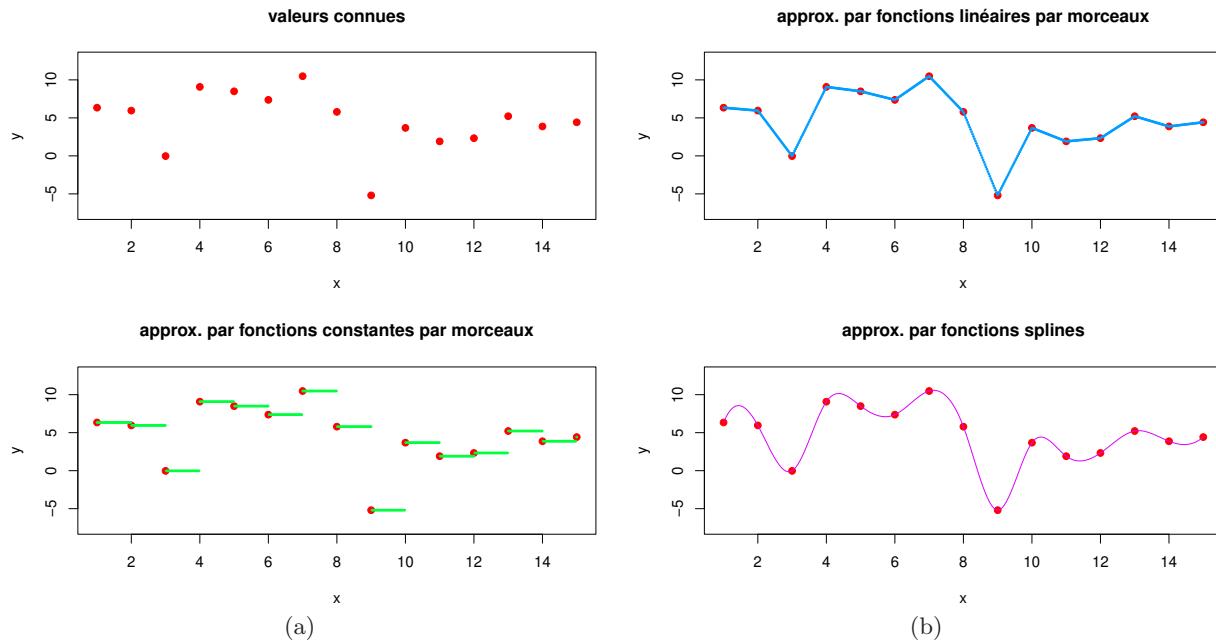


FIGURE 6.2 – Exemples de méthodes simples de reconstruction de fonctions

La première étape est alors la reconstruction des fonctions. Si les mesures sont supposées sans erreur, alors elles peuvent être interpolées par une méthode prenant en compte les hypothèses sur la

nature des fonctions(Cf. figure 6.2).

Dans le cas contraire, le recours à une procédure de lissage est plus indiqué.

Le lissage de données est l'objet d'une foule de techniques [RS97b]. La méthode la plus simple est la méthode de lissage linéaire, combinant les mesures par combinaisons linaires des mesures pondérées par des fonctions simples. Viennent ensuite dans l'ordre de sophistication croissante les méthodes de lissage par pondération locale telles que les méthodes de lissage par noyaux (Cf. figure 6.3a). Le lissage peut également être réalisé par l'ajustement d'un développement dans une base de l'espace de fonctions, base choisie pour son adéquation avec les caractéristiques des fonctions étudiées ou pour la qualité de reconstruction qu'elle permet d'atteindre. Cet ajustement minimise en général l'erreur commise à l'aide du critère des moindres carrés.

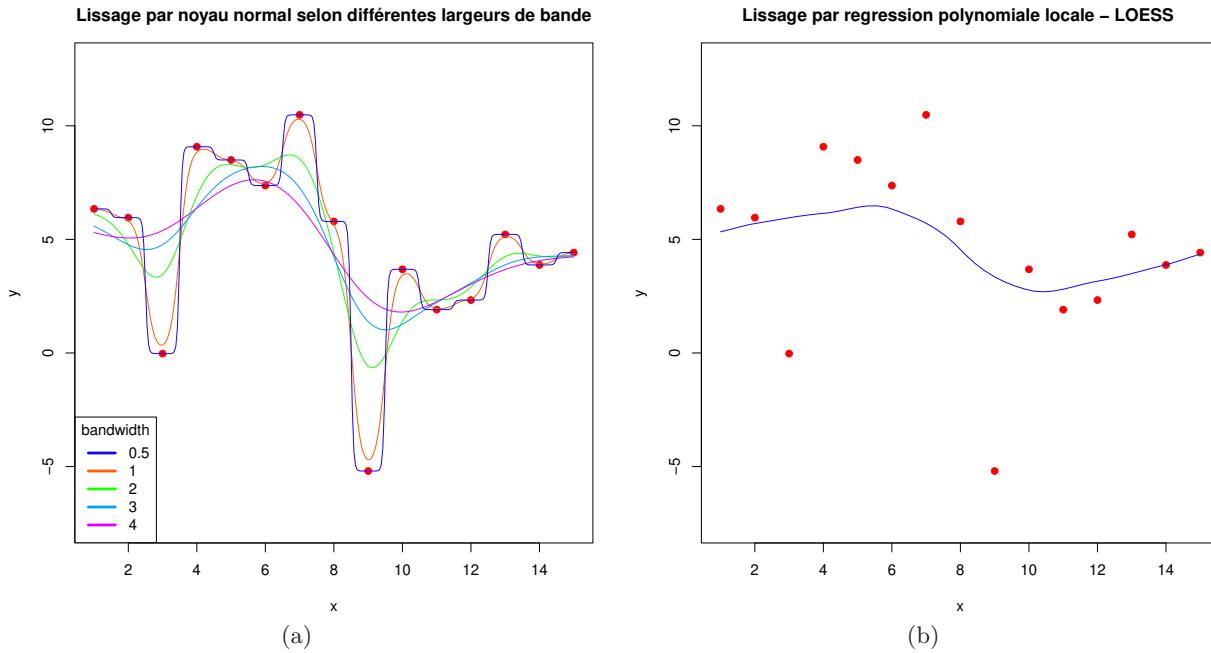


FIGURE 6.3 – Exemples de lissage

Les méthodes précédentes utilisent en général les valeurs de la fonction. Il est des situations dans lesquelles des contraintes sur les valeurs des dérivées doivent également entrer dans le processus de reconstruction afin d'aboutir, par exemple, à une fonction lissée dont la dérivée minimise les variations brutales quitte à ne pas “coller” parfaitement aux mesures. Ce type de lissage utilise l'approche par régularisation. Pratiquement, cela conduit à gérer le lissage sur les données de la fonction et les contraintes portant sur les autres éléments variationnels en même temps. Ce compromis s'opère par le biais d'optimisation d'une fonctionnelle réunissant quantitativement les deux objectifs concurrents. [RH00] (resp. [DR99]) étudient le paramétrage de ce type de lissage lorsque les splines (resp. ondelettes) sont utilisées pour la composition de la fonction.

Problèmes de recalage

Qu'est ce que le problème de recalage de fonctions ? La figure 6.4 montre le graphique de quatre fonctions dont les variations ou la dynamique semblent proches, d'un point de vue global. Pourtant nous pouvons observer que l'une des courbes semble mal synchronisée par rapport aux autres courbes. Si notre intérêt se porte sur ces décalages, sur leur recherche, cette représentation est adaptée pour mettre en évidence l'information cherchée. Mais dans le cas où ce qui est intéressant est l'ensemble des valeurs de chaque fonction en certains points remarquables, comme ceux mis en évidence sur la figure

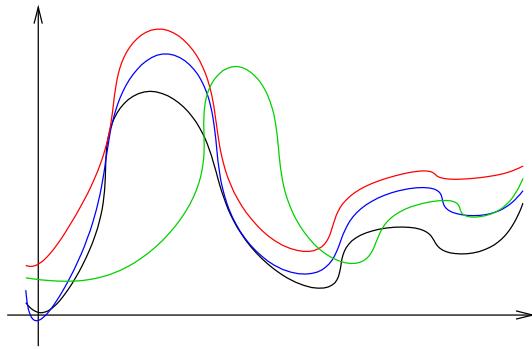


FIGURE 6.4 – Exemple de courbes asynchrones

6.5 alors il est fait implicitement allusion à un référentiel commun dans lequel il est intéressant de se placer pour l'observation des courbes. Il est possible de rechercher ce référentiel commun en réajustant les courbes à l'aide de points facilement identifiables dans toutes les courbes. [RBG95, RWF95] se sont

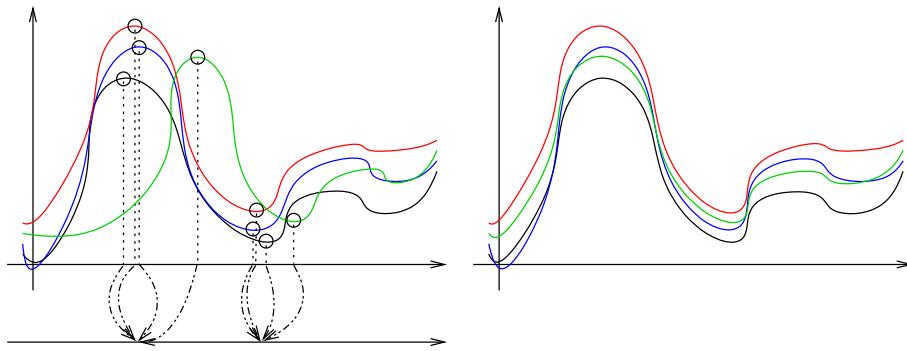


FIGURE 6.5 – Recalage de courbes asynchrones

penchés sur le problème du recalage de données bio-mécaniques dans leurs applications respectives. [KG92] analyse l'importance de ce recalage pour une définition cohérente de la moyenne de courbes, ainsi que d'autre concepts statistiques.

Ces quelques lignes nous montrent que le prétraitement et le formatage de données fonctionnelles prennent une place dans un processus d'analyse qui n'est en rien comparable avec le centrage et la réduction des variables pour une ACP classique. Toutefois, nous ne voulons pas donner ici une place importante à cet aspect du traitement des données. Nous préférons nous focaliser sur le seul développement de l'ACP au contexte fonctionnel. Pour la suite de cette partie, nous supposerons que les données sont connues sous une forme adaptée à leur analyse et que les éventuels problèmes de recalage ont été réglés.

6.2 Statistique et données fonctionnelles

La nature de l'information extraite par une ACP repose sur des concepts statistiques simples tels que ceux de moyenne, variance, covariance. Ces concepts peuvent s'adapter au contexte fonctionnel d'une façon naturelle [RS97b, KG92] :

- La moyenne d'une famille de N applications $f_i : T \rightarrow \mathbb{R}$, $t \mapsto f_i(t)$ peut être définie comme la fonction qui à t associe la moyenne des f_i en ce point.

$$\bar{f}(t) = \frac{1}{N} \sum_i f_i(t) \quad (6.2)$$

Nous pouvons observer que cette définition trouve un grand intérêt pour des courbes recalées comme celles de la figure 6.5, alors que dans un cas comme celui illustré par la figure 6.4, la moyenne ainsi définie lisse l'information.

- De la même façon, la variance d'une famille de N applications $f_i : T \rightarrow \mathbb{R}$, $t \mapsto f_i(t)$ peut être définie comme la fonction qui à t associe la variance des f_i en ce point.

$$\text{var}_f(t) = \frac{1}{N} \sum_i (f_i(t) - \bar{f}(t))^2 \quad (6.3)$$

- Similairement, la covariance et la corrélation peuvent être définies comme suit :

$$\text{cov}_f(t_1, t_2) = \frac{1}{N} \sum_i (f_i(t_1) - \bar{f}(t_1))(f_i(t_2) - \bar{f}(t_2)) \quad (6.4)$$

$$\text{corr}_f(t_1, t_2) = \frac{\text{cov}_f(t_1, t_2)}{\sqrt{\text{var}_f(t_1) \text{var}_f(t_2)}} \quad (6.5)$$

- La covariance est habituellement étudiée entre deux variables. Etant donnée une seconde famille d'applications $g_i : T \rightarrow \mathbb{R}$, $t \mapsto g_i(t)$, nous définissons la covariance et la corrélation croisées par

$$\text{cov}_{fg}(t_1, t_2) = \frac{1}{N} \sum_i (f_i(t_1) - \bar{f}(t_1))(g_i(t_2) - \bar{g}(t_2)) \quad (6.6)$$

$$\text{corr}_{fg}(t_1, t_2) = \frac{\text{cov}_{fg}(t_1, t_2)}{\sqrt{\text{var}_f(t_1) \text{var}_g(t_2)}} \quad (6.7)$$

Ces concepts sont autant de jalons supplémentaires en commun entre les différents cadres de l'ACP.

6.3 Cadre d'analyse des données fonctionnelles

Transposons maintenant le cadre de l'ACP multivariée au contexte fonctionnel. Nous allons dans un premier temps nous restreindre au cas où seule une fonctionnelle réelle est considérée sur la population observée et nous étendrons cet environnement dans les sections suivantes. Dans ce cas particulier et en adoptant un point de vue purement calculatoire, nous pouvons présenter le passage du contexte de l'ACP multivariée à notre cadre comme une simple substitution d'un indice continu t à l'index discret j [RS97b]¹. Le nouveau “tableau” de données peut se noter comme suit :

$$F = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}$$

où f_i est la fonction mesurée sur l'individu i .

Comme dans le cadre classique, chaque individu est identifié à sa fonction f_i , l'associant à un point de l'espace $L_2(T)$. Comme dans le cadre classique, nous appelons cet espace l'espace des individus. Cet espace étant muni du produit scalaire défini ci-dessus, on peut définir une distance entre ses éléments. Chaque individu f_i est associé à un poids $p_i > 0$, l'ensemble des poids vérifiant $\sum_{i=1}^n p_i = 1$. On désigne par $N = \{(f_i, p_i), i = 1, \dots, n\}$ le nuage de points pondérés.

La fonctionnelle peut aussi être assimilée à un point de l'espace vectoriel $L_2^n(T)$, mais cela ne présente pas d'intérêt tant que l'on se restreint à l'étude d'une seule fonctionnelle sur la population. Nous reviendrons plus tard sur ces considérations. Toutes les structures nécessaires à la définition du problème de recherche de composantes principales sont identifiées. Passons maintenant à la position du problème dans notre cadre simplifié.

1. La justification de ceci se trouve notamment dans le possible recours à la discrétisation du problème pour la recherche de la solution.

Chapitre 7

Transport de l'ACP dans le cadre fonctionnel

Une des originalités de l'ACPF par rapport à l'ACP, c'est qu'elle peut être appliquée à une seule variable. Une telle analyse permet de déterminer les principaux modes de variabilité dans des données, modes qui fournissent alors une décomposition intéressante pour la compression des données par exemple. Cette potentialité est peut-être la plus connue, notamment dans le domaine du traitement de l'image où l'ACP est appliquée sous l'appellation de la transformation de Karhunen-Loeve [JAI89]. Cela constitue en fait presque un problème lorsqu'il s'agit de motiver son application à plusieurs variables simultanément : pourquoi faire une analyse qui offre des résultats d'une interprétation délicate (car inhabituels) lorsque que l'on peut appliquer le même protocole d'analyse à chaque variable séparément de façon à obtenir des résultats de nature familière... L'information cherchée n'est pas la même dans les deux cas. L'analyse multidimensionnelle permet un sondage des structures relationnelles qui existent entre les variables.

Cette potentialité d'analyse monodimensionnelle engendre une seconde difficulté dans une application multidimensionnelle de l'ACPF. En effet, dans une analyse monodimensionnelle, les modes de variation de la variable sont orthogonaux deux à deux. Ce fait se généralise au cas multidimensionnel d'une façon naturelle et pourtant anti-intuitive : Les modes de variations de l'ensemble des variables sont orthogonaux deux à deux, mais les modes de variations d'une variable extraits des modes de variation de l'ensemble des variables ne sont pas orthogonaux deux à deux ! Nous reviendrons sur ce fait lorsque des notations introduites en nombre suffisant permettront une description plus concise.

Ceci n'ôte pas à l'ACP monofonctionnelle ses qualités pour une introduction de l'ACP au cadre fonctionnel.

7.1 Problème dans le cas monofonctionnel

Nous laissons de côté pour l'instant les interprétations qui pourraient nous guider dans la formulation du problème de l'ACP fonctionnelle. Nous présenterons la démarche comme l'adaptation conceptuelle qui découle de l'ACP classique, ce qui nous permettra de découvrir la nature des résultats au fur et à mesure, comme dans une processus de découverte.

7.1.1 Formulation du problème

Considérons F une variable aléatoire du second ordre à valeur dans l'espace hilbertien séparable $L_2(T)$. Soit d un entier. Cherchons H un sous-espace de $L_2(T)$ de dimension d permettant d'obtenir une projection du nuage de fonctions dont l'inertie est maximale. Nous disposons des mêmes outils que dans le cadre de l'ACP multivariée : produit scalaire et distance. La formulation 2.5 proposée dans la partie II pour le problème de l'ACP est encore valable ici. Nous savons que le sous-espace

cherché passe par le centre de gravité du nuage, qui n'est autre que la fonction obtenue en moyennant les fonctions f_i , i.e. $\sum_i p_i f_i$. Supposons que le centre de gravité soit la fonction nulle. Le centrage des données permet de satisfaire cette hypothèse. Dans ce cas, le sous-espace cherché est le sous-espace vectoriel H satisfaisant

$$\max_H \sum_i p_i \|\widehat{f_i^H}\|^2 \quad (7.1)$$

i.e. dans $L_2(T)$

$$\max_H \sum_i p_i \int_T \widehat{f_i^H}^2 \quad (7.2)$$

Nous verrons dans le chapitre suivant que les propriétés de H permettent de décomposer ce problème en sous-problèmes à résoudre dans l'ordre donné :

1. Recherche de u_1 dans $L_2(T)$ tel que la quantité $\sum_i p_i \int_T (f_i u_1)^2$ soit maximale sous la contrainte $\int_T u_1^2 = \|u_1\|^2 = 1$
- ⋮
- d. Recherche de u_d dans $L_2(T)$ tel que la quantité $\sum_i p_i \int_T (f_i u_d)^2$ soit maximale sous les contraintes $\int_T u_d^2 = \|u_d\|^2 = 1$ et $\forall i < d, \int_T u_i u_d = \langle u_i, u_d \rangle = 0$

Nous retrouvons là une propriété déjà connue dans le cadre multivarié. Passons à présent au cas bifonctionnel.

7.2 Problème dans le cas bifonctionnel

Il est naturel de s'intéresser au cas où plusieurs fonctions par individu sont mesurées. Plaçons nous dans le cas bifonctionnel, où deux fonctionnelles F et G sont étudiées. Le $i^{\text{ème}}$ individu est identifié avec le 2-vecteur de fonctions qui lui correspondent :

$$\begin{pmatrix} f_i \\ g_i \end{pmatrix}$$

C'est un point de l'espace $L_2^2(T)$. Adjoignons à cet espace le produit scalaire suivant :

$$\forall \begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \in L_2^2(T), \langle \begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \rangle_{L_2^2(T)} = \langle x, u \rangle_{L_2(T)} + \langle y, v \rangle_{L_2(T)} = \langle x, u \rangle + \langle y, v \rangle$$

Dans le cas où l'espace H cherché est de dimension 1, alors le problème 7.2 s'écrit :

$$\max_{\|u\|+\|v\|=1} \sum_i p_i \left[\langle f_i, u_1 \rangle^2 + 2\langle f_i, u_1 \rangle \langle g_i, v_1 \rangle + \langle g_i, v_1 \rangle^2 \right]$$

Nous verrons plus tard que grâce aux propriétés de H , le problème général se décline en sous-problèmes comme suit :

1. Recherche de $\begin{pmatrix} u_1 \\ v_1 \end{pmatrix}$ dans $L_2^2(T)$ de norme 1 et tel que la quantité

$$\sum_i p_i \left[\langle f_i, u_1 \rangle^2 + 2\langle f_i, u_1 \rangle \langle g_i, v_1 \rangle + \langle g_i, v_1 \rangle^2 \right]$$

soit maximale

⋮

d. Recherche de $\begin{pmatrix} u_d \\ v_d \end{pmatrix}$ dans $L_2^2(T)$ vérifiant $\forall i \leq d$, $\langle \begin{pmatrix} u_i \\ v_i \end{pmatrix}, \begin{pmatrix} u_d \\ v_d \end{pmatrix} \rangle = \delta_{id}$ et tel que la quantité

$$\sum_i p_i \left(\sum_{j=1}^d \langle \begin{pmatrix} f_i \\ g_i \end{pmatrix}, \begin{pmatrix} u_j \\ v_j \end{pmatrix} \rangle \right)^2$$

soit maximale

Avant de passer au cas général, revenons sur notre remarque portant sur l'orthogonalité des modes de variation. Les contraintes des sous-problèmes ci-dessus peuvent se résumer par

$$\forall i, j \in \llbracket 1, d \rrbracket, \langle \begin{pmatrix} u_i \\ v_i \end{pmatrix}, \begin{pmatrix} u_j \\ v_j \end{pmatrix} \rangle = \delta_{ij} \quad (7.3)$$

A chaque étape, le nouveau vecteur de fonction $\begin{pmatrix} u_i \\ v_i \end{pmatrix}$ est orthogonal aux autres. Mais ceci n'apporte aucune information sur les quantités $\langle u_i, u_j \rangle$. De ce fait, une certaine redondance peut s'infiltre dans les "sous-modes" de variation. Ainsi les décompositions obtenues pour une variables seule n'ont pas de propriétés particulières en termes d'orthogonalité.

7.3 Problème dans le cas multifonctionnel

Supposons que p fonctionnelles, notées F_1, \dots, F_p , soient mesurées sur notre population de n individus. Chaque individu est entièrement caractérisé par le p -vecteur de fonctions qui lui correspond :

$$\begin{pmatrix} f_{1i} \\ \vdots \\ f_{pi} \end{pmatrix}$$

Ainsi, l'information recueillie sur notre population est un nuage de points pondérés de $L_2^n(T)$. Sur cet espace, nous définissons le produit scalaire suivant :

$$\forall \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \in L_2^n(T), \langle \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \rangle_{L_2^n(T)} = \langle x_1, y_1 \rangle_{L_2(T)} + \cdots + \langle x_p, y_p \rangle_{L_2(T)}$$

Le problème général se décompose alors :

1. Recherche de $\begin{pmatrix} u_1^1 \\ \vdots \\ u_p^1 \end{pmatrix}$ dans $L_2^n(T)$ tel que la quantité $\sum_i p_i \langle \begin{pmatrix} u_1^1 \\ \vdots \\ u_p^1 \end{pmatrix}, \begin{pmatrix} f_{1i} \\ \vdots \\ f_{pi} \end{pmatrix} \rangle^2$ soit maximale sous la contrainte $\|u_1^1\|^2 + \cdots + \|u_p^1\|^2 = 1$

\vdots

d. Recherche de $\begin{pmatrix} u_1^d \\ \vdots \\ u_p^d \end{pmatrix}$ dans $L_2^n(T)$ vérifiant $\forall i \leq d$, $\langle \begin{pmatrix} u_1^i \\ \vdots \\ u_i^d \end{pmatrix}, \begin{pmatrix} u_1^d \\ \vdots \\ u_p^d \end{pmatrix} \rangle = \delta_{id}$ tel que la quantité $\sum_i p_i \langle \begin{pmatrix} u_1^d \\ \vdots \\ u_p^d \end{pmatrix}, \begin{pmatrix} f_{1i} \\ \vdots \\ f_{pi} \end{pmatrix} \rangle^2$ soit maximale

Chapitre 8

Solution dans le cas monofonctionnel

Revenons sur le cas monofonctionnel. Nous allons traiter dans un premier temps le cas pour lequel nous disposons d'une base de l'espace des individus. Ce cas particulier est intéressant car il permet de retomber sur la formulation classique du problème de l'ACP et la recherche de solution se fait donc selon le procédé décrit à la section 2.7. Nous présenterons ensuite la solution théorique du problème puis nous nous intéresserons à l'approche par discréétisation qui fournit probablement le moyen le plus simple d'effectuer une ACPF. Nous nous attarderons enfin sur la nature des résultats obtenus dans le cadre monofonctionnel afin de percevoir la nature de l'information extraite par notre analyse dans ce cas particulier.

8.1 Réduction du problème par l'emploi d'une base de l'espace des individus

Supposons que nous disposions d'une base $\{\phi_n\}$ orthonormée de l'espace des valeurs de la variable aléatoire fonctionnelle considérée. Il est alors possible d'exprimer la fonction de chaque individu comme une combinaison linéaire des fonctions de base. Le problème est que ces combinaisons linéaires sont en général infinies. Il est possible d'obtenir des approximations de ces fonctions avec un sous-ensemble fini des fonctions de bases, pour un seuil de qualité choisi. Etant donné un seuil de qualité, supposons un tel sous-ensemble connu et dénoté K . Identifions à présent chaque fonction individuelle à son approximation. Posons

$$f_i = \sum_{k \in K} c_{ik} \phi_k$$

i.e. sous forme matricielle :

$$\begin{bmatrix} \vdots \\ f_i \\ \vdots \end{bmatrix} = \begin{bmatrix} & & \vdots \\ \cdots & c_{ik} & \cdots \\ & & \vdots \end{bmatrix} \begin{bmatrix} \vdots \\ \phi_k \\ \vdots \end{bmatrix}$$

En tenant compte de la base orthonormée, la première étape du problème consiste en la recherche de $u_1 = \sum_{k \in K} u_{1k} \phi_k$ dans l'espace engendré par les fonctions de base limitées par K telle que la quantité $\sum_i p_i \int_T f_i u_1$ soit maximale sous la contrainte $\int_T u_1^2 = 1$. Mais d'une part

$$\sum_i p_i \int_T (f_i u_1)^2 = \sum_i p_i \sum_k c_{ik}^2 u_{1k}^2$$

ce que nous pouvons écrire de façon matricielle

$$\begin{bmatrix} \vdots \\ u_{1i} \\ \vdots \end{bmatrix}' \begin{bmatrix} \cdots & c_{ik} & \cdots \\ \cdots & \vdots & \cdots \end{bmatrix}' \begin{bmatrix} \ddots & & \\ & p_i & \\ & & \ddots \end{bmatrix} \begin{bmatrix} \cdots & c_{ik} & \cdots \\ \cdots & \vdots & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ u_{1i} \\ \vdots \end{bmatrix}$$

D'autre part,

$$\int_T u_1^2 = 1 \text{ s'écrit } \sum_i u_{1i}^2 = 1$$

Nous retrouvons ainsi la formulation exacte du problème de l'ACP classique pour la recherche du premier axe principal. La solution correspond à la fonction qui est combinaison linéaire des fonctions de base limitées à K et qui a pour coefficients ceux du vecteur donnant l'axe principal. Il est facile de montrer que les étapes suivantes conduisent à la suite de la formulation de l'ACP classique avec pour données les coefficients de développement des fonctions individuelles sur les fonctions de base.

Voici donc une situation pour laquelle il est trivial de se ramener au cadre classique de l'ACP afin de pratiquer une ACPF. Ceci vient d'être montré dans le cas monofonctionnel. Le cas multifonctionnel permet la même adaptation dans le cas où une base de l'espace produit est choisie. L'approche par discréétisation présentée ci-dessous permet, elle aussi, de ramener la recherche des composantes principales fonctionnelles à la recherche de composantes principales d'une certaine matrice.

8.2 Solution théorique

Comme nous le citions en introduction, J.KLEFFE a prouvé dans [KLE73] que toute variable aléatoire à valeur dans un espace hilbertien séparable pouvait être décomposée en composantes principales. Il est bien entendu parvenu à ce résultat en définissant les composantes principales comme les solutions de notre problème, généralisation du problème de l'ACP au cadre fonctionnel. Ces composantes principales fonctionnelles possèdent beaucoup de propriétés qui étaient vérifiées par leurs analogues dans le cadre classique. C'est le cas notamment de la pluralité de leur caractérisation. Présentées ici comme solution d'un problème d'optimisation, elles se définissent également comme les éléments propres d'un certain opérateur. Cette équivalence des définitions repose sur l'équivalence qui existe entre les deux problèmes généraux suivants :

$$\max_{\|x\|=1} \langle x, Ax \rangle \tag{8.1}$$

$$Ax = \lambda x \tag{8.2}$$

Cette équivalence bien connue dans les espaces euclidiens, et que nous avons utilisée dans la partie précédente, s'étend aux espaces hilbertiens séparables [RN56].

L'opérateur dont les composantes sont éléments propres est l'opérateur de covariance V dont la caractérisation par rapport à la variable étudiée peut être consultée dans [BES91]. Nous ne donnons pas plus de détails sur l'opérateur V et sur les différentes formulations du problème, préférant donner la priorité à l'étude des moyens pratiques de recherche des composantes. Une présentation détaillée de la résolution théorique du problème peut être trouvée dans [KLE73, AAV99]. Retenons simplement la solution générale de notre problème. Notons (λ_k) les valeurs propres de V et (u_k) les vecteurs propres associés. Supposons que les valeurs propres soient numérotées dans l'ordre décroissant de leurs valeurs. Etant donné d , le sous-espace H de $L_2(T)$ et de dimension d maximisant l'inertie de la projection du nuage de courbes est la somme directe des sous-espaces de dimension 1 engendrés par d vecteurs propres correspondant aux d plus grandes valeurs propres de V :

$$H = \Delta u_1 \oplus \Delta u_2 \oplus \cdots \oplus \Delta u_d \tag{8.3}$$

et l'inertie de H est donnée par

$$I(H) = \lambda_1 + \lambda_2 + \cdots + \lambda_d \tag{8.4}$$

8.3 Approximation de la solution par discréétisation du problème

[BES91] établit la validité du recours à la discréétisation pour la recherche d'approximation des composantes principales, et cela est plutôt réconfortant. En effet, des données issues d'expérimentation ne sont jamais connues autrement que de façon discrète et ce sont principalement des approximations des fonctions qui sont candidates à l'ACP. BESSE insère son propos dans le cadre de l'utilisation des splines comme moyen d'approximation. La convergence de l'ajustement spline vers les composantes est prouvé.

D'une façon générale, l'abordage d'un problème continu par le biais d'une approximation obtenue par une discréétisation est une pratique courante dans l'étude des problèmes fonctionnels. Prenons le cas simple d'une variable fonctionnelle aléatoire à valeurs dans $L_2(T)$. Etant donnée une grille de l'intervalle T faite de n valeurs régulièrement espacées afin de discréétiser les fonctions observées, nous considérons les fonctions individuelles par le biais des vecteurs de valeurs de ces fonctions en les points de la grille. Cela conduit à une $N \times n$ -matrice X de données. Notons D la matrice diagonale dont le $i^{\text{ème}}$ est la pondération p_i associée à l'individu i . Le produit scalaire est défini par la matrice symétrique M dont les termes sont les coefficients du schéma d'intégration numérique choisi pour approcher les intégrales $\int_T f(t)g(t)dt$.

Prenons l'exemple du schéma d'intégration basé sur les trapèzes, comme illustré sur la figure 8.1, et adapté à une fonction f connue aux points x_1 à x_n .

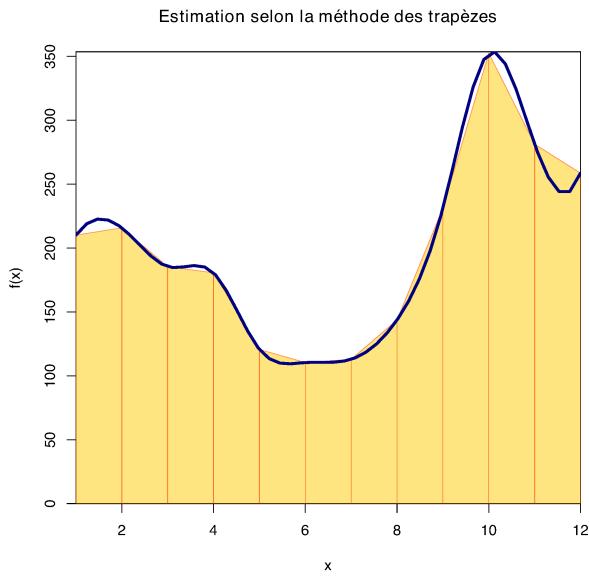


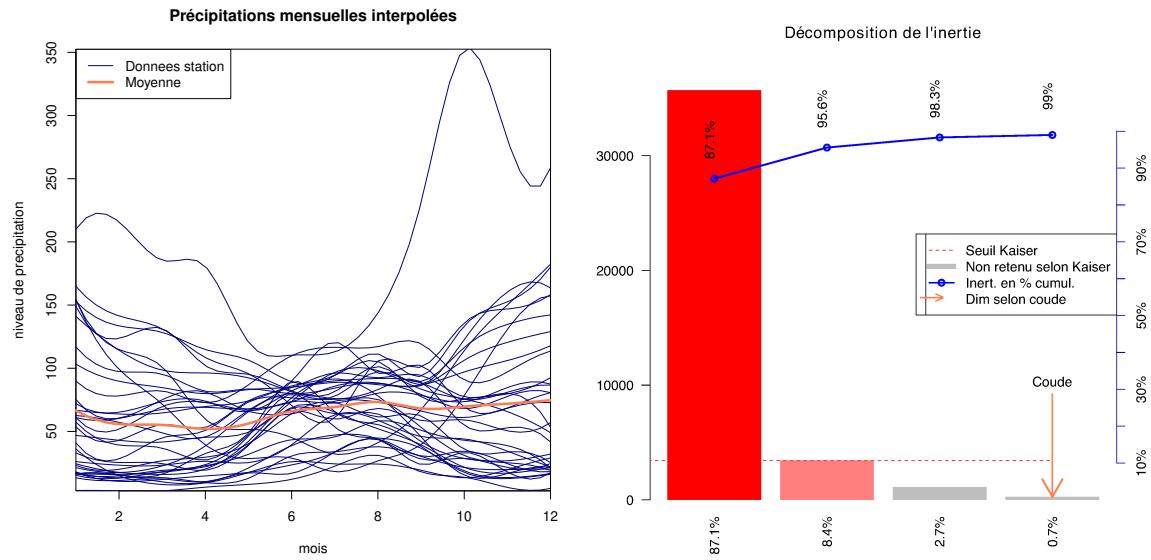
FIGURE 8.1 – Schéma d'intégration basé sur les trapèzes

$$\text{La matrice } D \text{ associée est} \begin{bmatrix} \frac{x_2-x_1}{2} & & & & \\ & \frac{x_3-x_1}{2} & & & \\ & & \ddots & & \\ & & & \frac{x_n-x_{n-2}}{2} & \\ & & & & \frac{x_n-x_{n-1}}{2} \end{bmatrix}$$

L'opérateur de covariance peut alors être approché par la matrice $N^{-1}(X - \bar{X})' D(X - \bar{X})$ comme dans le cadre classique de l'ACP [RS97b] et notre problème se ramène donc à nouveau au cadre classique de l'ACP. La détermination des éléments propres de $(X - \bar{X})' D(X - \bar{X})M$ permet d'approcher les composantes principales cherchées.

8.4 Application

Nous proposons dès à présent un court aperçu des résultats produits par l'ACPF sur un échantillon de données d'une variable aléatoire fonctionnelle. Ces données sont tirées du site indiqué plus bas¹. Elles correspondent à des mesures des précipitations relevées en différentes localités au cours d'une période de temps. Les variations du niveau de précipitation en chaque lieu sont données par les



(a) Un ensemble de courbes de précipitation et leur moyenne illustrée en gras (b) Valeurs propres et pourcentage de l'inertie associé

FIGURE 8.2 – Courbes et valeurs propres

discrétisations dérivant de leurs mesures ponctuelles. Les courbes reconstruites apparaissent sur la figure 8.2a. La moyenne de ces courbes est indiquée en gras et avec une couleur distincte, comme indiqué par la légende.

Pour la plupart des courbes représentées, une bonne approximation peut être fournie par la somme de la moyenne et d'une combinaison linéaire bien choisie de fonctions. L'ACPF appliquée aux courbes centrées révèle que les deux premières composantes principales u_1 et u_2 , dont les variations sont données par les figures 8.3a et 8.3b permettent une restitution de 95% de l'inertie de la famille de courbes, avec 87% pour u_1 et 8% pour u_2 (figure 8.2b). Les coefficients des décompositions des courbes centrées sur l'espace vectoriel porté par u_1 et u_2 conduit, comme dans le cadre d'une ACP classique, à une représentation des individus dans le plan porté par u_1 et u_2 (figure 8.4). A partir de ces premiers éléments, que pouvons-nous retenir ? Comment construire notre interprétation de ces résultats ? L'attention doit d'abord être portée sur la moyenne des précipitations dans le temps : sachant que l'intervalle de temps correspond à l'année, et qu'il débute en janvier, le niveau des précipitations semble le plus bas au printemps, et plus fort en été. En hiver, une petite remontée s'observe. Passons ensuite au premier mode de variation. La fonction qui lui est associée à des valeurs négatives. Il ne faut pas perdre de vue que le signe de ces valeurs n'a pas d'importance, car nous sommes intéressés par l'espace vectoriel de dimension 1 défini par cette fonction. L'écart que crée cette fonction engendre une variation, par rapport à la moyenne, qui est importante en début d'année et en automne. Ces variations vont dans le même sens. Comment cela se traduit pour les individus ? D'après la position de *thunderb* sur la figure 8.4, les précipitations seront plus faibles par rapport à la moyenne en hiver et au printemps et leur niveau se rapprochera progressivement des moyennes saisonnières en dehors de ces variations. Pour *yarmouth*, c'est l'inverse qui se produit. Les précipitations sont accentuées au prin-

1. <http://www.psych.mcgill.ca/misc/fda/downloads/FDAfun/>

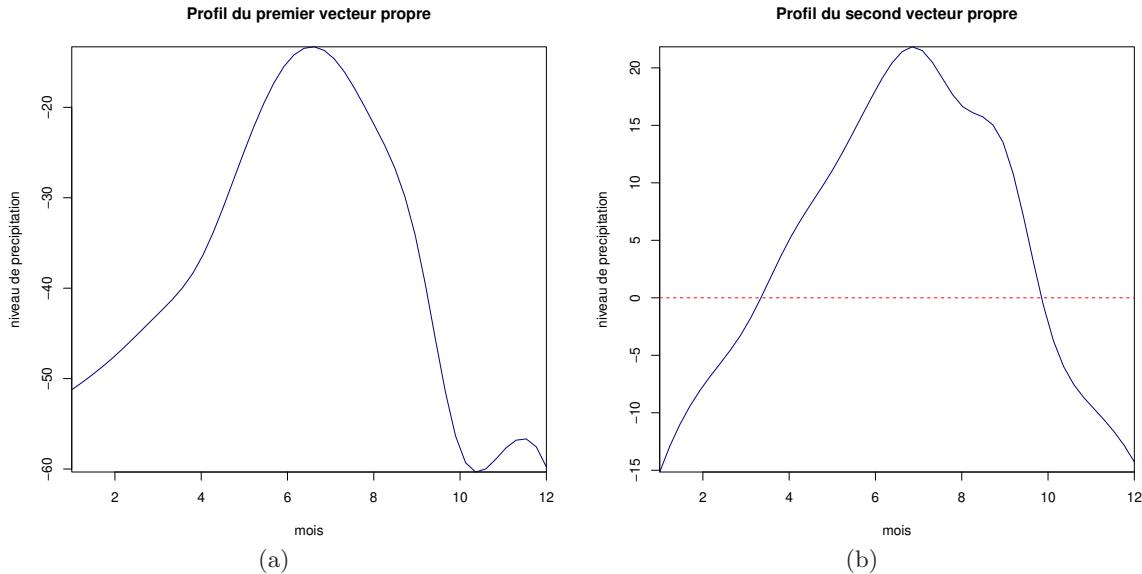


FIGURE 8.3 – Premier et second modes de variation

temps et en hiver. La seconde composante principale correspond à une perturbation de la moyenne qui tend à accentuer l'écart observé entre l'été d'une part et le printemps et l'hiver d'autre part. Pour un individu comme *yarmouth* dont la décomposition comporte une composante positive pour la fonction représentée par la figure 8.3b, les précipitations en été sont plus fortes que la moyenne et plus faibles en hiver.

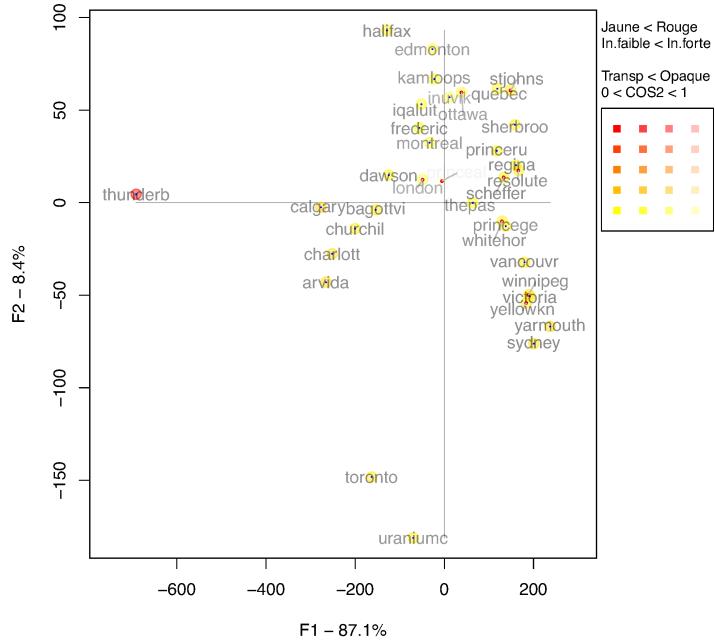


FIGURE 8.4 – Projections des individus dans le plan porté par u_1 et u_2

Les individus choisis ci-dessus peuvent ne pas être les plus représentatifs, malgré leur apparentes positions extrêmales. Il faudrait s'assurer de la qualité de représentation de leur projection. Comme pour une ACP classique, il est souhaitable de calculer pour chaque individu un certain nombre d'indicateurs, apportant plus de fiabilité que les graphiques présentés. Mais quels indicateurs...? Les

indicateurs proposés pour l'ACP s'utilisent ici sans adaptation particulière. Signalons par ailleurs que [RS97b] proposent d'autres développements graphiques, notamment pour l'analyse de la dynamique des composantes principales : la visualisation de la moyenne accompagnée de la même moyenne diminuée et augmentée selon chaque mode de variation principal rend très bien compte des effets de ces types d'écart.

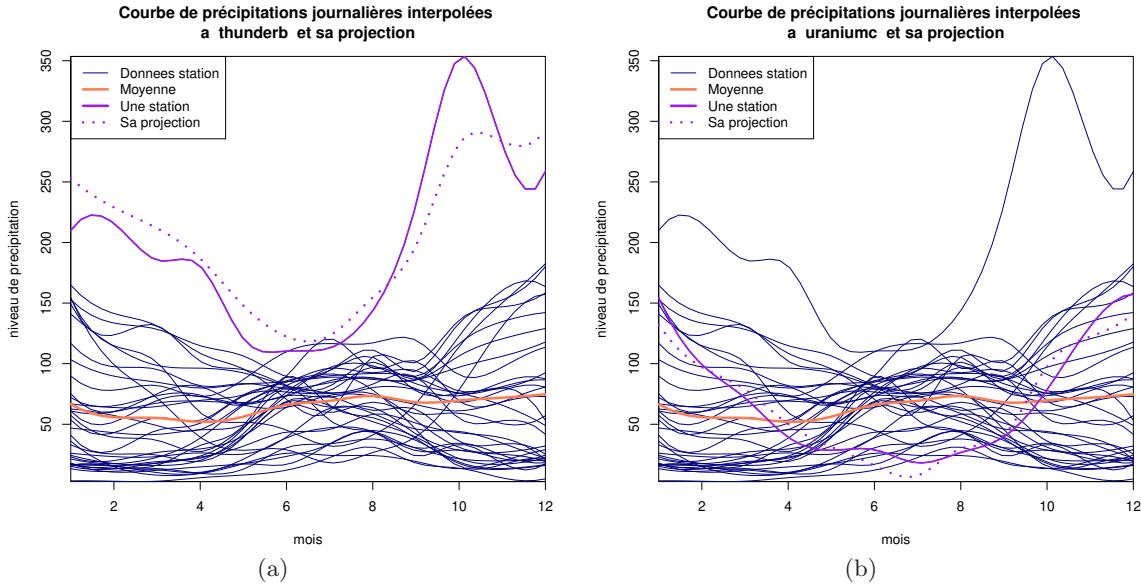


FIGURE 8.5 – Projections des stations de Thunderb et Uraniumc

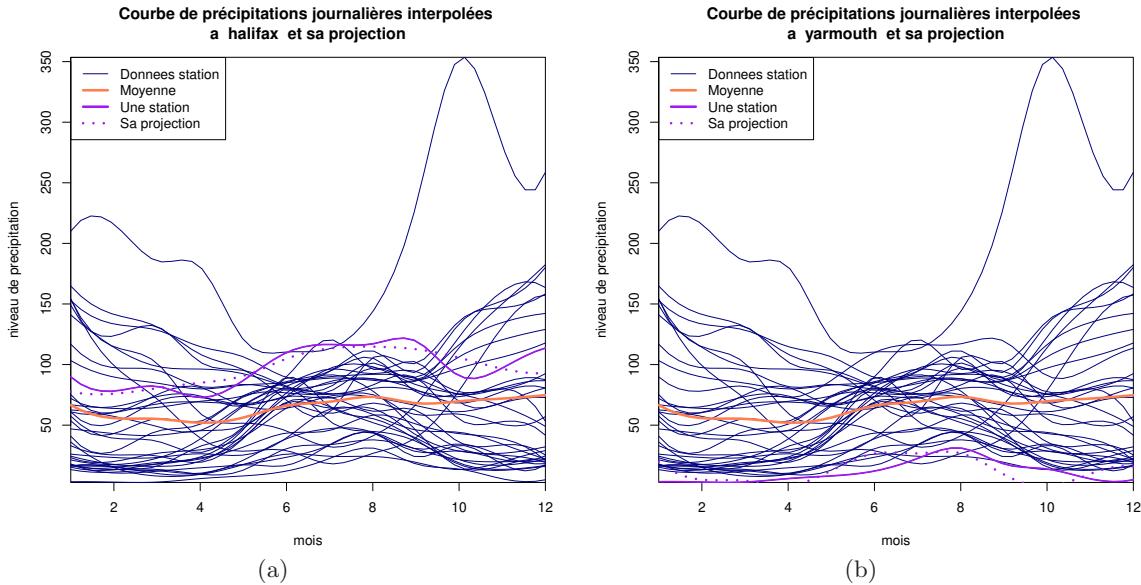


FIGURE 8.6 – Projections des stations de Halifax et Yarmouth

Cette courte application montre la grande similarité d'analyse existant entre les deux cadres d'application de l'ACP : la recherche des valeurs propres permet de déterminer la dimension qu'il est intéressant de retenir pour H , puis les individus peuvent être observés et comparés dans le sous-espace de projection. Les nouveautés sont d'une part l'absence de projection dans l'espace des variables, ce qui semble normal, puisqu'une seule variable est étudiée et d'autre part la présence de graphiques

détaillés des composantes principales. Ces graphiques font cohabiter les aspects statiques (projections des individus) et dynamique (modes de variations) présents dans les courbes. Nous allons à présent voir comment ce protocole bien ficelé se laisse complètement déborder dans le cas multifonctionnel, ce qui justifiera l'enrichissement du processus d'analyse pour une meilleure synthétisation de l'information.



Appendices

Annexe A

Preuves de la partie ACP multivariée dans le cadre euclidien simplifié

A.1 Relation de Huygens

Montrons que $M^t(H) = M^t(H_g) + d^2(H, H_g)$ en utilisant le fait que $\widehat{x_i^H} = \widehat{x_i^H} + (\widehat{g^H} - \widehat{g^H})$:

$$\begin{aligned}
M^t(H) &= \sum_i p_i d^2(x_i, \widehat{x_i^H}) \\
&= \sum_i p_i \left\langle x_i - \widehat{x_i^H} + (\widehat{g^H} - \widehat{g^H}), x_i - \widehat{x_i^H} + (\widehat{g^H} - \widehat{g^H}) \right\rangle \\
&= \sum_i p_i \left\langle x_i - \widehat{x_i^H}, x_i - \widehat{x_i^H} \right\rangle + 2 \sum_i p_i \left\langle x_i - \widehat{x_i^H}, \underbrace{\widehat{g^H} - \widehat{g^H}}_g \right\rangle + \left\| \underbrace{\widehat{g^H} - \widehat{g^H}}_g \right\|^2 \\
&= \sum_i p_i \left\langle x_i - \widehat{x_i^H}, x_i - \widehat{x_i^H} \right\rangle + 2 \left\langle \sum_i p_i x_i - \sum_i p_i \widehat{x_i^H}, g - \widehat{g^H} \right\rangle + \|g - \widehat{g^H}\|^2 \\
&= \sum_i p_i \left\langle x_i - \widehat{x_i^H}, x_i - \widehat{x_i^H} \right\rangle + 2 \underbrace{\left\langle g - \underbrace{\widehat{g^H}}_0, g - \widehat{g^H} \right\rangle}_{+ \|g - \widehat{g^H}\|^2} + \|g - \widehat{g^H}\|^2 \\
M^t(H) &= M^t(H_g) + d^2(H, H_g)
\end{aligned}$$

A.2 $Q_H = 2I_g - 2M^t(H)$

Montrons que $Q_H = 2I_g - 2M^t(H)$ en utilisant la relation de Huygens, et où $I_g = \sum_i p_i d(x_i, g)^2$:

$$\begin{aligned}
 \sum_i \sum_j p_i p_j d(\widehat{x_i^H}, \widehat{x_j^H})^2 &= \sum_j p_j \sum_i p_i d(\widehat{x_i^H}, \widehat{x_j^H})^2 \\
 &= \sum_j p_j \left[\sum_i p_i d(\widehat{x_i^H}, \widehat{g^H})^2 + d(\widehat{g^H}, \widehat{x_j^H})^2 \right] \\
 &= \sum_i p_i d(\widehat{x_i^H}, \widehat{g^H})^2 + \sum_j p_j d(\widehat{g^H}, \widehat{x_j^H})^2 \\
 &= 2 \sum_i p_i d(\widehat{x_i^H}, \widehat{g^H})^2 \\
 &= 2 \sum_i p_i \left[d(x_i, \widehat{g^H})^2 - d(x_i, \widehat{x_i^H})^2 \right] \\
 &= 2 \sum_i p_i d(x_i, g)^2 + 2d(g, \widehat{g^H})^2 - 2 \sum_i p_i d(x_i, \widehat{x_i^H})^2 \\
 &= 2 \sum_i p_i d(x_i, g)^2 + 2 \left[\underbrace{d(g, \widehat{g^H})^2}_{d(H, H_g)^2} - \underbrace{\sum_i p_i d(x_i, \widehat{x_i^H})^2}_{M^t(H)} \right] \\
 Q_H &= 2I_g - 2M^t(H_g)
 \end{aligned}$$

A.3 Positivité de la matrice d'inertie du nuage

Nous avons

$$V = \begin{pmatrix} & \vdots & \\ \cdots & \sum_k x_k^i p_k x_k^j & \cdots \\ & \vdots & \end{pmatrix}$$

D'où, $\forall p = 1, \dots, p$

$$e_p' V e_p = \sum_k x_k^p p_k x_k^p = \sum_k p_k (x_k^p)^2 \geq 0$$

A.4 $I(H) = \text{Tr}(VP)$

Etant donné un sous-espace vectoriel H de dimension d , $\{u_1, \dots, u_d\}$ une base orthonormée de H et P une matrice de la projection orthogonale sur H , montrons que $I(H) = \text{Tr}(VP)$ où $V = X'DX$:

- Les colonnes de P sont $\widehat{e_i^H} = \sum_{l=1}^d \langle e_i, u_l \rangle u_l = \sum_{l=1}^d u_l u_l$.
- D'une part, nous avons

$$\begin{aligned}
 I(H) &= \sum_{i=1}^n p_i \sum_{l=1}^d \langle x_i, u_l \rangle^2 = \sum_{i=1}^n p_i \sum_{l=1}^d u_l' x_i x_i' u_l = \sum_{l=1}^d u_l' \left(\sum_{i=1}^n p_i x_i x_i' \right) u_l \\
 I(H) &= \sum_{l=1}^d u_l' X' D X u_l
 \end{aligned}$$

— D'autre part, nous observons que

$$\begin{aligned} \text{Tr}(VP) &= \sum_{i=1}^p V'_i P^i = \sum_{i=1}^p V'_i \sum_{l=1}^d u_{l_i} u_l = \sum_{i=1}^p \sum_{l=1}^d u_{l_i} V'_i u_l \\ &= \sum_{l=1}^d \left(\sum_{i=1}^p u_{l_i} V'_i \right) u_l \\ \text{Tr}(VP) &= \sum_{l=1}^d u'_l V u_l \end{aligned}$$

D'où le résultat.

A.5 $I(H_1 \oplus H_2) = I(H_1) + I(H_2)$

Etant donnés deux sous-espaces vectoriels orthogonaux H_1 et H_2 , montrons que

$$I(H_1 \oplus H_2) = I(H_1) + I(H_2)$$

Il suffit de constater que $\|\widehat{x^{H_1 \oplus H_2}}\|^2 = \|\widehat{x^{H_1}}\|^2 + \|\widehat{x^{H_2}}\|^2$.

A.6 Solution de $\max_{\sum_i a_i^2 = 1} \sum_{i=1}^p \lambda_i a_i^2$

— Borne supérieure de $\max_{\sum_i a_i^2 = 1} \sum_{i=1}^p \lambda_i a_i^2$:

$$\sum_{i=1}^p \underbrace{\lambda_i}_{\leq \lambda_1} a_i^2 \leq \lambda_1 \sum_{i=1}^p a_i^2 \quad (\text{A.1})$$

$$\sum_{i=1}^p \lambda_i a_i^2 \leq \lambda_1 \quad (\text{A.2})$$

— Cette borne est atteinte pour $a_1 = 1$ et $a_i = 0, \forall i \neq 1$, i.e. pour $u = u_1$.

A.7 Lemme fondamental

Soit $k < p$; si F_k est le sous-espace vectoriel de dimension k d'inertie maximale, alors le sous-espace vectoriel de dimension $k+1$ d'inertie expliquée maximale est $F_{k+1} = F_k \oplus \Delta u$, où Δu est la droite vectorielle orthogonale à F_k d'inertie expliquée maximale.

En effet, soit E_{k+1} un sous-espace vectoriel de dimension $k+1$. Comme $\dim(F_k^\perp) = n-k$, $\dim(F_k^\perp \cap E_{k+1}) \geq 1$. Soit b dans $F_k^\perp \cap E_{k+1}$, avec $\|b\| = 1$, et G tel que $E_{k+1} = \Delta_b \oplus G$. Posons $F = \Delta_b \oplus F_k$. Nous avons

$$I(E_{k+1}) = I(\Delta_b) + I(G) \quad (\text{A.3})$$

$$I(F) = I(\Delta_b) + I(F_k) \quad (\text{A.4})$$

Or $I(F_k) \geq I(G)$, donc $I(F) \geq I(E_{k+1})$.

Maintenant, choisissons b dans F_k^\perp tel que $I(\Delta_b)$ soit maximal; on voit que $b = \pm u_{k+1}$. Posons $F_{k+1} = \Delta_b \oplus F_k$.

$$I(F_{k+1}) = I(\Delta_{u_{k+1}}) + I(F_k) \quad (\text{A.5})$$

Il est clair que

$$\forall b \in F_k^\perp, I(\Delta_b) \leq I(\Delta_{u_{k+1}}) \quad (\text{A.6})$$

donc $I(F_{k+1}) \geq I(E_{k+1})$.

Annexe B

ACP multivariée dans le cadre euclidien général

Dans le cadre euclidien général, le produit scalaire est défini à partir d'une matrice M symétrique, définie positive et de rang égal à la dimension de l'espace :

$$\forall x, y \in \mathbb{R}^p, \langle x, y \rangle = x' M y$$

Avant d'exposer la solution dans ce cadre, nous allons rappeler quelques définitions et des résultats classiques, transposés à notre nouveau contexte.

B.0.1 Rappels

Définition B.0.1 Soit A une matrice carrée et M une matrice symétrique définie positive. On dit que A est M -symétrique si elle vérifie

$$MA = A'M$$

Théorème B.0.1 Soit A une matrice M -symétrique. Alors A est diagonalisable, de valeurs propres réelles. Par ailleurs, les sous-espaces propres sont deux à deux M -orthogonaux.

B.0.2 Cas où $d = 1$

Plaçons nous dans le cas où nous cherchons la direction d'un sous-espace affine de dimension 1 passant par g . Etant donné un vecteur u , avec $\|u\| = 1$, notons Δu la droite portée par u et passant par g .

Expression de l'inertie expliquée par Δu

Nous avons

$$\begin{aligned} I(\Delta u) &= \sum_i p_i \left\| \widehat{x_i^{\Delta u}} - g \right\|^2 = \sum_i p_i \langle x_i - g, u \rangle^2 = \sum_i p_i u' M (x_i - g) (x_i - g)' M u \\ &= u' M \left(\sum_i p_i (x_i - g) (x_i - g)' \right) M u \\ I(\Delta u) &= u' M V M u \end{aligned}$$

où selon nos notations, $V = Y'DY$.

Détermination du premier axe principal d'inertie

Nous cherchons un vecteur u de \mathbb{R}^p , M -normé, tel que l'inertie expliquée par la droite Δu soit maximale, cette inertie ayant pour expression :

$$I(\Delta u) = u' M V M u \quad (\text{B.1})$$

La $p \times p$ -matrice VM étant M -symétrique, elle est diagonalisable et il existe une base de vecteurs propres M -orthogonaux $\{u_1, \dots, u_p\}$. Par ailleurs, il est rapide de vérifier que ses valeurs propres sont positives : soit $u \in \mathbb{R}^p$. D'une part $u' M V M u = (Mu)' V (Mu) \geq 0$ car V est positive, d'autre part, pour tout vecteur propre v de VM , $v' M V M v = v' M \lambda v = \lambda \|v\|^2$. Donc $\lambda \|v\|^2 \geq 0$ i.e. $\lambda \geq 0$.

Notant les valeurs propres $\{\lambda_1, \dots, \lambda_p\}$, nous pouvons supposer, à une permutation près, que $\lambda_1 \geq \dots \geq \lambda_p$. Ainsi u_1 correspond à la valeur propre la plus grande, etc.

Posons $u = \sum_{i=1}^p a_i u_i$. Donnons une nouvelle formulation à notre problème B.1 :

$$\begin{aligned} I(\Delta u) &= u' M V M u \\ &= \sum_{i=1}^p a_i u_i' M \left(VM \sum_{j=1}^p a_j u_j \right) \\ &= \sum_{i=1}^p a_i u_i' M \left(\sum_{j=1}^p a_j \lambda_j u_j \right) \\ &= \sum_{i=1}^p a_i \langle u_i, \sum_{j=1}^p a_j \lambda_j u_j \rangle \\ &= \sum_{i=1}^p a_i^2 \lambda_i \\ I(\Delta u) &= \sum_{i=1}^p \lambda_i a_i^2 \end{aligned}$$

De plus

$$\begin{aligned} u' M u = 1 &\Leftrightarrow \sum_{i=1}^p a_i u_i' M \sum_{j=1}^p a_j u_j = 1 \\ &\Leftrightarrow \sum_{i=1}^p a_i \langle u_i, \sum_{j=1}^p a_j u_j \rangle = 1 \\ u' M u = 1 &\Leftrightarrow \sum_{i=1}^p a_i^2 = 1 \end{aligned}$$

Notre problème s'écrit donc :

$$\max_{\sum_i a_i^2 = 1} \sum_{i=1}^p \lambda_i a_i^2 \quad (\text{B.2})$$

auquel correspond la solution $a_1 = 1$, i.e. $u = u_1$.

preuve : voir A.6 page 88.

Ainsi, la droite Δu recherchée est portée par un vecteur propre associé à la plus grande valeur propre. Δu est appelée premier axe principal d'inertie.

B.0.3 Cas général

Lemme fondamental

Lemme B.0.1 Soit $k < p$; si F_k est le sous-espace affine de dimension k d'inertie maximale, alors le sous-espace affine de dimension $k+1$ d'inertie expliquée maximale est $F_{k+1} = F_k \oplus \Delta u$, où Δu est la droite affine orthogonale à F_k d'inertie expliquée maximale.

preuve : voir A.7 page 88.

Ainsi, la recherche des sous-espaces F_k s'effectue de proche en proche :

— pour déterminer la droite $F_1 = \Delta u_1$, on cherche le vecteur u_1 solution du problème :

$$\max_{\|u\|=1} I(\Delta u)$$

— pour déterminer le plan F_2 , on sait que $F_2 = \Delta u_1 \oplus \Delta u_2$ où Δu_2 est solution du problème :

$$\max_{\substack{\|u\|=1 \\ u \perp u_1}} I(\Delta u)$$

— etc...

Les droites $\Delta u_1, \Delta u_2, \dots$ sont appelées premier, second, ... axe principal d'inertie du nuage. Il arrive dans la littérature que l'appellation axe principal d'inertie soit attribuée aux vecteurs u_1, u_2, \dots

Détermination du k-ième axe principal d'inertie

Par récurrence, il est facile de montrer que le k-ième axe principal d'inertie est porté par le k-ième vecteur propre de VM .

En conséquence, nous avons $I(\Delta u_k) = \lambda_k$.

Solution du problème général

Finalement, la solution au problème $\max_{\substack{H \\ \dim(H)=k}} I(H)$ est donnée par :

$$F_k = \Delta u_1 \oplus \Delta u_2 \oplus \cdots \oplus \Delta u_k \quad (\text{B.3})$$

avec

$$I(F_k) = \lambda_1 + \lambda_2 + \cdots + \lambda_k \quad (\text{B.4})$$

Nous avons ainsi déterminé, pour une dimension donnée, la direction des sous-espaces maximisant l'inertie expliquée et l'inertie que chacun d'eux porte.

Annexe C

Preuves de la partie ACP multivariée dans le cadre euclidien général & données de l'analyse budget/temps de Michel Jambu

C.1 expression matricielle d'une projection M -orthogonale

Soit E un K -ev et F un sous ev de E .

C.2 $I(H) = \text{Tr}(VP)$

Etant donné un sous-espace vectoriel H de dimension d , $\{u_1, \dots, u_d\}$ une base orthonormée de H et P une matrice de la projection orthogonale sur H , montrons que $I(H) = \text{Tr}(VP)$ où $V = MX'DXM$:

- Les colonnes de P sont $\widehat{e_i^H} = \sum_{l=1}^d \langle e_i, u_l \rangle u_l = \sum_{l=1}^d u_l u_l$.
- D'une part, nous avons

$$\begin{aligned} I(H) &= \sum_{i=1}^n p_i \sum_{l=1}^d \langle x_i, u_l \rangle^2 = \sum_{i=1}^n p_i \sum_{l=1}^d u_l' M x_i x_i' M u_l = \sum_{l=1}^d u_l' M \left(\sum_{i=1}^n p_i x_i x_i' \right) M u_l \\ I(H) &= \sum_{l=1}^d u_l' M X' D X M u_l \end{aligned}$$

- D'autre part, nous observons que

$$\begin{aligned} \text{Tr}(VP) &= \sum_{i=1}^p V_i' P^i = \sum_{i=1}^p V_i' \sum_{l=1}^d u_{l_i} u_l = \sum_{i=1}^p \sum_{l=1}^d u_{l_i} V_i' u_l \\ &= \sum_{l=1}^d \left(\sum_{i=1}^p u_{l_i} V_i' \right) u_l \\ \text{Tr}(VP) &= \sum_{l=1}^d u_l' V u_l \end{aligned}$$

D'où le résultat.

C.3 $I(H_1 \oplus H_2) = I(H_1) + I(H_2)$

Etant donnés deux sous-espaces vectoriels orthogonaux H_1 et H_2 , montrons que

$$I(H_1 \oplus H_2) = I(H_1) + I(H_2)$$

Il suffit de constater que $\|x^{\widehat{H_1 \oplus H_2}}\|^2 = \|x^{H_1}\|^2 + \|x^{H_2}\|^2$.

C.4 Solution de $\max_{\sum_i a_i^2=1} \sum_{i=1}^p \lambda_i a_i^2$

— Borne supérieure de $\max_{\sum_i a_i^2=1} \sum_{i=1}^p \lambda_i a_i^2$:

$$\sum_{i=1}^p \underbrace{\lambda_i}_{\leq \lambda_1} a_i^2 \leq \lambda_1 \sum_{i=1}^p a_i^2 \quad (\text{C.1})$$

$$\sum_{i=1}^p \lambda_i a_i^2 \leq \lambda_1 \quad (\text{C.2})$$

— Cette borne est atteinte pour $a_1 = 1$ et $a_i = 0$, $\forall i \neq 1$, i.e. pour $u = u_1$.

C.5 Lemme fondamental

Soit $k < p$; si F_k est le sous-espace vectoriel de dimension k d'inertie maximale, alors le sous-espace vectoriel de dimension $k+1$ d'inertie expliquée maximale est $F_{k+1} = F_k \oplus \Delta u$, où Δu est la droite vectorielle orthogonale à F_k d'inertie expliquée maximale.

En effet, soit E_{k+1} un sous-espace vectoriel de dimension $k+1$. Comme $\dim(F_k^\perp) = n-k$, $\dim(F_k^\perp \cap E_{k+1}) \geq 1$. Soit b dans $F_k^\perp \cap E_{k+1}$, avec $\|b\| = 1$, et G tel que $E_{k+1} = \Delta_b \oplus G$. Posons $F = \Delta_b \oplus F_k$. Nous avons

$$I(E_{k+1}) = I(\Delta_b) + I(G) \quad (\text{C.3})$$

$$I(F) = I(\Delta_b) + I(F_k) \quad (\text{C.4})$$

Or $I(F_k) \geq I(G)$, donc $I(F) \geq I(E_{k+1})$.

Maintenant, choisissons b dans F_k^\perp tel que $I(\Delta_b)$ soit maximal; on voit que $b = \pm u_{k+1}$. Posons $F_{k+1} = \Delta_b \oplus F_k$.

$$I(F_{k+1}) = I(\Delta_{u_{k+1}}) + I(F_k) \quad (\text{C.5})$$

Il est clair que

$$\forall b \in F_k^\perp, I(\Delta_b) \leq I(\Delta_{u_{k+1}}) \quad (\text{C.6})$$

donc $I(F_{k+1}) \geq I(E_{k+1})$.

C.6 Données tirées de [JAM76]

	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS
"HAUS"	610.	140.	60.	10.	120.	95.	115.	760.	175.	315.
"FAUS"	475.	90.	250.	30.	140.	120.	100.	775.	115.	325.
"FNAU"	10.	0.	495.	110.	170.	110.	130.	785.	160.	430.
"HMUS"	615.	141.	65.	10.	115.	90.	115.	765.	180.	305.
"FMUS"	179.	29.	421.	87.	161.	112.	119.	776.	143.	373.
"HCUS"	585.	115.	50.	0.	150.	105.	100.	760.	150.	385.
"FCUS"	482.	94.	196.	18.	141.	130.	96.	775.	132.	336.
"Hawe"	652.	100.	95.	7.	57.	85.	150.	807.	115.	330.
"FAWE"	510.	70.	307.	30.	80.	95.	142.	815.	87.	262.
"FNAW"	20.	7.	567.	87.	112.	90.	180.	842.	125.	367.
"HMWE"	655.	97.	97.	10.	52.	85.	152.	807.	122.	320.
"FMWE"	168.	22.	529.	69.	102.	83.	174.	825.	119.	392.
"HCWE"	642.	105.	72.	0.	62.	77.	140.	812.	100.	387.
"FCWE"	389.	34.	262.	14.	92.	97.	147.	848.	84.	392.
"HAYO"	650.	140.	120.	15.	85.	90.	105.	760.	70.	365.
"FAYO"	560.	105.	375.	45.	90.	90.	95.	745.	60.	235.
"FNAY"	10.	10.	710.	55.	145.	85.	130.	815.	60.	380.
"HMYO"	650.	145.	112.	15.	85.	90.	105.	760.	80.	357.
"FMYO"	260.	52.	576.	59.	116.	85.	117.	775.	65.	295.
"HCYO"	615.	125.	95.	0.	115.	90.	85.	760.	40.	475.
"FCYO"	433.	89.	318.	23.	112.	96.	102.	774.	45.	409.
"HAES"	650.	142.	122.	22.	76.	94.	100.	764.	96.	334.
"FAES"	578.	106.	338.	42.	106.	94.	92.	752.	64.	228.
"FNAE"	24.	8.	594.	72.	158.	92.	128.	840.	86.	398.
"HMES"	652.	133.	134.	22.	68.	94.	102.	762.	122.	310.
"FMES"	434.	77.	431.	60.	117.	88.	105.	770.	73.	229.
"HCES"	627.	148.	68.	0.	88.	92.	86.	770.	58.	463.
"FCES"	433.	86.	296.	21.	128.	102.	94.	798.	58.	379.

Annexe D

Algorithme NIPALS

Présentation de l'algorithme NIPALS - *Nonlinear estimation by Iterative Partial Least Squares* - dans le cas de données complètes puis dans le cas de données manquantes. La présentation tirée de l'ouvrage «La régression PLS : théorie et pratique» de Tenenhaus a été adaptée aux notations en usage ici.

L'algorithme NIPALS permet d'estimer les composantes principales à partir de la matrice de données, y compris lorsque des données manquent. Il est basé sur le principe de la méthode de la puissance itérée et repose sur la possibilité d'écrire X , la matrice des données que l'on suppose centrées, comme suit :

$$X = \sum_l c_l u'_l$$

où c_l et u'_l sont respectivement la l^e composante principale et la transposée du l^e vecteur propre de la matrice de covariance (ou corrélation si les données sont réduites). Cette écriture est justifiée car tout individu x_i peut s'exprimer dans la base des u_l par $x_i = \sum_l \langle x_i, u_l \rangle u_l$, d'où :

$$X = \begin{pmatrix} \vdots \\ \sum_{l=1}^p \langle x_i, u_l \rangle u'_l \\ \vdots \end{pmatrix}$$

Ci-dessous, X est supposée être de rang a .

D.1 Sans données manquantes

Etape 1 : $X_0 \leftarrow X$

Etape 2 : pour $h = 1 \dots a = rg(X)$,

 2.1 : $C_h \leftarrow$ première colonne de X_{h-1}

 2.2 : itérer jusqu'à convergence de C_h

 2.2.1 : $u_h \leftarrow \frac{^t X_{h-1} C_h}{(^t C_h C_h)}$ (D.1)

 2.2.2 : normer u_h

 2.2.3 : $C_h \leftarrow \frac{X_{h-1} u_h}{^t u_h u_h}$

 2.3 : $X_h \leftarrow X_{h-1} - C_h {}^t u_h$

Afin de simplifier les notations utilisées ici - ce qui facilitera la lisibilité de sa transposition au cas de données manquantes, nous pouvons ré-ecrire cet algorithme comme suit :

Etape 1 : $\tilde{X} \leftarrow X$

Etape 2 : pour $h = 1 \dots a = rg(X)$,

 2.1 : $C_h \leftarrow$ première colonne de \tilde{X}

 2.2 : itérer jusqu'à convergence de C_h

$$2.2.1 : u_h \leftarrow \frac{^t\tilde{X}C_h}{(^tC_hC_h)} \quad (D.2)$$

 2.2.2 : normer u_h

$$2.2.3 : C_h \leftarrow \frac{\tilde{X}u_h}{^tu_hu_h}$$

 2.3 : $\tilde{X} \leftarrow \tilde{X} - C_h^t u_h$

D.2 Avec données manquantes

Etape 1 : $\tilde{X} \leftarrow X$

Etape 2 : pour $h = 1 \dots a = rg(X)$,

 2.1 : $C_h \leftarrow$ première colonne de \tilde{X}

 2.2 : itérer jusqu'à convergence de C_h

$$2.2.1 : u_h \leftarrow \frac{^t\tilde{X}C_h}{(^tC_hC_h)} \text{ devient : pour } j = 1 \dots p,$$

$$u_{jh} \leftarrow \frac{\sum_{i \in A} x_{ij} c_{ih}}{\sum_{i \in A} c_{ih}^2}, \text{ où } A = \{i : x_{ij} \text{ et } c_{ih} \text{ existent}\} \quad (D.3)$$

 2.2.2 : normer u_h

$$2.2.3 : C_h \leftarrow \frac{\tilde{X}u_h}{^tu_hu_h} \text{ devient : pour } i = 1 \dots n,$$

$$c_{ih} \leftarrow \frac{\sum_{j \in B} x_{ij} u_{jh}}{\sum_{j \in B} u_{jh}^2}, \text{ où } B = \{j : x_{ij} \text{ existe}\}$$

 2.3 : $X_h \leftarrow X_{h-1} - C_h^t u_h$

Annexe E

Rappel sur le χ^2

La distance du χ^2 nous est utile pour l'analyse factorielle des correspondances. Il n'est peut-être pas inutile de faire quelques rappels sur le χ^2 . Ci-dessous, nous rappelons comment le χ^2 est utilisé pour mesurer l'écart à une certaine «indépendance», d'abord dans l'étude de la répartition des modalités d'une variable aléatoire observé dans un échantillon puis dans l'étude des relations des modalités de deux variables. Pour une présentation plus détaillée, le lecteur pourra consulter [SAP90].

E.1 Le test du χ^2

Soit une variable aléatoire X à k modalités. La i^{e} modalité a la probabilité p_i d'apparition. Supposons à présent qu'un échantillon de la population conduise à l'obtention de k classes d'effectifs $N_1 \dots N_k$, la i^{e} classe étant associée à la i^{e} modalité de X .

La répartition des effectifs N_i dérive-t-elle de la distribution de probabilité d'apparition des modalités ? Posons $\sum_i N_i = n$. L'effectif espéré de la i^{e} classe est np_i .

La quantité $\sum_i \frac{(N_i - np_i)^2}{np_i}$ mesure l'écart entre les effectifs espérés et ceux observés. Intuitivement, les derniers sont d'autant plus proches des premiers que cette quantité est faible.

Pratiquement, l'hypothèse que les répartitions sont les mêmes sera rejetée, avec un risque d'erreur de α , dès que la valeur de la quantité ci-dessus excédera la valeur du χ^2 à $k - 1$ ddl pour le risque α .

Remarquons que la quantité ci-dessus s'exprime également facilement en fonction des fréquences observées, notées $f_i = N_i/n$, et des probabilités connues :

$$\sum_i \frac{(N_i - np_i)^2}{np_i} = n \sum_i i \frac{(N_i/n - p_i)^2}{p_i} = n \sum_i \frac{(f_i - p_i)^2}{p_i}$$

E.2 Le χ^2 comme mesure de l'écart à l'indépendance

Reprendons les notations du chapitre 3. Nous considérons deux variables qualitatives notées V et W .

Partant des mesures effectuées sur une population de N individus, nous disposons du tableau de contingence et du tableau associé des fréquences suivants :

V	W	$w_1 \dots w_j \dots w_p$		
v_1				
\vdots		\vdots		
v_i		$\dots n_{ij} \dots$	$n_{i.}$	\rightarrow
\vdots		\vdots		
v_n				
		$\dots n_{.j} \dots$	N	

V	W	$w_1 \dots w_j \dots w_p$		
v_1				
\vdots		\vdots		
v_i		$\dots \frac{n_{ij}}{N} \dots$		$\frac{n_{i.}}{N}$
\vdots		\vdots		
v_n				
		$\dots n_{.j}/N \dots$		1

Considérons le tableau des profils lignes Z :

$$Z = \begin{pmatrix} & \vdots & \\ \dots & \frac{f_{ij}}{f_{i.}} & \dots \\ & \vdots & \end{pmatrix}$$

Lorsque tous les profils lignes sont identiques, il est clair que la connaissance de la modalité de W pour un individu n'apporte aucune information sur sa modalité pour V .

Dans ce cas, tous les éléments d'une colonne sont égaux entre eux et en particulier égaux à la fréquence observée pour la j^e modalité de W :

$$\forall i, j, \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$$

Remarquons que cette relation donne

$$\forall i, j, \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N}$$

Par suite, l'égalité des profils lignes entraîne l'égalité des profils colonnes et réciproquement. L'indépendance observée se traduit par l'une des relations ci-dessus ou par

$$\forall i, j, n_{ij} = \frac{n_{i.} n_{.j}}{N}$$

Une mesure de l'écart à l'indépendance de V et W est donnée par

$$d^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{N} \right)^2}{\frac{n_{i.} n_{.j}}{N}}$$

ou encore, en utilisant les fréquences

$$d^2 = \sum_i \sum_j N \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$$

Il est trivial qu'en cas d'indépendance observée, cette quantité s'annule.

E.2.1 Bornes de l'écart à l'indépendance

Déterminons à présent une borne supérieure pour d^2 .

$$d^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{N} \right)^2}{\frac{n_{i.} n_{.j}}{N}}$$

$$\begin{aligned}
 &= N \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{N} \right)^2}{n_{i\cdot}n_{\cdot j}} \\
 &= N \sum_i \sum_j \frac{n_{ij}^2 - 2\frac{n_{ij}n_{i\cdot}n_{\cdot j}}{N} + \frac{n_{i\cdot}^2n_{\cdot j}^2}{N^2}}{n_{i\cdot}n_{\cdot j}} \\
 &= N \sum_i \sum_j \left[\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 2\frac{n_{ij}}{N} + \frac{n_{i\cdot}n_{\cdot j}}{N^2} \right] \\
 &= N \left[\sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - \underbrace{2 \sum_i \sum_j \frac{n_{ij}}{N}}_1 + \underbrace{\sum_i \sum_j \frac{n_{i\cdot}n_{\cdot j}}{N^2}}_1 \right] \\
 d^2 &= N \left[\sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1 \right]
 \end{aligned}$$

Comme $\frac{n_{ij}}{n_{\cdot j}} \leq 1$ et $\frac{n_{ij}}{n_{i\cdot}} \leq 1$, nous avons $\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} \leq \frac{n_{ij}}{n_{i\cdot}}$.

Par suite, nous obtenons la majoration

$$\sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} \leq \sum_i \sum_j \underbrace{\frac{n_{ij}}{n_{i\cdot}}}_{=1} = \sum_i 1 = n , \text{ d'où } d^2 \leq N(n-1)$$

De même,

$$\sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} \leq \sum_j \sum_i \underbrace{\frac{n_{ij}}{n_{\cdot j}}}_{=1} = \sum_j 1 = p , \text{ d'où } d^2 \leq N(p-1)$$

Ainsi $d^2 \leq N \inf(n-1, p-1)$.

Nous obtenons ainsi une borne supérieure pour d^2 . Nous allons à présent montrer que cette borne peut être atteinte dans certains cas, mais pas dans tous les cas. Le problème sera abordé de la façon suivante : un couple de distributions en effectifs des variables V et W étant donné pour une population, quelle peut être l'inertie maximale d'un tableau de contingence dont les marges coïncident avec ces distributions.

Considérons dans un premier temps le couple de distributions en effectifs indiqué par le tableau suivant :

	W	w_1	w_2	w_3	
V					
v_1	?	?	?		3
v_2	?	?	?		4
	1	1	5		7

La borne supérieure proposée plus égale $7 \times \min(2-1, 3-1) = 7$. Seuls quatre tableaux de contingence disposent des mêmes marges :

$\begin{array}{ ccc c } \hline 1 & 1 & 1 & 3 \\ \hline 0 & 0 & 4 & 4 \\ \hline 1 & 1 & 5 & 7 \\ \hline \end{array}$	$\rightarrow d^2 \approx 3,73$	$\begin{array}{ ccc c } \hline 1 & 0 & 2 & 3 \\ \hline 0 & 1 & 3 & 4 \\ \hline 1 & 1 & 5 & 7 \\ \hline \end{array}$	$\rightarrow d^2 \approx 2,10$	$\begin{array}{ ccc c } \hline 0 & 1 & 2 & 3 \\ \hline 1 & 0 & 3 & 4 \\ \hline 1 & 1 & 5 & 7 \\ \hline \end{array}$	$\rightarrow d^2 \approx 2,10$
		$\begin{array}{ cc c } \hline 0 & 0 & 3 \\ \hline 1 & 1 & 2 \\ \hline 1 & 1 & 5 \\ \hline \end{array}$	$\rightarrow d^2 \approx 2,10$		

Ainsi, pour ce couple de distribution, l'inertie la plus forte est d'environ 3,73. Le lecteur pourra s'entraîner sur les tableaux correspondant au couple suivant :

V	W	w_1	w_2	w_3	
v_1	?	?	?	?	3
v_2	?	?	?	?	4
v_3	?	?	?	?	5
		1	2	9	12

La borne supérieure donne la valeur $12 \times \min(3 - 1, 3 - 1) = 24$. Le lecteur pourra vérifier qu'au mieux, un tableau pourra donner 12 comme inertie maximale.

Ainsi, il est prouvé que cette borne excède parfois l'inertie que peut prendre un tableau, ces marges étant connues. Il est toutefois des cas où cette borne est atteinte. C'est le cas lorsque une dépendance fonctionnelle existe entre les variables V et W .

Plus précisément, supposons que $p \leq n$ et que W est fonctionnellement lié à V (i.e. connaissant la valeur pour V , la valeur pour W est déduite). Cela signifie que pour chaque modalité de V (donc chaque ligne), il existe une seule modalité de W observée (i.e. une case non nulle, et cette case a pour valeur $n_{i.}$). Cela s'écrit : $\forall i, \exists j_i : n_{ij_i} = n_{i.}, \forall j \neq j_i, n_{ij} = 0$

Par suite :

$$\begin{aligned}
 \sum_i \sum_j \frac{n_{ij}^2}{n_{i.} n_{.j}} &= \sum_j \sum_i \frac{n_{ij}^2}{n_{i.} n_{.j}} \text{ car nous allons sommer par colonnes} \\
 &= \sum_j \sum_{i \in K_j} \frac{n_{ij}^2}{n_{i.} n_{.j}} \text{ où } K_j \text{ est l'ensemble des indices pour lesquels } n_{ij} \text{ n'est pas nul} \\
 &= \sum_j \sum_{i \in K_j} \frac{n_{i.}^2}{n_{i.} n_{.j}} \text{ car les } n_{ij} \text{ non nuls égalent } n_{i.}, \text{ ce qui entraîne la simplification} \\
 &= \sum_j \sum_{i \in K_j} \frac{n_{i.}}{n_{.j}} \\
 &= \sum_j \frac{1}{n_{.j}} \underbrace{\sum_{i \in K_j} n_{i.}}_{=n_{.j}} \text{ car nous sommes les éléments non nuls de la } j^{\text{e}} \text{ colonne} \\
 &= \sum_j 1 \\
 \sum_i \sum_j \frac{n_{ij}^2}{n_{i.} n_{.j}} &= p
 \end{aligned}$$

Ainsi, cette borne est la meilleure possible dans le cas de dépendance fonctionnelle. La question naturelle maintenant est de savoir si la réalisation de cette borne implique la dépendance fonctionnelle... C'est effectivement le cas, et la preuve est laissée au courageux lecteur parvenu à ce stade.

E.2.2 Significativité de l'écart

La valeur d^2 est à comparer avec la valeur donnée par la table proposée en annexe F (ou par une table plus précise), le nombre de degré de liberté étant $(n - 1)(p - 1)$ et un seuil d'erreur pour le rejet de l'hypothèse étant fixé. Par exemple, supposons que nous calculons pour d^2 la valeur 134,23 alors que le nombre de ddl est 20 et que le seuil d'erreur est fixé à 0,005 (i.e. 0,5%). La valeur critique lue dans le tableau est 40, valeur largement inférieure à celle calculée. Nous pouvons rejeter l'hypothèse d'indépendance des modalités de V et W. Il y a donc des relations à identifier et à expliquer entre les modalités...

La quantité calculée mesure un certain écart entre le tableau de contingence d'élément de base n_{ij} et le tableau des effectifs théoriques (sous l'hypothèse d'indépendance) dont l'élément de base est $\frac{n_{i.} \cdot n_{.j}}{N}$. Le calcul du tableau d'élément $\frac{\left(n_{ij} - \frac{n_{i.} \cdot n_{.j}}{N}\right)^2}{\frac{n_{i.} \cdot n_{.j}}{N}} \frac{1}{d^2}$ permet de mettre en évidence les associations significatives entre les modalités des deux variables V et W qui ont fortement contribué à la formation de d^2 . Par ailleurs, le signe de $\left(n_{ij} - \frac{n_{i.} \cdot n_{.j}}{N}\right)$ permet de savoir s'il y a sur-représentation des individus associant les modalités correspondantes (signe positif) ou sous-représentation.

E.3 Distance du χ^2

Montrons que la fusion de deux modalités conduit à la réduction de I_g , l'inertie du nuage initial. Plus précisément, supposons, sans restreindre la généralité, que les modalités fusionnées soient w_{p-1} et w_p . Notons I_g^* l'inertie du nuage transformé. Nous voulons montrer que $I_g \geq I_g^*$.

Comparer I_g et I_g^* revient à comparer

$$\sum_i \left[\frac{n_{ip-1}^2}{n_{i.} \cdot n_{.p-1}} + \frac{n_{ip}^2}{n_{i.} \cdot n_{.p}} \right] \text{ et } \sum_i \frac{(n_{ip-1} + n_{ip})^2}{n_{i.} \cdot (n_{.p-1} + n_{.p})}$$

Vérifions la supériorité du premier par rapport au second :

$$\begin{aligned} \sum_i \frac{(n_{ip-1} + n_{ip})^2}{n_{i.} \cdot (n_{.p-1} + n_{.p})} &\stackrel{?}{\leq} \sum_i \left[\frac{n_{ip-1}^2}{n_{i.} \cdot n_{.p-1}} + \frac{n_{ip}^2}{n_{i.} \cdot n_{.p}} \right] \\ \sum_i \frac{(n_{ip-1} + n_{ip})^2}{n_{i.}} &\stackrel{?}{\leq} \sum_i \left[\frac{n_{ip-1}^2}{n_{i.} \cdot n_{.p-1}} + \frac{n_{ip}^2}{n_{i.} \cdot n_{.p}} \right] (n_{.p-1} + n_{.p}) \\ \sum_i \frac{n_{ip-1}^2}{n_{i.}} + \sum_i \frac{2n_{ip-1}n_{ip}}{n_{i.}} + \sum_i \frac{n_{ip}^2}{n_{i.}} &\stackrel{?}{\leq} \sum_i \frac{n_{ip-1}^2}{n_{i.} \cdot n_{.p-1}} (n_{.p-1} + n_{.p}) + \sum_i \frac{n_{ip}^2}{n_{i.} \cdot n_{.p}} (n_{.p-1} + n_{.p}) \\ \sum_i \frac{n_{ip-1}^2}{n_{i.}} + \sum_i \frac{2n_{ip-1}n_{ip}}{n_{i.}} + \sum_i \frac{n_{ip}^2}{n_{i.}} &\stackrel{?}{\leq} \sum_i \frac{n_{ip-1}^2}{n_{i.}} + \sum_i \frac{n_{ip-1}^2}{n_{i.}} \frac{n_{.p}}{n_{.p-1}} + \sum_i \frac{n_{ip}^2}{n_{i.}} + \sum_i \frac{n_{ip}^2}{n_{i.}} \frac{n_{.p-1}}{n_{.p}} \\ \sum_i \frac{2n_{ip-1}n_{ip}}{n_{i.}} &\stackrel{?}{\leq} \sum_i \frac{n_{ip-1}^2}{n_{i.}} \frac{n_{.p}}{n_{.p-1}} + \sum_i \frac{n_{ip}^2}{n_{i.}} \frac{n_{.p-1}}{n_{.p}} \\ 0 &\stackrel{?}{\leq} \sum_i \frac{n_{ip-1}^2}{n_{i.}} \frac{n_{.p}}{n_{.p-1}} - \sum_i \frac{2n_{ip-1}n_{ip}}{n_{i.}} + \sum_i \frac{n_{ip}^2}{n_{i.}} \frac{n_{.p-1}}{n_{.p}} \\ 0 &\stackrel{?}{\leq} \sum_i \frac{i \left(\sqrt{\frac{n_{.p}}{n_{.p-1}}} n_{ip-1} - \sqrt{\frac{n_{.p-1}}{n_{.p}}} n_{ip} \right)^2}{n_{i.}} \end{aligned}$$

Ainsi, nous avons bien $I_g^* \leq I_g$. Par ailleurs, cette inégalité se transforme en égalité lorsque les profils associés à w_{p-1} et w_p sont identiques. Il suffit de poser $n_{ip-1} = kn_{ip}$ et d'observer l'annulation du terme ci-dessus.

Annexe F

Table de valeurs du χ^2

La table ci-dessous donne, en fonction de la valeur du nombre de degrés de liberté (abrégé par ddl dans la suite) et en fonction de P , la valeur du χ^2 telle que la probabilité pour une variable aléatoire suivant une loi du χ^2 de dépasser cette valeur est P . La figure ci-dessous illustre cela :

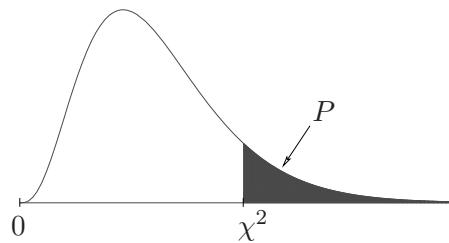


FIGURE F.1 – L'aire hachurée égale la probabilité P de dépasser la valeur du χ^2 , le nombre de degrés de liberté étant fixé.

Les valeurs données par la table peuvent être cherchées à partir des formules exactes suivantes :

- pour un # de ddl ν pair, $P(\chi_\nu^2 > x) = \sum_{i=0}^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right) \frac{\left(\frac{x}{2}\right)^i}{i!}$
- pour un # de ddl ν impair, $P(\chi_\nu^2 > x) = \dots$

ddl	$1 - P$													
	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995	0.999
1	0.000	0.000	0.001	0.004	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.07	12.83	15.09	16.75	20.51
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.603	3.053	3.816	4.575	5.578	7.584	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	3.074	3.571	4.404	5.226	6.304	8.438	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.565	4.107	5.009	5.892	7.041	9.299	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.075	4.660	5.629	6.571	7.790	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.601	5.229	6.262	7.261	8.547	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.142	5.812	6.908	7.962	9.312	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.697	6.408	7.564	8.672	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.265	7.015	8.231	9.390	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.844	7.633	8.907	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.434	8.260	9.591	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.31
21	8.034	8.897	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.643	9.542	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.260	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.886	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.65	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
31	14.46	15.66	17.54	19.28	21.43	25.39	30.34	35.89	41.42	44.99	48.23	52.19	55.00	61.10
32	15.13	16.36	18.29	20.07	22.27	26.30	31.34	36.97	42.58	46.19	49.48	53.49	56.33	62.49
33	15.82	17.07	19.05	20.87	23.11	27.22	32.34	38.06	43.75	47.40	50.73	54.78	57.65	63.87
34	16.50	17.79	19.81	21.66	23.95	28.14	33.34	39.14	44.90	48.60	51.97	56.06	58.96	65.25
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27	66.62
36	17.89	19.23	21.34	23.27	25.64	29.97	35.34	41.30	47.21	51.00	54.44	58.62	61.58	67.98
37	18.59	19.96	22.11	24.07	26.49	30.89	36.34	42.38	48.36	52.19	55.67	59.89	62.88	69.35
38	19.29	20.69	22.88	24.88	27.34	31.81	37.34	43.46	49.51	53.38	56.90	61.16	64.18	70.70
39	20.00	21.43	23.65	25.70	28.20	32.74	38.34	44.54	50.66	54.57	58.12	62.43	65.48	72.06
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.1	118.5	124.3	129.6	135.8	140.2	149.4
120	83.85	86.92	91.57	95.70	100.6	109.2	119.3	130.1	140.2	146.6	152.2	159.0	163.6	173.6
140	100.7	104.0	109.1	113.7	119.0	128.4	139.3	150.9	161.8	168.6	174.6	181.8	186.8	197.4
160	117.7	121.3	126.9	131.8	137.5	147.6	159.3	171.7	183.3	190.5	196.9	204.5	209.8	221.0
180	134.9	138.8	144.7	150.0	156.2	166.9	179.3	192.4	204.7	212.3	219.0	227.1	232.6	244.4
200	152.2	156.4	162.7	168.3	174.8	186.2	199.3	213.1	226.0	234.0	241.1	249.4	255.3	267.5
240	187.3	192.0	199.0	205.1	212.4	224.9	239.3	254.4	268.5	277.1	284.8	293.9	300.2	313.4
300	240.7	246.0	253.9	260.9	269.1	283.1	299.3	316.1	331.8	341.4	349.9	359.9	366.8	381.4
400	330.9	337.2	346.5	354.6	364.2	380.6	399.3	418.7	436.6	447.6	457.3	468.7	476.6	493.1

Annexe G

Prolongements de l'annexe E

G.1 Preuve laissée au lecteur

Nous avons donné la preuve que la dépendance fonctionnelle de W à V implique que l'inertie du tableau de contingence issu de l'observation de ces variables sur une population est maximale. Réciproquement, si l'inertie d'un tableau de contingence égale $N \times \min(n - 1, p - 1)$, alors les observations permettent de conclure qu'une dépendance de W par rapport à V est observée. Prouvons cela en supposant que $p \leq n$:

— Tout d'abord, nous avons la majoration suivante

$$\begin{aligned} \sum_i \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} &= \frac{1}{n_{\cdot j}} \sum_i \frac{n_{ij}^2}{n_{i\cdot}} \\ &\leq \frac{1}{n_{\cdot j}} \sum_i n_{ij} \quad \text{car } n_{ij} \leq n_{i\cdot} \\ &\leq \frac{1}{n_{\cdot j}} n_{\cdot j} \\ \sum_i \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} &\leq 1 \end{aligned}$$

- Supposons ensuite que $d^2 = N(p - 1)$, ce qui revient à supposer que $\sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} = p$ est vérifié. Supposons de plus qu'il existe j^* tel que $\sum_i \frac{n_{ij^*}^2}{n_{i\cdot} n_{\cdot j^*}} < 1$. Ces deux points impliquent $\sum_{j \neq j^*} \sum_i \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} > p - 1$. Or pour tout j , $\sum_i \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} \leq 1$, donc $\sum_{j \neq j^*} \sum_i \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} \leq p - 1$. Donc, $\sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} = p$ implique que pour tout j , $\sum_i \frac{n_{ij^*}^2}{n_{i\cdot} n_{\cdot j^*}} = 1$.
- Supposons que pour tout j , $\sum_i \frac{n_{ij^*}^2}{n_{i\cdot} n_{\cdot j^*}} = 1$.

$$\begin{aligned} \sum_i \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} = 1 &\Leftrightarrow \frac{1}{n_{\cdot j}} \sum_i \frac{n_{ij}^2}{n_{i\cdot}} = 1 \\ &\Leftrightarrow \sum_i \frac{n_{ij}^2}{n_{i\cdot}} = n_{\cdot j} \end{aligned}$$

$$\begin{aligned} &\Leftrightarrow \sum_i n_{ij} \frac{n_{ij}}{n_{i\cdot}} = n_{\cdot j} \\ &\Leftrightarrow \sum_{i \in K_j} n_{ij} \frac{n_{ij}}{n_{i\cdot}} = n_{\cdot j} \text{ où } K_j = \{i | n_{ij} \neq 0\} \end{aligned}$$

Supposons qu'il existe $i^* \in K_j$ tel que $n_{i^*j} < n_{i^*\cdot}$. Ceci implique que

$$\sum_{i \in K} n_{ij} \frac{n_{ij}}{n_{i\cdot}} < \sum_{i \in K} n_{ij} = n_{\cdot j}$$

ce qui contredit l'hypothèse de départ. Donc, $\forall j, \sum_i \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} = 1 \Rightarrow \forall j, \forall i \in K_j, n_{ij} = n_{i\cdot}$. Ceci implique que chaque ligne ne comporte qu'un élément non nul, égal à la marge associée. Ceci caractérise bien la dépendance fonctionnelle de W par rapport à V .

Ainsi, la réalisation de la borne supérieure $N \min(n - 1, p - 1)$ implique que W est liée fonctionnellement à V .

G.2 Développements annexes

Le problème de la détermination de la valeur maximum de l'inertie des tableaux constructibles, les marges étant fixées nous paraît intéressant.

Nous proposons ci-dessous d'examiner l'existence de liens entre des conditions sur les marges et l'inertie maximale qu'un tableau de contingence respectant ces marges peut atteindre. Nous considérerons deux cas : le cas où les variables V et W ont les mêmes nombres de modalités et le cas où ces nombres diffèrent. Dans chaque cas, nous essaierons d'identifier des informations permettant de caractériser des marges permettant une dépendance fonctionnelle et les marges empêchant toute dépendance fonctionnelle.

G.2.1 Cas où $n = p$

Cas avec possibilité de dépendance fonctionnelle

Supposons que les marges considérées permettent une dépendance fonctionnelle de W par rapport à V . Cela signifie que la connaissance de la modalité de V permet la déduction de la modalité de W . Par suite, à toute modalité de V correspond une seule modalité de W . Notons $f : \{v_i | i = 1..n\} \rightarrow \{w_j | j = 1..n\}$ la fonction qui associe à v_i la modalité w_j . Comme toute modalité de W à au moins un correspondant (car toute modalité de W est observée au moins une fois), la fonction f est surjective. De plus, toute modalité de V à également un correspondant. Comme il y a $n = p$ modalités pour chaque variable, la fonction est nécessairement bijective.

De plus, l'effectif associé à une modalité de V est le même pour la modalité associée de W . Ainsi, les marges sont les mêmes, à l'ordre près.

Nous pouvons ainsi retenir que dans le cas où $n = p$, il ne peut y avoir de dépendance fonctionnelle que si les marges sont égales, à l'ordre près. Nous savons déjà que dans le cas de dépendance fonctionnelle, l'inertie maximale est $N \times n$.

Cas sans possibilité de dépendance fonctionnelle

Comment construire un tableau de plus grande inertie dans le cas où les marges diffèrent à l'ordre près.

Observons le cas ci-dessous :

- Dans un premier temps, l'observation des marges nous permet de déterminer des bornes minimales. Par exemple, le troisième élément de la première ligne peut «descendre» jusqu'à 0, car les troisièmes éléments des lignes suivantes permettent de réaliser le 9 en marge de la troisième colonne. En revanche, le troisième élément de la seconde ligne ne peut être inférieur à 1 car les troisièmes éléments des autres lignes ne permettent d'atteindre que 8 au maximum. De façon similaire, toutes les bornes minimales peuvent être déterminer.

?	?	?	3	0	0	0	3
?	?	?	4	0	0	1	4
?	?	?	5	0	0	2	5
1	2	9	12	1	2	9	12

- A partir de cette matrice de bornes minimales, nous pouvons redéfinir les marges et reconstruire le problème avec les nouvelles marges mais en restreignant les tableaux cherchés à ceux possédant des éléments non nuls là où les marges sont non nulles.

?	?	?	3	1	2	0	3
?	?	? > 0	3	0	0	3	3
?	?	? > 0	3	0	0	3	3
1	2	6	9	1	2	6	9

- Les tableaux cherchés correspondent aux tableaux trouvés précédemment auxquels est ajoutée la matrice de bornes minimales.

1	2	0	3	1	2	0	3
0	0	3	3	0	0	4	4
0	0	3	3	0	0	5	5
1	2	6	9	1	2	9	12

Hélas, évidemment, tout ceci est à vérifier. Mais l'intuition est la suivante : l'inertie maximale correspond à la dépendance fonctionnelle. Si celle-ci n'est pas réalisable, l'inertie maximale correspond à une situation se rapprochant le plus possible de la situation de dépendance fonctionnelle. La proximité ici pourrait être définie par une correspondance entre partitions des marges, les partitions comportant le plus de partie possible tout en respectant la condition suivante : la somme des marges constituant une partie doit être égale la somme des marges de la partie en correspondance.

Plus précisément, ici, la partition sur les marges des colonnes est $\{\{1, 2\}, \{9\}\}$ et celle des marges des colonnes est $\{\{3\}, \{4, 5\}\}$. $\{1, 2\}$ correspond à $\{3\}$, et $1 + 2 = 3$. De même, $\{9\}$ correspond à $\{4, 5\}$ et $4 + 5 = 9$.

Investigation à poursuivre...

G.2.2 Cas où $n \geq p$

Cas avec possibilité de dépendance fonctionnelle

Cas sans possibilité de dépendance fonctionnelle

Développements ?...

Table des matières détaillée

Notations	1
I Analyses factorielles dans le cadre multivarié	2
1 Préparation des données	3
1.1 Contrôle et synthèse des données	3
1.2 Recherche de relations	4
1.3 Etablir une causalité	6
2 L'analyse en composantes principales	8
2.1 L'ACP, outil d'exploration et de synthèse	8
2.2 L'ACP présentée comme problème de visualisation	12
2.3 Solution au problème dans le cadre euclidien simplifié	15
2.4 Solution au problème dans le cadre euclidien général	17
2.5 Composantes principales : de l'espace des individus vers l'espace des variables	18
2.6 Récapitulatif des formulations du problème de l'ACP	21
2.7 Mise en œuvre et interprétation des résultats	23
2.8 Davantage d'exemples	35
3 L'analyse factorielle des correspondances	36
3.1 L'AFC pour rechercher des liens dans une paire de variables qualitatives	36
3.2 L'analyse de tableaux de contingence	37
3.3 ACP des profils lignes	40
3.4 ACP des profils colonnes	41
3.5 Exploitation des relations de dualité	42
3.6 L'analyse des profils lignes non-centrés	44
3.7 Interprétation des résultats	46
4 L'analyse factorielle des correspondances multiples	51
4.1 Cadre	51
4.2 Projections des nuages d'individus et de modalités	56
4.3 Mise en œuvre et interprétation des résultats	57
4.4 Equivalence de l'analyse d'un tableau de Burt	59
4.5 Trois méthodes pour l'analyse d'un couple de variables	60
5 L'analyse factorielle des données mixtes	61
5.1 Cadre	61
5.2 Projections des nuages d'individus et de variables et modalités	64
5.3 Mise en œuvre et interprétation des résultats	64

II Analyses factorielles dans le cadre fonctionnel	69
6 Introduction au cadre fonctionnel	70
6.1 Données fonctionnelles	70
6.2 Statistique et données fonctionnelles	73
6.3 Cadre d'analyse des données fonctionnelles	74
7 Transport de l'ACP dans le cadre fonctionnel	75
7.1 Problème dans le cas monofonctionnel	75
7.2 Problème dans le cas bifonctionnel	76
7.3 Problème dans le cas multifonctionnel	77
8 Solution dans le cas monofonctionnel	78
8.1 Réduction du problème par l'emploi d'une base de l'espace des individus	78
8.2 Solution théorique	79
8.3 Approximation de la solution par discréétisation du problème	80
8.4 Application	81
Appendices	86
A Preuves de la partie ACP multivariée dans le cadre euclidien simplifié	86
A.1 Relation de Huygens	86
A.2 $Q_H = 2I_g - 2M^t(H)$	86
A.3 Positivité de la matrice d'inertie du nuage	87
A.4 $I(H) = \text{Tr}(VP)$	87
A.5 $I(H_1 \oplus H_2) = I(H_1) + I(H_2)$	88
A.6 Solution de $\max_{\sum_i a_i^2=1} \sum_{i=1}^p \lambda_i a_i^2$	88
A.7 Lemme fondamental	88
B ACP dans le cadre euclidien général	90
C Suite des preuves et données de l'analyse Budget/Temps	93
C.1 expression matricielle d'une projection M -orthogonale	93
C.2 $I(H) = \text{Tr}(VP)$	93
C.3 $I(H_1 \oplus H_2) = I(H_1) + I(H_2)$	94
C.4 Solution de $\max_{\sum_i a_i^2=1} \sum_{i=1}^p \lambda_i a_i^2$	94
C.5 Lemme fondamental	94
C.6 Données tirées de [JAM76]	95
D Algorithme NIPALS	96
D.1 Sans données manquantes	96
D.2 Avec données manquantes	97
E Rappel sur le χ^2	98
E.1 Le test du χ^2	98
E.2 Le χ^2 comme mesure de l'écart à l'indépendance	98
E.3 Distance du χ^2	102
F Table de valeurs du χ^2	104

G Prolongements de l'annexe E	106
G.1 Preuve laissée au lecteur	106
G.2 Développements annexes	107
Table des matières détaillée	109

Bibliographie

- [AAV99] F.A. OCAN A, A.M. AGUILERA, and M.J. VALDERRAMA. Functional principal components analysis by choice of norm. *Journal of multivariate analysis*, 71 :262–276, 1999.
- [ABF⁺] T. ANTONIADOU, P. BESSE, A.L. FOUGERES, C. LE GALL, and D. STEPHENSON. L’oscillation atlantique nord et son influence sur le climat européen. [www](#).
- [ADZ90] J.P. AURAY, G. DURU, and A. ZIGHED. *Analyse des données multidimensionnelles*. Editions Alexandre Lacassagne, 1990.
- [AND63] T.W. ANDERSON. Asymptotic theory for principal component analysis. *Ann. Math. Statis.*, 34 :122–148, 1963.
- [BC73] J.P. BENZECRI and COLLABORATEURS. *L’analyse de données -2- L’analyse des correspondances*. Dunod, 1973.
- [BC96] P. BESSE and H. CARDOT. Approximation spline de la prévision d’un processus fonctionnel autorégressif d’ordre 1. *Revue Canadienne de Statistique/Canadian Journal of Statistics*, 24 :467–487, 1996.
- [BCFF88] P. BESSE, H. CAUSSINUS, L. FERRE, and J. FINE. Principal component analysis and optimization of graphical displays. *statistics*, 19(2) :301–312, 1988.
- [BER00] C. BERNARD. *Introduction à la médecine expérimentale*. Flammarion, 1800.
- [BESA] P. BESSE. Insight of a dreamed pca. available at www-sv.cict.fr/lsp/Besse/.pub/dreampca.ps.
- [BESb] P. BESSE. Models for multivariate data analysis. [www](#).
- [BES91] P. BESSE. Approximation spline de l’analyse en composantes principales d’une variable aléatoire hilbertienne. *Annales de la Faculté de sciences de Toulouse*, XII(3) :329–349, 1991.
- [BUR55] C. BURT. L’analyse factorielle : méthodes et résultats. In *Analyse factorielle et ses applications*, 1955.
- [CAR97a] H. CARDOT. *Contribution à l’estimation et à la prévision statistique de données fonctionnelles*. PhD thesis, Université Paul SABATIER - Toulouse III, 1997.
- [CAR97b] H. CARDOT. Convergence en moyenne quadratique de l’analyse en composantes principales fonctionnelle en présence d’observations discrétilisées. Technical report, Laboratoire de Statistique et Probabilité. de Toulouse, mar 1997.
- [CAR00] H. CARDOT. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Nonparametric Statistics*, 12 :503–538, 2000.
- [CAU86] H. CAUSSINUS. Models and uses of principal components analysis. *Multidimensional data analysis*, 1986.
- [CIB83] P. CIBOIS. *L’analyse factorielle*. Que sais-je ? Presses Universitaires de France, 1983.
- [CNR55] CNRS, editor. *L’analyse factorielle et ses applications*. Editions du centre national de la recherche scientifique, 1955.

- [DPR82] J. DAUXOIS, A. POUSSE, and Y. ROMAIN. Asymptotic theory for the principal component analysis of a vector random function : some applications to statistical inference. *Journal of Multivariate analysis*, 12 :136–154, 1982.
- [DR99] L. DECHEVSKY and J.O. RAMSAY. Penalized wavelet estimation with besov regularity constraints. Technical Report Research report CRM-2602, Université de Montréal, 1999.
- [FL98] J. Q. FAN and S. K. LIN. Test of significance when data are curves. *Journal of the American Statistical Association*, 93(443) :1007–1021, jan 1998.
- [FLU97] B. FLURY. A first course in multivariate statistics. Springer Texts in Statistics. Springer, 1997.
- [HEC97] N. HECKMAN. The theory and application of penalized least squares methods or reproducing kernel hilbert spaces made easy. <ftp://newton.stat.ubc.ca/pub/nancy/PLS.ps>, aug 1997.
- [HOT33a] H. HOTELLING. Analysis of a complex of statistical variables into principal components. *The journal of educational psychology*, 24 :417–441, september 1933.
- [HOT33b] H. HOTELLING. Analysis of a complex of statistical variables into principal components. *The journal of educational psychology*, 24 :498–520, october 1933.
- [JAI89] A.K. JAIN. *Fundamentals of digital image processing*. Prentice-Hall International Editions, 1989.
- [JAM76] MICHEL JAMBU. Comparaison d'un modèle factoriel et d'un modèle hiérarchique - application à l'analyse des budgets-temps. *Consommation*, pages 69–108, 1976.
- [JH99] G. JAMES and T. HASTIE. Principal component models for sparse functional data. Technical report, Department of statistics, Stanford University, jun 1999.
- [JOL86] I.T. JOLLIFFE. *Principal component analysis*. Springer, 1986.
- [KG92] Alois KNEIP and Theo GASSER. Statistical tools to analyse data representing a sample of curves. *The annals of statistics*, 20(3) :1266–1305, 1992.
- [KLE73] J. KLEFFE. Principal components of random variables with values in a separable hilbert space. *Mathematische operationsforchung und statistik*, 4 :391–406, 1973.
- [KLMR00] A. KNEIP, X. LI, K. MACGIBBON, and J. RAMSAY. Curve registration by local regression. *The Canadian Journal of Statistics*, 28 :in press, 2000.
- [LEF83] J. LEFEBVRE. *Introduction aux analyses statistiques multidimensionnelles*. MASSON, 1983.
- [LL84] L. LEGENDRE and P. LEGENDRE. *Ecologie numérique - tome 1 - Le traitement des données écologiques*. Masson, 1984.
- [LMP95] L. LEBART, A. MORINEAU, and M. PIRON. *Statistique exploratoire multidimensionnelle*. Dunod, 1995.
- [LR98] X. LI and J.O. RAMSAY. Curve registration. *Journal of the Royal statistical society*, 60 :351–363, 1998. Serie B.
- [OBWS89] R.O. OLSEN, E.N. BIDEN, M.P. WYATT, and D.H. SUTHERLAND. Invited paper - gait analysis and the bootstrap. *The annals of statistics*, 17(4) :1419–1440, 1989.
- [OTN99] R.T. OGDEN, K. TAKEZAWA, and S. NINOMIYA. Functional data analysis for remote sensing images. available at www.stat.sc.edu/~ogden/papers/other.html, march 1999.
- [PAG13] J. PAGÈS. *Analyse factorielle multiple avec R*. EDP Sciences, 2013.
- [PEA01] K. PEARSON. On lines and planes of closest fit to systems of points in space. *Philosophical magazine*, 2 :559–572, 1901.

- [PEE55] E.A PEEL. Résumé du symposium d'uppsala "l'analyse factorielle en psychologie". In Analyse factorielle et ses applications, 1955.
- [RAO58] C.R. RAO. Some statistical models for comparison of growth curves. Biometrics, 14 :1–17, 1958.
- [RBG95] J.O. RAMSAY, R.D. BOCK, and T. GASSEUR. Comparison of height acceleration curves in the fels, zurich, and berkeley growth data. Annals of Human Biology, 22 :413–426, 1995.
- [RD91] J.O. RAMSAY and C.J. DALZEL. Some tools for functionnal data analysis (with discussion). Journal of the Royal Statistical Society, (53) :539–572, 1991.
- [RH96] J.O. RAMSAY and N. HECKMAN. Some theory for l-spline smoothing. available at citesear.nj.nec.com/ramsay96some.html, 1996.
- [RH00] J.O. RAMSAY and N. HECKMAN. Penalized regression with modelbased penalties. The Canadian Journal of Statistics, 28 :in press, 2000.
- [RN56] F. RIESZ and B.S. NAGY. Functional analysis. London : Blackie, 1956.
- [RS91] J.A. RICE and B. SILVERMAN. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society, 53 :233–243, 1991. Serie B.
- [RS97a] J.O. RAMSAY and B.W. SILVERMAN. Functional Data Analysis. Springer, 1997.
- [RS97b] J.O. RAMSAY and B.W. SILVERMAN. Functionnal Data Analysis. Springer Series in Statistics. Springer, 1997. ISBN 0-387-94956-9.
- [RS02] J.O. RAMSAY and B.W. SILVERMAN. Applied Functional Data Analysis : Methods and Case Studies. Springer Series in Statistics. Springer, 2002.
- [RWF95] J.O. RAMSAY, X. WANG, and R. FLANAGAN. A functional data analysis of the pinch force of human fingers. Applied Statistics, 44 :17–30, 1995.
- [SAP90] G. SAPORTA. Probabilités, analyse des données et statistique. EditionsTechnip, 1990.
- [SIL95] B.W. SILVERMAN. incorporating parametric effects into functional principal components analysis. Journal of the Royal Statistical Society, 75 :673–690, 1995. Serie B.
- [SIL96] B.W. SILVERMAN. Smoothed functional principal components analysis by choice of norm. Annals of statistics, 24 :1–24, 1996.
- [SLDB96] G. SAPORTA, F. LAVALLARD, F. DAZY, and J.F. LE BARZIC. Analyse de données évolutives. Technip, 1996.
- [SPE04] C. SPEARMAN. The proof measurement association between two things. The American journal of psychology, 15 :72–101, 1904.
- [THU55] L.L. THURSTONE. Problèmes actuels et méthodes nouvelles en analyse factorielle. In Analyse factorielle et ses applications, pages 31–44. Editions du centre national de recherches scientifique, 1955.
- [VOL81] M. VOLLE. Analyse des données. Economica, 1981.