

Sixième partie

Annexes

Annexe A

L'analyse des correspondances

| | | |
|-------|--|-----|
| A.1 | Introduction | 735 |
| A.2 | Rappels de mécanique | 736 |
| A.3 | Le problème de l'analyse des tableaux de nombres positifs | 740 |
| A.3.1 | Les principes généraux | 741 |
| A.3.2 | Le cas des tableaux binaires totalement disjonctifs | 743 |
| A.4 | Nuages et produits scalaires dans \mathbb{R}^n et \mathbb{R}^p | 745 |
| A.5 | L'analyse dans \mathbb{R}^p | 746 |
| A.6 | L'analyse dans \mathbb{R}^n | 749 |
| A.7 | Lien entre les analyses dans \mathbb{R}^p et \mathbb{R}^n | 750 |
| A.8 | Retour à la dépendance | 754 |
| A.9 | Le problème de l'interprétation | 755 |
| A.10 | Contributions absolues et relatives | 756 |
| A.11 | Des variables supplémentaires | 758 |

A.1. Introduction

Des techniques d'analyse statistique variées sont souvent utilisées lors d'études didactiques réalisées à l'aide de questionnaires soumis à une population importante d'élèves. Le but est de découvrir des idées significatives parfois dissimulées dans des résultats expérimentaux touffus. Il s'agit pour l'essentiel de techniques classificatoires : analyse en composantes principales, analyse des correspondances, analyse hiérarchique ou encore analyse implicative. On consultera à ce sujet les ouvrages mentionnés dans la bibliographie.

A l'aide de ces techniques statistiques on obtient parfois des résultats fins nous permettant d'éclairer en profondeur notre compréhension des phénomènes d'apprentissage. Il importe cependant de signaler qu'il s'agit de techniques lourdes nécessitant un investissement important en temps de travail ainsi que l'emploi de logiciels spécialisés. De plus, quand on dispose des résultats d'une de ces analyses, il reste encore à les interpréter, ce qui n'est pas nécessairement l'opération la plus facile. On se rend compte alors de ce que le questionnaire de départ doit avoir été conçu avec beaucoup de soin, de façon à faire apparaître les réponses à des questions que l'on se pose. Avec le risque de ne rien extraire d'autre des résultats que ce que l'on y a mis soi-même à travers les questionnaires.

Mais même si les questionnaires ont été rédigés avec tout le soin et la rigueur nécessaires, l'interprétation des résultats reste délicate et doit être menée avec beaucoup de prudence. C'est encore plus vrai si les questionnaires, comme c'est le cas de ceux que nous analysons dans ce travail, n'ont pas été conçus en vue de faire l'objet d'une telle analyse et si de plus ils n'ont été soumis qu'à une population limitée d'individus.

A.2. Rappels de mécanique

Soit V un espace vectoriel de dimension p muni d'un produit scalaire euclidien noté $(x|y)$. La norme du vecteur x est notée $\|x\|$ (et définie par $\|x\|^2 = (x|x)$).

Un point *massif* de V est un couple (x, m) où $x \in V$ et $m \in \mathbb{R}_+$. On se donnera ci-dessous une famille finie $\mathcal{E} = (x_i, m_i)_{i=1}^n$ de points massifs de V , encore appelée un *nuage* de points massifs.

Le *centre de gravité* g de \mathcal{E} est le vecteur $g = \frac{1}{M} \sum_{i=1}^n m_i x_i$, où $M = \sum_{i=1}^n m_i$.

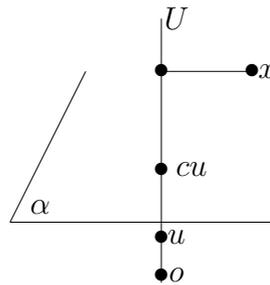
Si α est un *hyperplan affine* (c'est-à-dire le translaté d'un hyperplan vectoriel), on notera $d(x, \alpha)$ la distance de x à α .

Le *moment d'inertie* de \mathcal{E} par rapport à un hyperplan affine α est

$$\mathcal{M}(\alpha) = \sum_{i=1}^n m_i d^2(x_i, \alpha)$$

Il existe une formule plus commode pour exprimer $\mathcal{M}(\alpha)$. Pour l'obtenir, notons u un vecteur perpendiculaire à α , de longueur 1 : $\|u\| = 1$.

L'équation de l'hyperplan α est du type $(u|x) = c$ où $c \in \mathbb{R}$. Faire varier c revient à remplacer α par un hyperplan parallèle. Notons U la droite vectorielle engendrée par u . Quel que soit $p \in V$, la projection de p sur U est le vecteur $(p|u)u$. Le point de percée de U dans α est le vecteur cu . La distance de p à α vaut donc $|c - (p|u)|$.



Ainsi $d^2(p, \alpha) = c^2 - 2c(p|u) + (p|u)^2$ et

$$\mathcal{M}(\alpha) = \sum_{i=1}^n m_i (c^2 - 2c(x_i|u) + (x_i|u)^2)$$

$$= Mc^2 - 2M(g|u)c + \sum_{i=1}^n m_i(x_i|u)^2$$

Translatons α : cela revient à considérer c comme variable. $\mathcal{M}(\alpha)$ est un trinôme du second degré en c . Il admet un minimum pour $c = (g|u)$. L'hyperplan correspondant est d'équation $(u|x) = (u|g)$. Ainsi :

L'hyperplan d'une direction donnée pour lequel le moment d'inertie est minimum est celui qui comprend le centre de gravité g .

Dans la suite de cette annexe, on suppose $g = o$ ce qui revient à dire que nous avons effectué la translation $p \mapsto p - g$. Tous les hyperplans passant par g sont alors vectoriels. On va à présent essayer de déterminer quel est celui d'entre eux pour lequel le moment d'inertie est maximum.

Puisque $g = 0$, on a

$$\mathcal{M}(\alpha) = \sum_{i=1}^n m_i(x_i|u)^2$$

Si u et v sont deux vecteurs de V , définissons une fonction M sur $V \times V$ en posant

$$M(u, v) = \sum_{i=1}^n m_i(x_i|u)(x_i|v)$$

Il est clair que M est une forme bilinéaire symétrique, semi-définie positive sur V . De plus $\mathcal{M}(\alpha) = M(u, u)$. Rappelons qu'on appelle *grammienne* de M par rapport à une base (e_1, \dots, e_p) de V , la matrice M^e donnée par $M_{k\ell}^e = M(e_k, e_\ell)$.

On sait que l'espace V admet une base orthonormée (f_1, \dots, f_p) par rapport à laquelle la grammienne M^f de M est diagonale :

$$M^f = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}$$

Les droites-supports des vecteurs (f_1, \dots, f_p) sont en mécanique appelées les *axes principaux d'inertie* du nuage de points massifs. Les λ_i sont les moments d'inertie par rapport aux hyperplans vectoriels perpendiculaires aux axes principaux d'inertie. Ce sont donc des nombres positifs ou nuls.

Nous supposons toujours que les vecteurs f_1, \dots, f_p sont numérotés de façon à avoir $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$.

La détermination des vecteurs f_1, \dots, f_p se ramène au calcul des vecteurs propres et valeurs propres d'une transformation linéaire. En effet, on peut écrire

$$M(u, v) = \left(u \mid \sum_{i=1}^n m_i(x_i|v)x_i \right)$$

Par abus de notation, posons $M(v) = \sum_{i=1}^n m_i(x_i|v)x_i$. On a pour tout $u \in V$:

$$M(u, v) = (u|M(v))$$

L'application $v \mapsto M(v)$ est visiblement linéaire. De plus, elle est symétrique, ce qui signifie qu'on a, quels que soient u et v :

$$(u|M(v)) = (M(u)|v)$$

En effet $(M(u)|v) = (v|M(u)) = M(v, u) = M(u, v) = (u|M(v))$. Par rapport à n'importe quelle base orthonormée, (e_1, \dots, e_p) , la grammienne de la forme bilinéaire M coïncide avec la matrice de l'application linéaire M . En effet, les colonnes de cette dernière matrice sont les composantes des vecteurs $M(e_1), \dots, M(e_p)$. Ce sont donc bien les produits scalaires $(e_i|M(e_j)) = M(e_i, e_j)$.

Les vecteurs f_1, \dots, f_p sont donc les vecteurs propres de l'application linéaire M . Si u est le vecteur de norme 1, perpendiculaire à l'hyperplan vectoriel α , notons (u_1, \dots, u_p) sa coordonnée par rapport à (f_1, \dots, f_p) . On a

$$\mathcal{M}(\alpha) = M(u, u) = \sum_{i=1}^n \lambda_i u_i^2$$

Puisque $\sum_{i=1}^n u_i^2 = 1$, le nombre $\mathcal{M}(\alpha)$ est compris entre λ_p et λ_1 . La valeur maximum de $\mathcal{M}(\alpha)$ est λ_1 . Elle est atteinte pour $u = f_1$. Le même raisonnement montre que la valeur minimum de $\mathcal{M}(\alpha)$ vaut λ_p et est atteinte pour $u = f_p$.

Il existe d'autres moments d'inertie :

Le moment d'inertie de \mathcal{E} par rapport à l'origine est $K = \sum_{i=1}^n m_i \|x_i\|^2$. Il est immédiat que K est la trace de la matrice M^e , c'est-à-dire $\sum_{j=1}^p \lambda_j$.

Le moment d'inertie de \mathcal{E} par rapport à une droite vectorielle D est

$$\mathcal{M}(D) = \sum_{i=1}^n m_i d^2(x_i, D)$$

Si D^\perp est l'hyperplan vectoriel perpendiculaire à D , on a d'après le théorème de Pythagore $\|x_i\|^2 = d^2(x_i, D) + d^2(x_i, D^\perp)$.

Donc

$$\boxed{K = \mathcal{M}(D) + \mathcal{M}(D^\perp)}$$

La valeur minimum de $\mathcal{M}(D)$ vaut $K - \lambda_1$. Elle est atteinte lorsque D est la droite vectorielle engendrée par le vecteur propre f_1 . Cette droite — le premier axe principal d'inertie — est celle qui approche le mieux le nuage.

Quelle signification attribuer aux autres vecteurs propres f_2, \dots, f_p et valeurs propres λ_2, λ_p ? Projetons \mathcal{E} sur l'hyperplan perpendiculaire H_1 à f_1 : nous obtenons une nouvelle famille \mathcal{E}' de points massifs. Soit \mathcal{M}' la forme bilinéaire sur H_1 associée à \mathcal{E}' . Sa matrice par rapport à (f_2, \dots, f_p) est

$$\begin{pmatrix} \lambda_2 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix}$$

La plus grande valeur propre de cette matrice est λ_2 . Ainsi, λ_2 et f_2 jouent par rapport à \mathcal{E}' le même rôle que λ_1 et f_1 par rapport à \mathcal{E} . On interprète f_3, \dots, f_p et $\lambda_3, \dots, \lambda_p$ de façon analogue en projetant sur des sous-espaces de plus en plus petits.

Rappelons encore que, la base (f_1, \dots, f_p) étant orthonormée, le développement d'un vecteur x dans cette base est donné par

$$x = \sum_{j=1}^p (f_j | x) f_j$$

A.3. Le problème de l'analyse des tableaux de nombres positifs

La statistique fournit souvent des tableaux de nombres dont il faut essayer d'extraire le plus d'informations possibles de la façon la plus synthétique possible. Ce qui suit s'appliquera aux tableaux de nombres positifs.

A.3.1 Les principes généraux

On dispose donc d'un tableau P de taille $n \times p$, constitué de nombres positifs notés k_{ij} , ($i = 1, \dots, n$, $j = 1, \dots, p$). L'origine de ce tableau peut être très variée : il peut s'agir de la liste des résultats d'élèves aux différents items d'un test, de n observations de p variables etc. D'une façon générale, les lignes représentent des « variables » d'une certaine espèce, les colonnes des « variables » d'une autre espèce.

Deux variables d'espèces différentes « interagissent », les résultats de toutes les interactions possibles étant les nombres k_{ij} consignés dans le tableau donné. Deux variables de même espèce peuvent être bien corrélées ou non. Le but recherché est de décrire aussi simplement que possible les interactions entre variables d'espèces différentes, ainsi que les corrélations entre variables de même espèce. Si le tableau est de grandes dimensions, cela nécessite un traitement mathématique destiné à réduire le nombre de paramètres en ne conservant que les plus significatifs.

On va commencer par donner aux éléments du tableau une signification probabiliste. Dans ce but, introduisons une probabilité sur l'ensemble $\Omega = \{1, \dots, n\} \times \{1, \dots, p\}$ en divisant chaque nombre k_{ij} par $S = \sum_{i=1}^n \sum_{j=1}^p k_{ij}$. On pose $p_{ij} = k_{ij}/S$. Les p_{ij} sont positifs, de somme 1, ils définissent donc bien une probabilité sur Ω . Choisir un élément de Ω revient à choisir une case du tableau P , avec la probabilité associée à cette case.

A la ligne n° i , associons l'événement

$$A_i = \{(i, 1), \dots, (i, p)\}$$

De même, à la colonne n° j , associons l'événement

$$B_j = \{(1, j), \dots, (n, j)\}$$

On pose ensuite $p_{i\bullet} = \sum_{j=1}^p p_{ij}$ et $p_{\bullet j} = \sum_{i=1}^n p_{ij}$.

Les événements A_1, \dots, A_n constituent une partition de Ω en n classes (une par ligne du tableau de données). De même B_1, \dots, B_p constituent une deuxième partition de Ω dont les classes sont associées aux colonnes du tableau. Les nombres $p_{i\bullet}$ et $p_{\bullet j}$ sont respectivement la probabilité de A_i et celle de B_j , ce sont les probabilités marginales sur l'ensemble Ω . Le nombre p_{ij} est la probabilité de $A_i \cap B_j$. Notons au passage les relations

$$\sum_{i=1}^n p_{i\bullet} = 1 \quad \sum_{j=1}^p p_{\bullet j} = 1$$

qui montrent que les probabilités marginales constituent bien des probabilités sur $\{A_1, \dots, A_n\}$ et $\{B_1, \dots, B_p\}$.

Comme on le sait, ces partitions sont indépendantes si et seulement si quels que soient i et j :

$$\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(B_j)$$

c'est-à-dire si

$$p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$$

Nous pouvons admettre qu'aucun des nombres $p_{i\bullet}$ ou $p_{\bullet j}$ n'est nul : si on avait par exemple $p_{i\bullet} = 0$, alors tous les nombres k_{ij} de la i^{e} ligne du tableau T seraient nuls et on supprimerait cette ligne (sans intérêt) du tableau. Les deux partitions sont alors indépendantes si et seulement si le tableau U donné par

$$U_{ij} = \frac{p_{ij}}{p_{i\bullet} \cdot p_{\bullet j}}$$

ne contient que des nombres 1. En général, ce ne sera pas le cas, et on cherchera à analyser ce tableau.

Considérons la colonne n^o j du tableau U . Nous la notons $U_{\star j}$. Nous pouvons la considérer comme un vecteur de \mathbb{R}^n ou comme les valeurs d'une variable aléatoire définie sur la catégorie d'épreuves $\{A_1, \dots, A_n\}$. L'espérance mathématique de cette variable aléatoire est

$$\sum_{i=1}^n p_{i\bullet} U_{ij} = \sum_{i=1}^n \frac{p_{ij}}{p_{\bullet j}} = 1$$

De même, chaque ligne $U_{i\star}$ du tableau U peut être considérée comme une variable aléatoire d'espérance 1 sur la catégorie d'épreuves $\{B_1, \dots, B_p\}$. En fait, n'importe quel vecteur de \mathbb{R}^p peut être considéré comme une variable aléatoire sur $\{B_1, \dots, B_p\}$ et n'importe quel vecteur de \mathbb{R}^n comme une variable aléatoire sur $\{A_1, \dots, A_n\}$.

A.3.2 Le cas des tableaux binaires totalement disjonctifs

Nous n'appliquerons les principes énoncés au point précédent que dans un cas très particulier : celui de tableaux binaires disjonctifs.

Il s'agira en l'occurrence de résultats de questionnaires soumis à un ensemble de n élèves et comportant un certain nombre q de questions. Confronté à une question, chaque élève produit un et un seul des trois comportements suivants ⁽¹⁾

1. Il fournit une réponse considérée comme bonne.
2. Il fournit une réponse considérée comme mauvaise.
3. Il ne fournit pas de réponse.

A chacun de ces comportements va correspondre une colonne du tableau de données P . Celui-ci aura donc $p = 3q$ colonnes. Si l'élève n° i réussit la question n° h , on placera le nombre 1 dans la case située à l'intersection de la ligne n° i et de la colonne n° $3h$, et 0 dans les cases de la même ligne situées dans les colonnes n° $3h + 1$ et $3h + 2$. On procède de façon analogue si l'élève a échoué ou s'est abstenu à la question n° h .

Les données sont binaires : chacun des nombres du tableau vaut 0 ou 1. Elles sont aussi totalement disjonctives en ce sens que dans toute ligne, on a toujours un et un seul nombre 1 dans un triple de cases portant les numéros $3h$, $3h + 1$, $3h + 2$.

Ainsi, dans toute ligne figurent exactement q nombres 1, de sorte que la somme S de tous les éléments du tableau vaut nq et toutes les probabilités p_{ij} (voir le point précédent) valent 0 ou $\frac{1}{nq}$.

Il est clair que dans ce modèle, nous pouvons identifier un événement élémentaire $(i, j) \in \Omega$ à une « observation » : nous observons le comportement n° j chez l'élève n° i . L'événement A_i s'identifie à l'élève n° i (plus exactement à l'ensemble des comportements qu'il a eus). L'événement B_j s'identifie au comportement n° j (à l'ensemble des élèves qui ont eu le comportement n° j). \mathbb{R}^p est ainsi l'espace des élèves, cependant que \mathbb{R}^n est l'espace des comportements de réponse.

Puisque dans toute ligne figurent exactement q nombres 1, toutes les probabilités marginales $p_{i\bullet}$ valent $\frac{1}{n}$: tous les élèves ont même probabilité d'être choisis. Il n'en est pas de même pour les comportements : certains sont plus fréquents que d'autres. Comme dans le cas général, nous supprimons du tableau les comportements qui ne sont jamais observés c'est-à-dire les colonnes pour lesquelles la probabilité marginale $p_{\bullet j}$ vaut 0. Cela ne change rien ni à la valeur de S ni à celle des $p_{i\bullet}$.

⁽¹⁾ On se limite souvent à deux comportements : la réussite et l'échec.

Le tableau U mesure l'indépendance des deux partitions, l'une constituée par les élèves, l'autre par les comportements. Dire que ces deux partitions sont indépendantes signifie qu'en fait tous les élèves ont répondu de la même façon à toutes les questions :

$$U_{ij} = 1 \Leftrightarrow p_{ij} = \frac{1}{n} p_{\bullet j}$$

Comme $p_{\bullet j} \neq 0$, on a $p_{ij} \neq 0$ et par conséquent $p_{ij} = \frac{1}{nq}$. Ainsi toutes les données k_{ij} du tableau T épuré des colonnes qui ne contenaient que des 0, valent 1. Autrement dit tous les élèves ont choisi le même comportement à toutes les questions. On ne passera évidemment pas son temps à appliquer une quelconque technique statistique à un tel tableau.

Le but de l'analyse des correspondances est de dégager des conclusions à partir de l'observation des dépendances (au sens probabiliste du terme) entre élèves et comportements.

A.4. Nuages et produits scalaires dans \mathbb{R}^n et \mathbb{R}^p

Nous reprenons notre description de l'analyse des correspondances dans le cas général. Nous allons étudier les corrélations entre les colonnes $U_{\star j}$ d'une part, entre les lignes $U_{i\star}$ d'autre part. Commençons par choisir dans \mathbb{R}^n et \mathbb{R}^p des produits scalaires ayant une signification statistique : soient $u, v \in \mathbb{R}^n$ et $x, y \in \mathbb{R}^p$. Alors

$$(u|v) = \sum_{i=1}^n p_{i\bullet} u_i v_i, \text{ et } (x|y) = \sum_{j=1}^p p_{\bullet j} x_j y_j$$

Si u et v sont des variables aléatoires de moyenne nulle, $(u|v)$ est simplement la covariance de u et v . Deux vecteurs orthogonaux sont des variables aléatoires non corrélées. Le carré de la norme d'un vecteur est la variance de ce vecteur considéré comme étant une variable aléatoire.

- Dans \mathbb{R}^p , muni du produit scalaire ci-dessus, nous disposons d'un nuage de n points : les n lignes $U_{i\star}$. Ces points deviennent massifs si nous attribuons à chaque $U_{i\star}$ le poids $p_{i\bullet}$. Le nuage $(U_{i\star}, p_{i\bullet})_{i=1}^n$ est un bon représentant de l'ensemble des événements A_1, \dots, A_n . La technique décrite au point A.2 va permettre de trouver une base orthonormée de \mathbb{R}^p qui soit adaptée à ce nuage. On décomposera ainsi les lignes $U_{i\star}$ en combinaisons linéaires de variables aléatoires non corrélées.
- Dans \mathbb{R}^n , nous disposons d'un nuage de p points massifs $(U_{\star j}, p_{\bullet j})$, qui représente l'ensemble des événements B_1, \dots, B_p . On lui appliquera également la technique du point A.2, et on obtiendra une décomposition des colonnes $U_{\star j}$ en combinaisons linéaires de variables aléatoires non corrélées.

A.5. L'analyse dans \mathbb{R}^p

Commençons par ramener à l'origine le centre de gravité des points du nuage $(U_{i\bullet}, p_{i\bullet})_{i=1}^n$.

Ce centre de gravité (qu'on appellera en statistique le vecteur moyen) est

$$m = \sum_{i=1}^n p_{i\bullet} U_{i\bullet}$$

Donc pour tout j , $m_j = \sum_{i=1}^n p_{i\bullet} U_{ij} = 1$. Ainsi $m = (1, 1, \dots, 1)$. On remplacera donc le nuage $(U_{i\bullet}, p_{i\bullet})$ par le nuage $(V_{i\bullet}, p_{i\bullet})$ où

$$V_{ij} = \frac{p_{ij} - p_{i\bullet} p_{\bullet j}}{p_{i\bullet} p_{\bullet j}}$$

En tant que variable aléatoire, $V_{i\bullet}$ est d'espérance nulle puisque nous avons vu que $U_{i\bullet}$ était d'espérance 1.

Nous ne manipulerons plus que des variables aléatoires de moyenne nulle, c'est-à-dire des vecteurs u appartenant à l'hyperplan α_0 d'équation $\sum_{j=1}^p p_{\bullet j} u_j = 0$ (par rapport à la base canonique).

Le moment d'inertie du nuage $(V_{i\bullet}, p_{i\bullet})$ par rapport à l'origine sera appelé la *variance totale* du nuage :

$$\begin{aligned} K &= \sum_{i=1}^n p_{i\bullet} \|V_{i\bullet}\|^2 \\ &= \sum_{i=1}^n p_{i\bullet} \sum_{j=1}^p p_{\bullet j} \left(\frac{p_{ij} - p_{i\bullet} p_{\bullet j}}{p_{i\bullet} p_{\bullet j}} \right)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^p \frac{(p_{ij})^2}{p_{i\bullet} p_{\bullet j}} - 1 \end{aligned}$$

Le moment d'inertie par rapport à un hyperplan vectoriel α sera appelé la variance au long de la droite α^\perp , ou encore la *part de la variance* due à la droite α^\perp . Cette quantité mesure en effet la dispersion des projections des points du nuage sur la droite α^\perp .

Rechercher les axes d'inertie du nuage, c'est diagonaliser la forme bilinéaire symétrique et semi-définie positive \mathcal{M} donnée par

$$\mathcal{M}(u, v) = \sum_{i=1}^n p_{i\bullet} (V_{i\star}|u)(V_{i\star}|v)$$

où le produit scalaire est celui qui a été défini précédemment :

$$(V_{i\star}|u) = \sum_{j=1}^p p_{\bullet j} V_{ij} u_j$$

C'est aussi diagonaliser la transformation linéaire M donnée par

$$M(u) = \sum_{i=1}^n p_{i\bullet} (V_{i\star}|u) V_{i\star}$$

On notera (f_1, \dots, f_p) une base de \mathbb{R}^p constituée de vecteurs propres normés de M , et $\lambda_1, \dots, \lambda_p$ les valeurs propres correspondantes. Ces vecteurs propres seront appelés des *facteurs*. On les numérote dans l'ordre décroissant des valeurs propres : $\lambda_1 \geq \lambda_2 \geq \dots$.

Comme on l'a vu au point A.2, la somme des λ_i est la variance totale du nuage, cependant que λ_i est la part de la variance expliquée par le i^{e} facteur.

Comme tous les vecteurs $V_{i\star}$ appartiennent à l'hyperplan α_0 d'équation $\sum_{j=1}^p p_{\bullet j} u_j = 0$, nous connaissons automatiquement une direction propre de M la direction perpendiculaire à α_0 . En effet, si $u \perp \alpha_0$, on a $M(u) = 0$. La p^{e} valeur propre est donc $\lambda_p = 0$. Le vecteur propre normé correspondant est sans intérêt : la part de variance qu'il explique est nulle.

Parmi les valeurs propres de M , d'autres que λ_p pourraient être nulles. Supposons qu'il y ait exactement ρ valeurs propres non nulles. Chaque variable aléatoire $V_{i\star}$ est alors une combinaison linéaire des variables deux à deux non corrélées f_1, \dots, f_ρ :

$$V_{i\star} = \sum_{j=1}^{\rho} (V_{i\star}|f_j) f_j$$

Calculons la variance totale K dans le cas du tableau binaire disjonctif considéré précédemment. On a alors $p = 3q$, $S = nq$, pour tout i $p_{i\bullet} = \frac{1}{n}$ et $p_{ij} = \frac{k_{ij}}{S}$ où tous les k_{ij} valent 0 ou 1. Il vient ainsi

$$\begin{aligned} K &= \sum_{i=1}^n \sum_{j=1}^p \frac{(p_{ij})^2}{p_{i\bullet} p_{\bullet j}} - 1 \\ &= n \sum_{j=1}^p \frac{1}{p_{\bullet j}} \sum_{i=1}^n \frac{k_{ij}^2}{n^2 q^2} - 1 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{nq^2} \sum_{j=1}^p \frac{1}{p_{\bullet j}} \sum_{i=1}^n k_{ij} - 1 \\
&= \frac{1}{nq^2} \sum_{j=1}^p \frac{1}{p_{\bullet j}} nqp_{\bullet j} - 1 \\
&= \frac{1}{nq^2} npq - 1 = 2
\end{aligned}$$

La variance totale vaut dans le cas d'un questionnaire prévoyant pour chaque question trois comportements de réponse vaut 2 ⁽²⁾.

⁽²⁾ Si on ne s'intéressait qu'à deux comportements de réponse (réussite et échec), la variance totale vaudrait 1.

A.6. L'analyse dans \mathbb{R}^n

Dans \mathbb{R}^n , l'analyse est évidemment semblable à celle de \mathbb{R}^p : on permute le rôle des lignes et colonnes. On considère donc cette fois le nuage de points massifs $(U_{\star j}, p_{\bullet j})_{j=1}^p$ de \mathbb{R}^n . Son centre de gravité est le point $(1, \dots, 1) \in \mathbb{R}^n$. On le translate à l'origine. On est ramené à considérer le nuage $(V_{\star j}, p_{\bullet j})$ où comme au point précédent

$$V_{ij} = \frac{p_{ij} - p_{i\bullet} \cdot p_{\bullet j}}{p_{i\bullet} \cdot p_{\bullet j}}$$

Le nouveau nuage est inclus à l'hyperplan α_0 d'équation $\sum_{i=1}^n p_{i\bullet} u_i = 0$. Sa variance totale est la même que celle du nuage de \mathbb{R}^p . La forme bilinéaire à diagonaliser est cette fois

$$\mathcal{M}'(u, v) = \sum_{j=1}^p p_{\bullet j} (V_{\star j} | u) (V_{\star j} | v)$$

Il revient au même de diagonaliser la transformation linéaire de \mathbb{R}^p donnée par

$$M'(v) = \sum_{j=1}^p p_{\bullet j} (V_{\star j} | v) V_{\star j}$$

Une base orthonormée de \mathbb{R}^n constituée de vecteurs propres de M' sera notée (g_1, \dots, g_n) . De même qu'au paragraphe précédent, la dernière valeur propre vaut 0, le facteur correspondant est sans intérêt.

A.7. Lien entre les analyses dans \mathbb{R}^p et \mathbb{R}^n

Les deux transformations linéaires M et M' étant toutes deux définies par le tableau V , il doit exister un lien entre vecteurs propres de M et vecteurs propres de M' .

Ce lien permettra de n'appliquer qu'une fois la technique de diagonalisation. Pour le décrire, rappelons d'abord ce qu'est la transposée d'une application linéaire A de \mathbb{R}^p dans \mathbb{R}^n : c'est une application linéaire, notée A^τ de \mathbb{R}^n dans \mathbb{R}^p telle que

$$\forall u \in \mathbb{R}^n, \forall v \in \mathbb{R}^p : (A(u)|v) = (u|A^\tau(v))$$

On montre sans peine qu'une telle transformation linéaire existe et est unique, et cela quelle que soit A ; mais nous n'aurons besoin de ce résultat que dans le cas particulier suivant :

$$A : \mathbb{R}^p \rightarrow \mathbb{R}^n : u \mapsto \sum_{i=1}^n (V_{i\star}|u) e_i$$

où (e_1, \dots, e_n) est la base canonique \mathbb{R}^n . Calculons A^τ . Si $v \in \mathbb{R}^n$, on a

$$(A(u)|v) = \sum_{i=1}^n (V_{i\star}|u)(e_i|v) = \left(\sum_{i=1}^n (e_i|v) V_{i\star} \mid u \right) = \left(u \mid \sum_{i=1}^n (e_i|v) V_{i\star} \right)$$

Ainsi

$$A^\tau(v) = \sum_{i=1}^n (e_i|v) V_{i\star}$$

Calculons à présent $A^\tau \cdot A$ (application de \mathbb{R}^p dans \mathbb{R}^p) et $A \cdot A^\tau$ (application de \mathbb{R}^n dans \mathbb{R}^n) :

1. Si $u \in \mathbb{R}^p$:

$$\begin{aligned} A^\tau A(u) &= \sum_{i=1}^n (V_{i\star}|u) A^\tau(e_i) \\ &= \sum_{i=1}^n (V_{i\star}|u) \sum_{j=1}^n (e_j|e_i) V_{j\star} \end{aligned}$$

Or $(e_j|e_i) = 0$ si $j \neq i$ et $(e_i|e_i) = p_{i\bullet}$. Donc

$$A^T A(u) = \sum_{i=1}^n p_{i\bullet} (V_{i\star}|u) V_{i\star} = M(u)$$

2. Si $v \in \mathbb{R}^n$:

$$\begin{aligned} AA^T(v) &= \sum_{i=1}^n (e_i|v) A(V_{i\star}) \\ &= \sum_{i=1}^n \sum_{j=1}^n (e_i|v) (V_{j\star}|V_{i\star}) e_j \end{aligned}$$

Comme $(e_i|v) = p_{i\bullet} v_i$ et $(V_{j\star}|V_{i\star}) = \sum_{h=1}^p p_{\bullet h} V_{jh} V_{ih}$, il vient

$$\begin{aligned} AA^T(v) &= \sum_{h=1}^p p_{\bullet h} \left(\sum_{i=1}^n p_{i\bullet} v_i V_{ih} \right) \left(\sum_{j=1}^n V_{jh} e_j \right) \\ &= \sum_{h=1}^p (V_{\star h}|v) V_{\star h} = M'(v) \end{aligned}$$

Ainsi $M = A^T A$ et $M' = AA^T$. Ces relations ont plusieurs conséquences importantes.

- Si $u \in \mathbb{R}^p$ est vecteur propre de M ; de valeur propre $\lambda \neq 0$, alors $A(u)$ est vecteur propre de M' de même valeur propre λ .
En effet : $M(u) = \lambda u \Rightarrow A^T A(u) = \lambda u \Rightarrow AA^T A(u) = \lambda A(u)$. Donc $M' A(u) = \lambda A(u)$. Pour montrer que $A(u)$ est vecteur propre de M' , de valeur propre λ , il reste à établir $A(u) \neq 0$. Or si on avait $A(u) = 0$, il viendrait $M(u) = 0$, u serait vecteur propre de M de valeur propre nulle alors que $\lambda \neq 0$.
- De même, si v est vecteur propre de M' , de valeur propre non nulle, alors $A^T(v)$ est vecteur propre de M , avec la même valeur propre.
- On déduit de ce qui précède que l'application linéaire A induit un isomorphisme du sous-espace propre E_λ de \mathbb{R}^p associé à une valeur propre λ de M sur le sous-espace propre E'_λ de \mathbb{R}^n associé à la même valeur propre λ de M' .
En effet, on vient de voir que A applique E_λ dans E'_λ . Cette application est injective puisque, si $u \in E_\lambda$, on a prouvé l'implication $u \neq 0 \Rightarrow A(u) \neq 0$. Elle est aussi surjective : si $v \in E'_\lambda$, alors $v = \frac{1}{\lambda} M'(v) = A\left(\frac{1}{\lambda} A^T(v)\right)$.
- Puisque les sous-espaces E_λ et E'_λ ont même dimension, les applications linéaires M et M' ont même rang (même nombre de valeurs propres non nulles).

Connaissant les facteurs f_1, \dots, f_p dans \mathbb{R}^p , nous allons maintenant pouvoir trouver les facteurs g_1, \dots, g_p dans \mathbb{R}^n . En effet, $A(f_i)$ est vecteur propre de M' de même valeur propre λ_i que f_i . Donc $A(f_i)$ est un multiple de g_i . En fait, puisque g_i est normé, et n'est défini qu'à un nombre de module 1 près, on peut choisir : $g_i = \frac{A(f_i)}{\|A(f_i)\|}$.

Or $\|A(f_i)\|^2 = (A(f_i) | A(f_i)) = (A^T A(f_i) | f_i) = (M(f_i) | f_i) = \lambda_i \|f_i\|^2$.

Donc pour tout $i = 1, \dots, \rho$, $g_i = \frac{1}{\sqrt{\lambda_i}} A(f_i)$. De même $f_i = \frac{1}{\sqrt{\lambda_i}} A^\tau(g_i)$.

Reprenons alors la formule $V_{i\star} = \sum_{h=1}^{\rho} (V_{i\star}|f_h) f_h$. Si on note f_{hj} la h^e composante de f_h par rapport à la base canonique de \mathbb{R}^p , on en déduit

$$V_{ij} = \sum_{h=1}^{\rho} (V_{i\star}|f_h) f_{hj}$$

Or, si (e_1, \dots, e_p) est la base canonique de \mathbb{R}^p , on a $(f_h|e_j) = p_{\bullet j} f_{hj}$ et

$$\begin{aligned} (f_h|e_j) &= \frac{1}{\sqrt{\lambda_h}} (A^\tau(g_h)|e_j) \\ &= \frac{1}{\sqrt{\lambda_h}} (g_h|A(e_j)) \\ &= \frac{1}{\sqrt{\lambda_h}} \left(g_h \left| \sum_{i=1}^n (V_{i\star}|e_j) e_i \right. \right) \\ &= \frac{1}{\sqrt{\lambda_h}} \left(g_h \left| \sum_{i=1}^n p_{\bullet j} V_{ij} e_i \right. \right) \\ &= p_{\bullet j} \frac{1}{\sqrt{\lambda_h}} \left(g_h \left| \sum_{i=1}^n V_{ij} e_i \right. \right) \\ &= p_{\bullet j} \frac{1}{\sqrt{\lambda_h}} (g_h|V_{\star j}) \end{aligned}$$

Par conséquent

$$f_{hj} = \frac{1}{\sqrt{\lambda_h}} (g_h|V_{\star j})$$

De même

$$g_{hi} = \frac{1}{\sqrt{\lambda_h}} (f_h|V_{i\star})$$

(où les g_{hi} sont les composantes de g_h par rapport à la base canonique de \mathbb{R}^n).

Donc

$$V_{ij} = \sum_{h=1}^{\rho} \frac{1}{\sqrt{\lambda_h}} (V_{i\star}|f_h) (V_{\star j}|g_h)$$

ou

$$V_{ij} = \sum_{h=1}^{\rho} \sqrt{\lambda_h} g_{hi} f_{hj}$$

Retournant aux p_{ij} , on obtient

$$\begin{aligned} p_{ij} &= p_{i\bullet} p_{\bullet j} \left(1 + \sum_{h=1}^{\rho} \frac{1}{\sqrt{\lambda_h}} (V_{i\star} | f_h)(V_{i\star} | g_h) \right) \\ &= p_{i\bullet} p_{\bullet j} \left(1 + \sum_{h=1}^{\rho} \sqrt{\lambda_h} g_{hi} f_{hj} \right) \end{aligned}$$

On peut donc reconstruire le tableau initial, connaissant les probabilités marginales et les facteurs.

A.8. Retour à la dépendance

Comme on l'a déjà noté, l'égalité $p_{ij} = p_{i\bullet}p_{\bullet j}$ est vraie si et seulement si les événements A_i et B_j sont indépendants. Le terme $\sum_{h=1}^{\rho} \frac{1}{\sqrt{\lambda_h}} (V_{i\star}|f_h)(V_{i\star}|g_h)$ est donc une mesure de la dépendance mutuelle de A_i et B_j . Il se décompose en termes $\frac{1}{\sqrt{\lambda_h}} (V_{i\star}|f_h)(V_{i\star}|g_h)$.

On peut donner une interprétation de ces termes en introduisant fictivement des « événements modèles de type A », correspondant aux facteurs f_i et des « événements modèles de type B », correspondant aux facteurs g_j . Dans le cas d'un questionnaire administré aux élèves d'une classe, osera-t-on parler d'« élèves modèles » et de « comportements modèles » ?

$V_{i\star}$ étant une combinaison linéaire des f_j , on pourrait dire que A_i est une « superposition » des événements modèles de type A. B_j serait une superposition des événements modèles de type B. La dépendance entre l'événement modèle de type A associé à f_i et l'événement modèle de type B associé à g_j serait dans cette optique mesurée par 0 si $i \neq j$ et $\frac{1}{\sqrt{\lambda_i}}$ si $i = j$. La dépendance entre A_i et B_j s'obtient alors par combinaison linéaire.

Mathématiquement, cela revient à définir sur $\mathbb{R}^p \times \mathbb{R}^n$ une nouvelle forme bilinéaire :

$$D(u, v) = \sum_{h=1}^{\rho} \frac{1}{\sqrt{\lambda_h}} (u|f_h)(v|g_h)$$

Par exemple

$$\begin{aligned} D(V_{i\star}, V_{j\star}) &= \frac{p_{ij}}{p_{i\bullet}p_{\bullet j}} - 1 & D(V_{i\star}, g_h) &= \frac{1}{\sqrt{\lambda_h}} (V_{i\star}|f_h) \\ D(f_i, g_j) &= \frac{1}{\sqrt{\lambda_i}} \delta_{ij} & D(f_h, V_{\star j}) &= \frac{1}{\sqrt{\lambda_h}} (g_h|V_{\star j}) \end{aligned}$$

A.9. Le problème de l'interprétation

Décomposer un événement en une superposition d'événements modèles n'est intéressant que s'il est possible de donner une signification « claire » à ces derniers.

Par exemple si on parvient à décomposer le comportement d'un élève lors d'un examen en une superposition de comportements élémentaires faciles à étudier, on a progressé d'un pas important dans la connaissance que l'on a de cet élève, sur le type d'erreurs qu'il commet, sur son stade de développement mathématique.

Attribuer une signification aux événements élémentaires est donc un objectif important, mais pour lequel aucune technique (statistique ou autre) ne s'impose clairement. Dans la pratique on se contente d'essayer de trouver une telle signification pour les facteurs qui expliquent une partie non négligeable de la variance totale, considérant les autres comme une sorte de « bruit de fond ». Cela se fait en cherchant des caractéristiques communes aux événements dont la composante suivant le facteur considéré n'est pas négligeable.

Pour faire cette recherche, on s'aide généralement de graphiques cartésiens plans tel celui représentant les projections des vecteurs V_{i^*} sur le plan (f_1, f_2) et les projections des vecteurs V_{*j} sur le plan (g_1, g_2) .

Dans la pratique, on identifie les plans (f_1, f_2) et (g_1, g_2) (on identifie f_1 à g_1 , et f_2 à g_2). Cette identification se justifie par les formules qui viennent d'être établies. Par exemple, on note que si $u \in \mathbb{R}^p$, on a $D(u, g_h) = \frac{1}{\sqrt{\lambda_h}}(u|f_h)$. La position du vecteur u dans le plan des deux premiers facteurs fournit donc des renseignements tant sur les corrélations de ce vecteur u avec les facteurs f_1 et f_2 que sur ses dépendances avec g_1 et g_2 .

On réalise aussi parfois les graphiques correspondant aux facteurs 1 et 3 ou 2 et 3, etc. Mais, comme on l'a écrit plus haut, on ne va généralement pas très loin de cette façon vu la difficulté grandissante d'interprétation. De plus, la part de variance expliquée par les facteurs de rang élevé est généralement très faible, l'influence individuelle d'aucun de ces facteurs n'est importante et c'est bien pourquoi leur interprétation est difficile.

A.10. Contributions absolues et relatives

Pour interpréter plus facilement les différents facteurs on considère souvent les *contributions absolue et relative* des différents événements. Elles sont définies comme suit.

Notons f_{h1}, \dots, f_{hp} les composantes du facteur f_h par rapport à la base canonique de \mathbb{R}^p . Puisque f_h est un vecteur normé, on a

$$\sum_{j=1}^p p_{\bullet j} |f_{hj}|^2 = 1$$

(cela signifie encore que, en tant que variable aléatoire, f_h est de variance 1).

Le nombre $p_{\bullet j} |f_{hj}|^2$ est appelé la *contribution absolue* de l'événement B_j ($1 \leq j \leq p$) au facteur f_h ($1 \leq h \leq \rho$).

Ce nombre exprime le pourcentage de la variance de f_h , dû à l'événement B_j . Il permet de mesurer l'importance de cet événement par rapport à ce facteur. D'après les formules établies précédemment, on a :

$$p_{\bullet j} |f_{hj}|^2 = \frac{p_{\bullet j}}{\lambda_h} |(g_h | V_{\star j})|^2 = p_{\bullet j} (D(f_h, V_{\star j}))^2$$

Ainsi, le fait que la contribution absolue de B_j à f_h soit importante peut provenir de deux causes (bien entendu non exclusives l'une de l'autre) : l'événement B_j , peut être de grande probabilité ou très dépendant de f_h et bien corrélé avec g_h .

On définit de façon analogue la contribution absolue de l'événement A_i , ($1 \leq i \leq n$) au facteur g_h ($1 \leq h \leq \rho$) : c'est

$$p_{i\bullet} |g_{hi}|^2 = \frac{p_{i\bullet}}{\lambda_h} |(f_h | V_{i\star})|^2 = p_{i\bullet} (D(V_{i\star}, g_h))^2$$

Pour attribuer une signification à un facteur, on est clairement amené à s'intéresser aux événements (type A ou type B) qui en sont les plus proches (donc les plus dépendants ou les mieux corrélés). On peut dans ce but utiliser les contributions absolues, en éliminant les événements ayant une probabilité très grande. En effet, ceux-ci sont proches du centre de gravité du nuage et ne déterminent donc que faiblement la position des axes d'inertie. Il est à noter que dans le cas particulier des analyses binaires totalement disjonctives, tous les événements de type A (les élèves) ont même probabilité, de sorte que leurs contributions absolues sont proportionnelles aux carrés des dépendances.

Les contributions absolues expriment les parts de chaque événement dans chaque facteur. Les contributions relatives expriment plutôt les parts de chaque facteur dans chaque événement.

Puisque $V_{i\star} = \sum_{h=1}^{\rho} (V_{i\star}|f_h)f_h$ et $\|f_h\| = 1$, on a

$$\|V_{i\star}\|^2 = \sum_{h=1}^{\rho} |(V_{i\star}|f_h)|^2$$

Le nombre $\frac{|(V_{i\star}|f_h)|^2}{\|V_{i\star}\|^2}$ est appelé la *contribution relative* ⁽³⁾ de l'événement A_i ($1 \leq i \leq n$) au facteur f_h , ($1 \leq h \leq \rho$) (on devrait plutôt dire de f_h à A_i). Il exprime le pourcentage de la variance de $V_{i\star}$ due à f_h . C'est aussi le carré du coefficient de corrélation de $V_{i\star}$ et f_h . Notons la formule

$$\frac{|(V_{i\star}|f_h)|^2}{\|V_{i\star}\|^2} = \frac{\lambda_h |g_{hi}|^2}{\|V_{i\star}\|^2}$$

De même, la contribution relative de B_j ($1 \leq j \leq p$) à g_h ($1 \leq h \leq \rho$) est

$$\frac{|(V_{\star j}|g_h)|^2}{\|V_{\star j}\|^2} = \frac{\lambda_h |f_{hj}|^2}{\|V_{\star j}\|^2}$$

⁽³⁾ On parle aussi des « \cos^2 »

A.11. Des variables supplémentaires

Réaliser une analyse des correspondances consiste en définitive à choisir dans les espaces \mathbb{R}^n et \mathbb{R}^p des bases orthonormées qui soient adaptées à l'étude des nuages de points massifs associés à des variables de deux espèces en interaction.

Une fois ces changements de base effectués, rien n'empêche de décomposer un vecteur quelconque de \mathbb{R}^p ou \mathbb{R}^n par rapport à la nouvelle base, et de le positionner sur les différents graphiques réalisés. On dit qu'un tel vecteur est placé en *variable supplémentaire*.

L'intérêt d'introduire des variables supplémentaires réside dans la possibilité de prendre en considération certains comportements sans les faire intervenir dans la détermination des facteurs constituant la nouvelle base orthonormée.

Supposons par exemple qu'un examen soit corrigé en réussite–erreur–abstention : à chaque question nous associons trois variables principales. Mais le comportement d'erreur peut être dû à des erreurs différentes selon les élèves. La variable principale « erreur » ne tient pas compte de ces différences.

On ne peut cependant placer chaque type d'erreur en variable principale sans risquer de fausser l'analyse puisque ce comportement intervient déjà dans la variable erreur globale associée à la question. En plaçant un type d'erreur en variable supplémentaire, on peut le prendre en considération pour l'interprétation des résultats sans fausser l'analyse. Nous avons appliqué ce système au chapitre 21.

Références

[20], [32], [81], [94].