MATHÉMATIQUES ET SCIENCES HUMAINES

BRIGITTE LE ROUX

Analyse spécifique d'un nuage euclidien : application à l'étude des questionnaires

Mathématiques et sciences humaines, tome 146 (1999), p. 65-83

<http://www.numdam.org/item?id=MSH_1999__146__65_0>

Math. Inf. Sci. hum., (37e année, n° 146, *1999*, pp. 65-83)

ANALYSE SPÉCIFIQUE D'UN NUAGE EUCLIDIEN : APPLICATION À L'ÉTUDE DES QUESTIONNAIRES

Brigitte Le ROUX[1]

RÉSUMÉ — *Dans cet article, on propose une méthode d'analyse des correspondances spécifique qui permet de traiter des questionnaires où manquent certaines réponses, et ainsi de s'affranchir du carcan du codage disjonctif complet. La méthode d'analyse spécifique est présentée dans le cadre général de l'analyse géométrique des données pour un nuage euclidien, puis particularisée à un protocole multinumérique et à un questionnaire. En particulier, on montre que l'analyse en composantes principales (*ACP*) bipondérée est privilégiée dans cette approche, et que l'analyse des correspondances multiples (*ACM*) est équivalente à une* ACP *bipondérée sur variables indicatrices. Enfin, on compare analyse spécifique et analyse usuelle, en donnant des inégalités sur les valeurs propres et en étudiant la rotation des sous-espaces principaux lorsque l'on passe de l'analyse globale à l'analyse spécifique.*

MOTS CLÉS — *Analyse géométrique des données, nuage euclidien, analyse en composantes principales bipondérée, analyse des correspondances multiples spécifique, stabilité.*

SUMMARY — *Specific Analysis of a Euclidean Cloud : Application to the study of questionnaires.*

*In this paper, we propose a method of specific Correspondence Analysis which allows to treat questionnaires when some responses are missing, and thus to free oneself from the yoke of complete disjunctive encoding. The method of specific analysis is presented within the general framework of Geometric Data Analysis for a Euclidean cloud, then particularized to multinumerical protocols and to questionnaires. We show that, in this approach, biweighted Principal Component Analysis (*PCA*) is privileged and that Multiple Correspondence Analysis (*MCA*) is equivalent to a biweighted* PCA *on indicator variables. Finally, we compare the specific analysis to the conventional one by writing inequalities between eigenvalues and studying the rotation of principal subspaces when one goes from the global analysis to the specific one.*

KEYWORDS — *Geometric Data Analysis, Euclidean cloud, Biweighted Principal Component Analysis, Specific Multiple Correspondence Analysis, Stability.*

[1]Centre de Recherche en Informatique de Paris 5 (CRIP5), laboratoire SBC, Université René Descartes, 45 Rue des Saints Pères, 75270 PARIS Cedex ; e-mail : lerb@math-info.univ-paris5.fr

Je remercie M. Barbut de ses remarques sur une version antérieure de ce texte.

INTRODUCTION[2]

L'analyse des données provenant de réponses à un questionnaire se fait couramment par l'analyse des correspondances multiples (ACM). Mais il arrive que certaines modalités de réponse aient de faibles fréquences d'observations. On essaie alors, dans la mesure du possible, de les regrouper avec d'autres modalités d'une même question ; mais il n'est pas toujours possible d'effectuer des regroupements qui aient un sens. Il peut arriver aussi que l'un des premiers axes soit un axe spécifique, par exemple un axe des non–réponses, qui ne présente pas toujours d'intérêt, et qui peut sensiblement perturber les autres axes. D'où deux pratiques l'une et l'autre courantes :

1. On écarte de l'analyse les individus ayant choisi ces modalités.
2. On met en éléments supplémentaires les modalités de non–intérêt (modalités rares, non–réponses).

Dans le premier cas, on conserve les structures liées au codage disjonctif complet mais on risque de perdre la représentativité des individus. Dans le second cas, le tableau n'est plus sous forme disjonctive complète, ce qui introduit des propriétés indésirables pour la distance entre individus (cf. remarque §3.2), c'est pourquoi nous récusons cette pratique un peu brutale. Une solution de rechange, appelée AC *avec marge modifiée* a été proposée par B. Escofier, 1987, [6]. La solution que nous préconisons, dans cet article, est techniquement proche mais d'inspiration différente ; elle consiste à garder le nuage des individus comme référence et à procéder à une *analyse spécifique* du nuage dérivé tenant compte uniquement des *modalités d'intérêt.*

Plus généralement, la méthode d'*analyse spécifique d'un nuage euclidien* que nous proposons consiste à *regarder* un nuage (construit une fois pour toutes) selon un certain éclairage, c'est–à–dire à déterminer des axes principaux sous contraintes, par exemple en se restreignant à un sous–espace d'intérêt, comme le support d'un sous–nuage d'individus ou celui d'un nuage dérivé (nuage inter, etc.). Cette méthode appliquée à un protocole multinumérique relève de l'analyse en composantes principales sur variables instrumentales (ACPVI) introduite par Rao, 1964, [12] et développée par Sabatier, 1984, [14].

Au §1, nous rappellerons des propriétés sur les directions et variables principales d'un *nuage euclidien* selon le point de vue adopté par Benzécri, 1973, [1], TIIB n°2, §5 et Rouanet & Le Roux, 1993, [13], chap. V & VI, puis nous les appliquerons à un protocole multinumérique pour lequel le profil d'un individu a le statut d'une *mesure* (par exemple, un tableau de transition), d'où les propriétés et formules de l'analyse en composantes principales bipondérée (ACP bipondérée). Au §2, après avoir rappelé les formules de l'analyse des correspondances multiples (ACM), nous montrerons que l'ACM est une ACP bipondérée particulière. Au §3, nous présenterons l'*analyse*

[2]Ce travail a trouvé son origine dans une collaboration (en cours) avec Jean Chiche (CEVIPOF) sur la construction de l'espace politique français, qui nous a conduit à une communication aux XXXèmes journées de Statistique de Rennes, 1998, [10], ainsi qu'à une contribution à la conférence internationale "Empirical Investigation of Social Space" au Zentralarchiv für Empirische Sozialforschung à Cologne (Octobre 1998).

spécifique d'un nuage euclidien, en particulier celle d'un questionnaire. Au §4, nous comparerons les valeurs propres des deux analyses et étudierons la rotation des sous–espaces principaux lorsque l'on passe de l'analyse globale à l'analyse spécifique.

1. ACP BIPONDÉRÉE

Nous commencerons par rappeler la méthode de détermination conjointe des directions et variables principales d'un nuage euclidien, puis nous présenterons l'ACP bipondérée d'un protocole multinumérique.

1.1. DIRECTIONS ET VARIABLES PRINCIPALES D'UN NUAGE EUCLIDIEN

Nous reprenons les notations de Rouanet & le Roux, 1993, [13]. Soit $\mathcal{U}$ un espace affin euclidien de dimension K, et $\mathcal{V}$ l'espace vectoriel sous–jacent. Soit I un ensemble fini (non vide) et (M^I, n_I) un nuage de points pondérés de $\mathcal{U}$ ($\forall i \in I : n_i > 0$ et $n = \sum_{i\in I} n_i$). L'*ajustement linéaire d'un nuage euclidien* consiste à déterminer le sous–espace affin $\mathcal{L}$ de $\mathcal{U}$ de dimension p par rapport auquel l'inertie (moment d'ordre 2) du nuage M^I ($\sum_{i\in I} n_i d^2(\mathrm{M}^i, \mathcal{L})$) est minimum, c'est–à–dire, si H^i désigne la projection orthogonale du point M^i sur $\mathcal{L}$, tel que $\sum_{i\in I} n_i(\mathrm{M}^i\mathrm{H}^i)^2$ soit minimum. Pour cela, on ajuste successivement p droites principales, qui passent par le point moyen G du nuage, et qui forment une base orthogonale de $\mathcal{L}$.

Notons $\mathbb{R}^I$ l'espace (vectoriel) des variables sur I muni du produit scalaire $\forall x^I \in\mathbb{R}^I, \forall y^I \in\mathbb{R}^I$: $\langle x^I | y^I\rangle = \sum_{i\in I} f_i\, x^i\, y^i$ (avec $f_i = n_i/n$), et $(\delta_i^I)_{i\in I}$ sa base canonique. Les directions principales du nuage M^I s'obtiennent à partir de la décomposition singulière des homomorphismes adjoints *Vac* et *Vac** (*Vac* pour *Va*riable *C*ovariante) définis par (cf. [13], §VI-1-f) :

$$Vac: \begin{array}{ccc} \mathcal{V} & \longrightarrow & \mathbb{R}^I \\ \overrightarrow{\alpha} & \longmapsto & \alpha^I = \left(\langle \overrightarrow{\mathrm{GM}^i} | \overrightarrow{\alpha}\rangle\right)_{i\in I} \end{array} \quad (1) \qquad Vac^*: \begin{array}{ccc} \mathbb{R}^I & \longrightarrow & \mathcal{V} \\ x^I & \longmapsto & \sum_{i\in I} f_i x^i \overrightarrow{\mathrm{GM}^i} \end{array} \quad (2)$$

Les *formules de passage* du vecteur principal normé $\overrightarrow{\alpha}_\ell$ à la variable principale réduite z_ℓ^I et vice–versa (cf. [13], p. 140) sont telles que :

$$\begin{cases} Vac\ (\overrightarrow{\alpha_\ell}) = \xi_\ell z_\ell^i & \text{avec } \|\overrightarrow{\alpha_\ell}\| = 1 \\ Vac^*(z_\ell^I) = \xi_\ell\, \overrightarrow{\alpha_\ell} & \text{et } \sum_{i\in I} f_i (z_\ell^i)^2 = 1 (= \mathrm{Var}\, z_\ell^I) \end{cases} \quad (3)$$

La coordonnée (principale) du point M^i sur l'axe principal $(\mathrm{G}, \overrightarrow{\alpha_\ell})$ est égale à $\xi_\ell\, z_\ell^i$, et est notée y_ℓ^i ; la variance du nuage projeté est égale à $\lambda_\ell = \xi_\ell^2$.

1.2. ACP BIPONDÉRÉE

Soit $(x_k^I)_{k\in\mathcal{K}}$ un protocole multinumérique de K variables sur (I, n_I), et $\varpi_K = (\varpi_k)_{k\in\mathcal{K}}$ une pondération sur $\mathcal{K}$ ($\forall k \in \mathcal{K} : \varpi_k > 0$). Le profil $x_K^i = (x_k^i)_{k\in\mathcal{K}}$ de i, considéré ici comme une *mesure sur* $\mathcal{K}$, est élément de $\mathbb{R}_K$, espace des mesures sur $(\mathcal{K}, \varpi_K)$, muni du produit scalaire $\forall u_K \in\mathbb{R}_K, \forall v_K \in\mathbb{R}_K$: $\langle u_K | v_K\rangle = \sum_{k\in\mathcal{K}} u_k\, v_k/\varpi_k$;

ce profil est représenté par un point pondéré (M^i, n_i) de l'espace affin euclidien $\mathcal{U}$, muni du repère cartésien orthogonal $(0, (\overrightarrow{\delta}^k)_{k\in\mathcal{K}})$, avec $\|\overrightarrow{\delta}^k\| = 1/\sqrt{\varpi_K}$, le point M^i est défini par : $\overrightarrow{OM}^i = \sum_{k\in\mathcal{K}} x_k^i \overrightarrow{\delta}^k$, et la distance entre les points M^i et $M^{i'}$ est égale à :

$$d(i,i') = \left(\sum_{k\in\mathcal{K}} \frac{(x_k^i - x_k^{i'})^2}{\varpi_k}\right)^{1/2} \tag{4}$$

Relativement aux couples de bases $(\delta_i^I)_{i\in I}$ de $\mathbb{R}^I$ et $(\overrightarrow{\delta}^k)_{k\in\mathcal{K}}$ de $\mathcal{V}$, les applications Vac et Vac^* (cf. les définitions (1) et (2)) sont telles que :

$$Vac(\overrightarrow{\delta}^k) = \sum_{i\in I} \frac{(x_k^i - \overline{x}_k)}{\varpi_k} \delta_i^I \qquad Vac^*(\delta_i^I) = \sum_{k\in\mathcal{K}} f_i(x_k^i - \overline{x}_k)\overrightarrow{\delta}^k \tag{5}$$

où $\overline{x}_k = \sum_{i\in I} f_i x_k^i$ désigne la moyenne de la variable x_k^I.

Posons $\overrightarrow{\alpha_\ell} = \sum_{k\in\mathcal{K}} a_{k\ell}\overrightarrow{\delta}^k$, les *formules de passage* (3) s'écrivent alors :

$$\begin{cases} \sum_{k\in\mathcal{K}} (x_k^i - \overline{x}_k)\, a_{k\ell}/\varpi_k = \xi_\ell z_\ell^i = y_\ell^i & \text{avec } \sum_{k\in\mathcal{K}} (a_{k\ell})^2/\varpi_k = 1 \\ \sum_{i\in I} f_i\,(x_k^i - \overline{x}_k)\, z_\ell^i = \xi_\ell\, a_{k\ell} & \text{et } \sum_{i\in I} f_i (z_\ell^i)^2 = 1 \end{cases} \tag{6}$$

Propriétés. Les variables principales sont centrées $(\sum_{i\in I} n_i y_\ell^i = 0)$, leurs variances sont égales aux valeurs propres $(\mathrm{Var}\, y_\ell^I = \lambda_\ell = \xi_\ell^2)$.

Cas particuliers

• Si l'on effectue l'ACP bipondérée par $f_I = (n_i/n)_{i\in I}$ et $f_K = (n_k/n)_{k\in\mathcal{K}}$ du *tableau de transition* f_K^I associé à la mesure effectifs n_{IK} $(f_k^i = n_{ik}/n_i)$, on obtient les formules de transition de l'analyse des correspondances de n_{IK}, en remplaçant dans les formules de passage (6), d'une part x_k^i par f_k^i, d'autre part $\overline{x}_k$ et ϖ_k par f_k, et enfin $a_{k\ell}$ par $f_k z_\ell^k$.

• Si le profil de i est une *variable* sur (K, ϖ_K), notée $x^{iK} = (x^{ik})_{k\in\mathcal{K}}$, on obtient, en remplaçant dans les formules (6), d'une part x_k^i par x^{ik} et $\overline{x}_k$ par $\overline{x}^k = \sum_{i\in I} f_i x^{ik}$, d'autre part $1/\varpi_k$ par ϖ_k et $a_{k\ell}$ par a_ℓ^k, les formules de passage de l'ACP bipondérée d'un *tableau de notes* (cf. [13], §VII-2, p. 168).

2. ACM ET ACP BIPONDÉRÉE

L'analyse des correspondances multiples (ACM) s'applique à un protocole à valeurs dans le produit cartésien de plusieurs variables catégorisées (questionnaire).

On note $\mathcal{Q}$ l'ensemble des questions (Q son cardinal), $\mathcal{K}<q>$ l'ensemble des modalités de la question q (K_q son cardinal), avec $\mathcal{K}<q> \cap \mathcal{K}<q'> = \emptyset$ si $q \neq q'$, et $\mathcal{K} = \bigcup_{q\in\mathcal{Q}} \mathcal{K}<q>$ l'ensemble de toutes les modalités (K son cardinal) ; n_k l'effectif de la modalité k, et $f_k = n_k/n$ sa fréquence.

Soit I un ensemble de n individus ayant passé le questionnaire. Chaque individu donne *une réponse et une seule par question.* Après codage disjonctif, on obtient le protocole des K variables indicatrices $(\delta_k^I)_{k\in\mathcal{K}}$, avec $\delta_k^i = 1$ si i a choisi k et 0 sinon.

Rappelons qu'en ACM, l'*espace affin* $\mathcal{U}$ est muni de la métrique définie par $\varpi_k = f_k/Q$, avec $\|\overrightarrow{\delta}^k\| = 1/\sqrt{f_k/Q}$, que le nuage M^I est équipondéré ($n_i = 1$) et défini par $(\overrightarrow{\mathrm{OM}}^i = \sum_{k\in\mathcal{K}} (\delta_k^i/Q)\,\overrightarrow{\delta}^k)_{i\in I}$, que la distance $d(i,i')$ entre deux points M^i et $\mathrm{M}^{i'}$ est telle que :

$$d(i,i') = \Big(\sum_{k\in\mathcal{K}} \frac{(\delta_k^i - \delta_k^{i'})^2}{Q\,f_k}\Big)^{1/2} \tag{7}$$

et que les *formules de transition* (cf. [13], p. 265) entre les ℓ-èmes variables principales réduites z_ℓ^K sur K et z_ℓ^I sur I s'écrivent :

$$\begin{cases} \sum_{k\in\mathcal{K}<i>} z_\ell^k/Q = \xi_\ell\, z_\ell^i & \text{avec} \quad \sum_{k\in\mathcal{K}} f_k (z_\ell^k)^2/Q = 1 \\ \sum_{i\in I<k>} z_\ell^i/n_k = \xi_\ell\, z_\ell^k & \text{et} \quad \sum_{i\in I} (z_\ell^i)^2/n = 1 \end{cases} \tag{8}$$

où $\mathcal{K}<i>$ désigne l'ensemble des Q modalités choisies par l'individu i, et $I<k>$ l'ensemble des individus ayant choisi la modalité k.

Si l'on effectue l'ACP *bipondérée du protocole des K variables indicatrices* $(\delta_k^I)_{k\in\mathcal{K}}$, muni des pondérations $n_i = 1$ et $\varpi_k = Q\,f_k$: l'espace affin $\mathcal{U}$ est muni de la métrique $\varpi_k = Q f_k$ (avec $\|\overrightarrow{\delta}^k\| = 1/\sqrt{Q f_k}$), le nuage M^I est défini par $(\overrightarrow{\mathrm{OM}}^i = \sum_{k\in\mathcal{K}} \delta_k^i\,\overrightarrow{\delta}^k)_{i\in I}$, et la distance entre deux points M^i et $\mathrm{M}^{i'}$ (cf. formule 4) est égale à la distance de l'ACM (formule 7).

PROPOSITION 1 *L'*ACP *bipondérée du protocole des K variables indicatrices, muni des pondérations $n_i = 1$ et $\varpi_k = Q\,f_k$, est équivalente à l'*ACM *du protocole disjonctif complet, en ce sens qu'elle donne les valeurs propres λ_ℓ et les variables principales réduites (z_ℓ^I sur I et z_ℓ^K sur $\mathcal{K}$) de l'*ACM.

En effet, en remplaçant dans les formules (6) x_k^i par δ_k^i, f_i par $1/n$ et ϖ_k par $Q\,f_k$, on obtient les formules de passage suivantes :

$$\begin{cases} \sum_{k\in\mathcal{K}} (\delta_k^i - f_k)\, a_{k\ell}/(Q\,f_k) = \xi_\ell\, z_\ell^i & \text{avec} \quad \sum_{k\in\mathcal{K}} (a_{k\ell})^2/(Q\,f_k) = 1 \\ \sum_{i\in I} \delta_k^i\, z_\ell^i/n = \xi_\ell a_{k\ell} & \text{et} \quad \sum_{i\in I} (z_\ell^i)^2/n = 1 \end{cases} \tag{9}$$

après simplification de la 2ème équation, puisque $f_k \sum_{i\in I} z_\ell^i/n = 0$ (la variable principale z_ℓ^I est centrée). De cette équation, on déduit que, pour $\xi_\ell \neq 0$, $a_{K\ell} = (a_{k\ell})_{k\in\mathcal{K}}$ est un contraste sur K. En effet : $\sum_{k\in\mathcal{K}} \delta_k^i = Q$ (protocole disjonctif complet), donc $\xi_\ell \sum_{k\in\mathcal{K}} a_{k\ell} = \sum_{i\in I} (\sum_{k\in\mathcal{K}} \delta_k^i) z_\ell^i/n = Q \sum_{i\in I} z_\ell^i/n = 0$.

La première équation (9) se simplifie alors en : $\sum_{k\in\mathcal{K}} \delta_k^i\, a_{k\ell}/(Q f_k) = \xi_\ell z_\ell^i$.

En posant $z_\ell^k = a_{k\ell}/f_k$ dans les équations (9), on retrouve les formules de transition de l'ACM (cf. équations (8)), d'où l'équivalence.

On en déduit les coordonnées $(y_\ell^i = \xi_\ell\, z_\ell^i)_{i\in I}$ des individus, et $(y_\ell^k = \xi_\ell\, z_\ell^k)_{k\in\mathcal{K}}$ des modalités.

3. ANALYSE SPÉCIFIQUE

Nous poserons d'abord le problème de l'analyse spécifique pour un nuage euclidien, puis nous présenterons l'ACM spécifique.

3.1. ANALYSE SPÉCIFIQUE D'UN NUAGE EUCLIDIEN

L'analyse spécifique d'un nuage consiste à déterminer ses directions principales sous la contrainte de leur appartenance à un certain sous-espace.

On reprend la démarche générale rappelée au §1.1., appliquée à un sous-espace $\mathcal{L}$ de $\mathcal{A} \subset \mathcal{U}$ (on suppose que $\mathcal{A}$ passe par G). Notons A^i la projection orthogonale du point M^i sur $\mathcal{A}$ et B^i sa projection sur $\mathcal{A}^\perp$, supplémentaire orthogonal de $\mathcal{A}$ passant par G. On a :

$$\overrightarrow{GM}^i = \overrightarrow{GA}^i + \overrightarrow{GB}^i \quad \text{avec } \overrightarrow{GA}^i \perp \overrightarrow{GB}^i$$

Le point H^i, projection orthogonale de M^i sur $\mathcal{L}$, est aussi projection orthogonale de A^i sur $\mathcal{L}$, d'où : $(M^iH^i)^2 = (M^iA^i)^2 + (A^iH^i)^2$, et

$$\sum_{i\in I} n_i d^2(M^i, \mathcal{L}) = \sum_{i\in I} n_i d^2(M^i, \mathcal{A}) + \sum_{i\in I} n_i d^2(A^i, \mathcal{L})$$

En conséquence, le sous-espace $\mathcal{L}$, de dimension p, de $\mathcal{A}$ par rapport auquel l'inertie du nuage M^i est minimum est le premier sous-espace principal de dimension p associé au nuage A^I : les coordonnées spécifiques de M^i sont égales aux coordonnées principales de A^i ; la variance du nuage spécifique A^I est toujours inférieure ou égale à celle du nuage global M^I.

Remarque

L'application de l'analyse spécifique d'un nuage euclidien à une ACP bipondérée est directe. En termes statistiques, cela revient à déterminer des variables principales du nuage M^I qui sont des combinaisons linéaires des variables d'intérêt.

3.2. ACM SPÉCIFIQUE D'UN SOUS-ENSEMBLE DE MODALITÉS

Nous traiterons, dans la suite, le cas de l'analyse spécifique d'un questionnaire *par restriction* de l'ensemble des modalités à un sous-ensemble $\mathcal{K}'$ de modalités d'intérêt (de cardinal noté K'). On notera $\mathcal{K}_s$ l'ensemble des autres modalités et K_s son cardinal (avec donc $\mathcal{K} = \mathcal{K}' \cup \mathcal{K}_s$). On définira naturellement la distance entre M^i et $M^{i'}$ en restreignant, dans la formule (7), la sommation à $\mathcal{K}' \subset \mathcal{K}$:

$$d'(i,i') = \left(\sum_{k\in\mathcal{K}'} \frac{(\delta_k^i - \delta_k^{i'})^2}{Q f_k} \right)^{1/2} \tag{10}$$

L'équivalence entre ACM et ACP bipondérée nous amène à considérer l'espace euclidien $\mathcal{U}$ et le nuage M^I définis à la fin du §2., le sous-espace $\mathcal{A}$ engendré par $(\overrightarrow{\delta}^k)_{k\in\mathcal{K}'}$, et le nuage A^I projeté sur $\mathcal{A}$, avec $A^iA^{i'} = d'(i,i')$ (formule 10).

L'ACM spécifique du nuage M^I consiste à déterminer les directions principales du nuage A^I, c'est-à-dire à effectuer l'ACP bipondérée du protocole des K' variables indicatrices $(\delta_k^I)_{k\in\mathcal{K}'}$, muni des pondérations $n_i = 1$ et $\varpi_k = Q f_k$.

Après avoir remplacé dans les *formules de passage* (9), ξ_ℓ par $\sqrt{\mu_\ell}$, $a_{k\ell}/f_k$ par $t_\ell^k/\sqrt{\mu_\ell}$ et z_ℓ^i par $t_\ell^i/\sqrt{\mu_\ell}$ et en restreignant, dans la première équation, la sommation à K', on obtient :

$$\begin{cases} \sum_{k\in K'<i>} t_\ell^k/Q - \sum_{k\in K'} f_k\, t_\ell^k/Q = \sqrt{\mu_\ell}\, t_\ell^i & \text{avec } \sum_{k\in K'} \frac{f_k}{Q}(t_\ell^k)^2 = \mu_\ell \\ \sum_{i\in I<k>} t_\ell^i/n_k = \sqrt{\mu_\ell}\, t_\ell^k \quad (k \in K') & \text{et } \sum_{i\in I}(t_\ell^i)^2/n = \mu_\ell \end{cases} \tag{11}$$

Dans l'analyse spécifique d'un questionnaire, on prendra t_ℓ^i comme coordonnée de l'individu i, et t_ℓ^k comme coordonnée de la modalité k.

Propriétés

- Si $\overline{t_\ell^k}$ désigne la moyenne des ℓ–èmes coordonnées principales des points M^i correspondants aux individus ayant choisi la modalité k ($\overline{t_\ell^k} = \sum_{i\in I<k>} t_\ell^i/n_k$), on a, d'après la deuxième formule de (11) : $\overline{t_\ell^k} = \sqrt{\mu_\ell}\, t_\ell^k$.

  Cette formule de passage permet aussi de déterminer les coordonnées principales des *modalités supplémentaires.*

- Les modalités $k \in K<q>$ d'une question q déterminent une partition du nuage des individus, d'où $\sum_{k\in K<q>} n_k \overline{t_\ell^k} = 0$, et $\sum_{k\in K<q>} n_k t_\ell^k = 0$: les variables principales ne sont pas centrées sur $K'<q>$, elles les sont sur $K<q>$.

  Pour chaque question, les moyennes des coordonnées principales des modalités actives (modalités d'intérêt) et supplémentaires d'une même question sont nulles.

- $\sum_{k\in K'} f_k\,(t_\ell^k)^2/Q = \mu_\ell$

  La somme des carrés des coordonnées principales des modalités actives (pondérées par f_k/Q) est égale à la valeur propre.

- $\operatorname{Var} t_\ell^K = \sum_{k\in K'} f_k\,(t_\ell^k)^2/Q + \sum_{k\in K_s} f_k (t_\ell^k)^2/Q \geq \mu_\ell$.

3.2.1. Algorithme de calcul

1) On détermine la matrice symétrique $\mathbf{T}$, de terme général :

$$t_{kk'} = \frac{1}{Q} \times \frac{f_{kk'}}{\sqrt{f_k\, f_{k'}}} - \frac{1}{Q} \times \sqrt{f_k\, f_{k'}} \qquad (k \in K', k' \in K')$$

2) On diagonalise $\mathbf{T}$; d'où les valeurs propres μ_ℓ et les vecteurs propres normés $c_{K\ell}$ (avec $\sum_{k\in K'} c_{k\ell}^2 = 1$).

3) On calcule les coordonnées des modalités (variables principales sur K') :

$$t_\ell^k = \sqrt{\mu_\ell}\, c_{k\ell}/\sqrt{f_k/Q}$$

et les coordonnées des individus (variables principales sur I) :

$$y_\ell^i = \sum_{k \in K' <i>} c_{k\ell}/\sqrt{Q\, f_k} - \sum_{k \in K'} c_{k\ell}/\sqrt{f_k/Q}$$

On a $\sum_{k' \in K'} t_{kk'} \sqrt{f'_k} \neq 0$: $\mathbf{T}$ n'a pas de valeur propre triviale nulle.

3.2.2. Commentaires méthodologiques

1) Le choix de l'analyse spécifique plutôt que celui de l'AC du sous–protocole $(\delta_k^I)_{k \in K'}$ est motivé par la remarque suivante.

Si l'on effectue l'AC du sous–protocole, pour tout couple d'individus n'ayant choisi que des modalités de d'intérêt ($k \in K'$), la distance diminue tout en restant proportionnelle à la distance d de l'ACM ($= d\sqrt{N'/(nQ)}$, avec $N' = \sum_{k \in K'} n_k < nQ$) ; mais, pour un couple d'individus qui ne sont en désaccord que pour une question, l'un ayant choisi une modalité d'intérêt et l'autre pas, en raison de la normalisation, les modalités d'intérêt des autres questions et pour lesquelles ils sont d'accord interviennent dans la distance (propriété indésirable).

Par contre, si l'on procède à l'analyse spécifique, pour tout couple d'individus n'ayant choisi que des modalités d'intérêt, la *distance est inchangée* ($d' = d$), et pour deux individus qui ne sont en désaccord que pour une question, l'un ayant choisi une modalité d'intérêt et l'autre une modalité k de non–intérêt, le carré de la distance diminue de $1/(Q\, f_k)$, et *les modalités d'intérêt des autres questions pour lesquelles ils sont d'accord n'interviennent pas dans le calcul de la distance.*

2) Les résultats de l'ACM spécifique par restriction de modalités sont proches de ceux de l'AC *avec marge modifiée* proposée par B. Escofier, 1987, [6] mais l'approche est différente.

Dans l'AC avec marge modifiée, le paradigme est celui de l'AC : le profil d'un individu est $(\delta_k^i/Q)_{k \in K'}$, pondéré par $1/n > (/K' <i> /)/(nQ)$, et les profils des modalités sont ceux de l'ACM, à savoir $(\delta_k^i/n_k)_{i \in I}$, affectés des poids n_k/N'. On vérifie facilement que les coordonnées des modalités de l'AC avec marge modifiée sont les mêmes que celles de l'ACM spécifique, que les valeurs propres sont celles de l'ACM spécifique multipliées par nQ/N' (elles peuvent être supérieures à celles du nuage de référence M^I et même devenir supérieures à 1), et que les coordonnées des individus sont égales à celles de l'ACM spécifique multipliées par $\sqrt{nQ/N'}$.

En ACM spécifique, le nuage des individus, avec ses directions et variables principales, est gardé comme référence. En particulier, les valeurs propres de l'ACM spécifique sont inférieures à celles de l'ACM (donc inférieures à 1), ce qui assure la comparabilité des analyses ; les taux de variance peuvent augmenter (cf. infra §4).

4. COMPARAISON DES ANALYSES GLOBALE ET SPÉCIFIQUE

Dans ce paragraphe, on compare les valeurs propres et les sous–espaces principaux de l'analyse globale et de l'analyse spécifique. On donnera d'abord les résultats pour l'ACP bipondérée, qui généralisent au cas d'une pondération non élémentaire des variables ceux de Escofier & Le Roux, 1977, [9], puis pour l'ACM.

Notons M, A et B les endomorphismes symétriques $Vac \circ Vac^*$ de $\mathbb{R}^I$ associés respectivement aux nuages M^I (global), A^I (spécifique) et B^I (résiduel). Leurs valeurs propres seront respectivement notées λ_ℓ, μ_ℓ et β_ℓ (rangées par ordre décroissant, avec leur ordre de multiplicité) ; leurs vecteurs propres normés sont les variables principales réduites de chacun des nuages. En reprenant les définitions de Vac et Vac^* (cf. formules 5), on vérifie que $M = A + B$. On est donc ramené à l'étude des perturbations des valeurs propres et de la rotation des sous-espaces invariants d'un endomorphisme symétrique M par la suppression d'un endomorphisme symétrique B positif (cf. Annexe). Dans toute la suite, les endomorphismes M, A et B auront pour source et pour but le sous-espace de $\mathbb{R}^I$ image par Vac du support du nuage M^I, dont on notera L la dimension.

Rappelons tout d'abord que la position relative de deux sous-espaces dont l'un est de dimension r et l'autre de dimension supérieure ou égale à r est définie par r angles $\pi/2 \geq \theta_1 \geq \theta_2 \geq \ldots \geq \theta_r \geq 0$, appelés *angles canoniques* (cf. Dixmier, 1948, [5] ; Benzécri & al. 1973, [1], p.179). Dans toute la suite, on étudiera le *plus grand angle canonique*, noté θ, c'est-à-dire le plus grand angle entre tout couple de vecteurs des deux sous-espaces qui sont projections orthogonales l'un de l'autre.

On présentera d'abord le cas d'un nuage résiduel B^I de dimension quelconque, puis on affinera les résultats pour un nuage de dimension 1.

4.1. NUAGE RÉSIDUEL DE DIMENSION QUELCONQUE

En appliquant les théorèmes (1) et (2) (cf. Annexe), on compare les valeurs propres et les sous-espaces principaux des nuages M^I et A^I.

4.1.1. Cas du protocole multinumérique x^I_K

Si l'on écarte K_s variables de l'analyse, le nuage B^I a au plus K_s valeurs propres non nulles, le théorème 1 s'applique en posant $\beta_k = 0$ pour $k > K_s$. En particulier, les inégalités de Weyl s'écrivent :

$$\lambda_\ell - \beta_1 \leq \mu_\ell \leq \lambda_\ell$$

Pour $\ell \leq K_s + 1$, on a : $\mu_\ell \geq \max\{\lambda_{K_s - \ell + 2} \,;\, \lambda_\ell - \beta_1\}$.

D'après le théorème 2, en posant $\delta = (\lambda_\ell - \lambda_{\ell+1})$ pour l'étude du sous-espace principal associé aux ℓ premières valeurs propres, et $\delta = \inf\{(\lambda_{\ell-1} - \lambda_\ell), (\lambda_{\ell+r} - \lambda_{\ell+r+1})\}$ pour celle du sous-espace principal associé aux valeurs propres de rang $\ell, .. \ell + r$, on a :

$$\text{si} \quad \beta_1 < \delta \quad \text{alors :} \quad \theta < \pi/4 \quad \text{et} \quad \sin 2\theta \leq \frac{\beta_1}{\delta}$$

Si l'on écarte *deux variables* x^I_k et $x^I_{k'}$, on démontre facilement que l'on a :

$$\beta_1 = \tfrac{1}{2}\left((\mathrm{Cta}_k + \mathrm{Cta}_{k'}) + \sqrt{(\mathrm{Cta}_k - \mathrm{Cta}_{k'})^2 + 4\,\mathrm{Cov}^2(x^I_k | x^I_{k'})/(\varpi_k\,\varpi_{k'})}\right)$$

où $\mathrm{Cta}_k = \mathrm{Var}\, x^I_k / \varpi_k$ désigne la contribution absolue de x^I_k à la variance du nuage.

Si $K_s > 2$, et si l'on ne connait pas la plus grande valeur propre du nuage B^I, on pourra toujours la majorer par $\sum_{k \in K_s} \mathrm{Cta}_k$, variance du nuage résiduel.

Remarques

1) Les valeurs propres spécifiques μ_ℓ sont inférieures ou égales à celles du nuage global. De $\text{Var}\,\text{M}^I = \text{Var}\,\text{A}^I + \text{Var}\,\text{B}^I$ et des inégalités (3) du théorème 1, il résulte que l'on ne peut pas conclure sur le sens de variation des taux de variance.

2) La rotation des sous–espaces principaux est d'autant plus faible que l'écart entre les valeurs propres bordant les sous–espaces étudiés est plus grand, et que la plus grande valeur propre du nuage résiduel est petite.

4.1.2. Cas d'un questionnaire

Si l'on écarte de l'analyse K_s modalités, on applique les résultats précédents avec :

$$\text{Cta}_k = \frac{1 - f_k}{Q} \quad \text{et} \quad \frac{\text{Cov}^2(x_k^I | x_{k'}^I)}{\varpi_k\, \varpi_{k'}} = \frac{1}{Q^2} \times \frac{(f_{kk'} - f_k\, f_{k'})^2}{f_k\, f_{k'}}$$

Dans le cas de K_q *modalités d'une même question* q, le nuage résiduel a $K_q - 1$ valeurs propres égales à $1/Q$, les autres étant nulles. On remplacera donc, dans les formules β_1 par $1/Q$, d'où :

$$\lambda_\ell - 1/Q \leq \mu_\ell \leq \lambda_\ell$$

et pour $\ell \leq K_q$: $\mu_\ell \geq \max\{\lambda_{K_q - \ell + 1}\,,\ \lambda_\ell - 1/Q\}$.

L'angle θ entre les sous–espaces engendrés par les ℓ premières variables principales est tel que :

$$\text{si} \quad \lambda_\ell - \lambda_{\ell+1} > \frac{1}{Q} \quad \text{alors} \quad \theta < \pi/4 \quad \text{et} \quad \sin 2\theta \leq \frac{1}{Q} \times \frac{1}{\lambda_\ell - \lambda_{\ell+1}}$$

Remarques

1) Les valeurs propres de l'ACM spécifique, lors de la suppression d'une question, sont inférieures ou égales à celles de l'ACM, et sont d'autant plus proches que le nombre de questions est plus grand.

2) Plus le nombre de questions est grand et plus l'écart entre les valeurs propres est important, plus faible est la rotation du sous–espace principal associé aux ℓ premières valeurs propres.

3) Si, au lieu de faire une analyse spécifique, on effectue l'ACM des $Q-1$ questions, alors un facteur correctif $Q/(Q-1)$ apparaît dans les inégalités sur les valeurs propres, et on ne peut pas conclure sur le sens de variation des valeurs propres (cf. Escofier & Le Roux, 1975, [8], p. 13).

4.2. NUAGE RÉSIDUEL DE DIMENSION UN

Si le nuage résiduel est de dimension 1, par exemple si l'on écarte de l'analyse une seule variable x_k^I, ou dans le cas d'un questionnaire, une modalité k ou une question à 2 modalités, on pourra tenir compte de la position de cette variable ou des modalités par rapport aux axes principaux.

4.2.1. Cas d'une variable

Notons ψ_ℓ l'angle aigu tel que $\cos\psi_\ell$ est égal à la valeur absolue de la corrélation entre x_k^I et la ℓ–ème variable principale y_ℓ^I du nuage M^I, et Ψ_ℓ l'angle tel que $\cos^2\Psi_\ell =$

$\sum_{j=1}^{\ell} \cos^2 \psi_j$ (carré de la corrélation multiple entre x_k^I et les ℓ premières variables principales du nuage M^I).

Notons $\mathrm{Cta}_{k\ell}$ ($= \mathrm{Cta}_k \cos^2 \psi_\ell$) la contribution absolue des variables x_k^I et de y_ℓ^I à la variance globale, $\mathrm{Ctr}_{k\ell}$ ($= \mathrm{Cta}_{k\ell} / \mathrm{Var}\,\mathrm{M}^I$) leur contribution relative, et Ctr_k ($= \sum_{\ell=1}^{L} \mathrm{Ctr}_{k\ell}$) la contribution relative de x_k^I à la variance globale.

— *Valeurs propres et taux de variance.*

En appliquant le théorème 3, on obtient pour les valeurs propres :

$$\max\{\lambda_{\ell+1} ; \lambda_\ell - \sum_{j=1}^{\ell} \mathrm{Cta}_{kj} \leq \mu_\ell \leq \lambda_\ell \quad (1 \leq \ell < L)$$

et pour les taux de variance $\tau_\ell(M) = \sum_{j=1}^{\ell} \lambda_j / \mathrm{Var}\,\mathrm{M}^I$ du nuage global M^I et $\tau_\ell(A) = \sum_{j=1}^{\ell} \mu_j / \mathrm{Var}\,\mathrm{A}^I$ du nuage spécifique A^I :

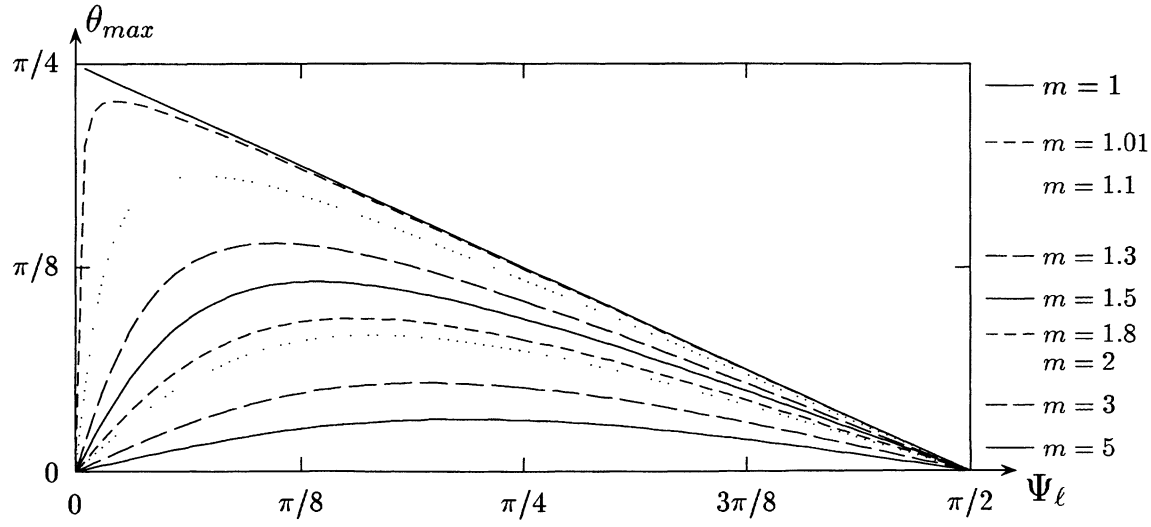
$$\frac{\tau_\ell(M) - \sum_{j=1}^{\ell} \mathrm{Ctr}_{kj}}{1 - \mathrm{Ctr}_k} \leq \tau_\ell(A) \leq \frac{1}{1 - \mathrm{Ctr}_k} \tau_\ell(M)$$

— *Sous-espace des ℓ premières variables principales.*

En appliquant le théorème 4, on a, en posant $m = (\lambda_\ell - \lambda_{\ell+1}) / \mathrm{Cta}_k$:

$$\bullet \text{ si } m > 1 \quad \text{alors : } \theta < \pi/4 \quad \text{et} \quad \tan 2\theta \leq \frac{\sin 2\Psi_\ell}{m - \cos 2\Psi_\ell} \tag{12}$$

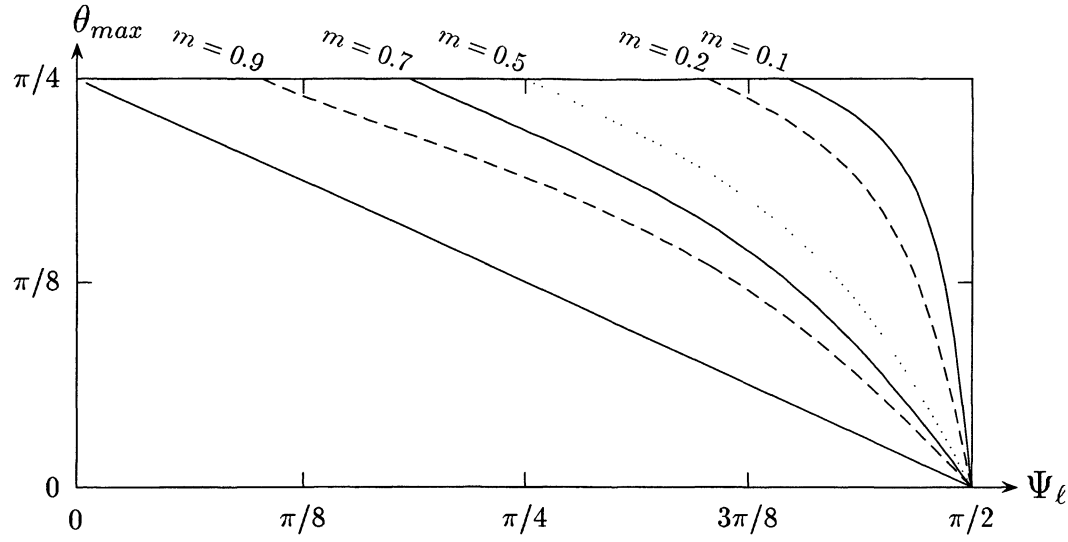
Le faisceau de courbes ci-après donne la borne θ_{max} de θ en fonction de Ψ_ℓ pour différentes valeurs de $m > 1$.



$$\bullet \text{ si } \cos^2 \Psi_\ell \leq m < 1 \quad \text{alors : } \theta < \pi/4 \quad \text{et} \quad \tan 2\theta \leq \frac{\sin 2\Psi_\ell}{m - \cos^2 \Psi_\ell} \tag{13}$$

La condition $\cos^2 \Psi_\ell \leq m < 1$ s'écrit aussi $\sum_{j=1}^{\ell} \text{Cta}_{kj} \leq \lambda_\ell - \lambda_{\ell-1} < \text{Cta}_k$.

Le faisceau de courbes ci-après donne la borne θ_{max} de θ en fonction de Ψ_ℓ pour différentes valeurs de $m < 1$.



— *Comparaison des variables principales de rang ℓ.*

D'après le théorème 5, l'angle θ (à l'orientation près) entre variables principales de rang ℓ est tel que :

si $\delta = \inf\{(\lambda_{\ell-1} - \lambda_\ell) - \sum_{j=1}^{\ell-1} \text{Cta}_{kj} \; ; \; (\lambda_\ell - \lambda_{\ell+1}) - \sum_{j=1}^{\ell+1} \text{Cta}_{kj}\} > 0$, alors :

$$\sin\theta \leq \frac{\text{Cta}_k}{2} \times \frac{\sin 2\psi_\ell}{\delta} \tag{14}$$

Remarques

1) Les valeurs propres μ_ℓ et λ_ℓ sont égales si la corrélation multiple entre la variable écartée x_k^I et les variables principales de rang inférieur ou égal à ℓ est nulle.

Le taux de variance des ℓ premiers axes de l'analyse spécifique est supérieur à celui de l'analyse globale si le carré de la corrélation multiple de x_k^I avec les ℓ premières variables principales de l'analyse globale est inférieur au taux de variance $\tau_\ell(M)$ de l'analyse globale.

2) Si la corrélation multiple de la variable écartée avec le sous-espace principal étudié est proche de 0 (Ψ_ℓ proche de $\pi/2$), la rotation de ce sous-espace restera petite ; si, au contraire, elle est proche de ± 1 (Ψ_ℓ proche de 0), la rotation pourra très vite devenir importante (cf. les faisceaux de courbes).

4.2.2. Cas du questionnaire

Dans le cas où l'on écarte *une modalité d'une question*, les formules précédentes s'appliquent avec $\text{Cta}_k = (1 - f_k)/Q$, $\text{Cta}_{k\ell} = (f_k/Q)(y_\ell^k)^2$ et $\cos^2 \psi_\ell = \frac{f_k}{1-f_k}(y_\ell^k)^2$.

Dans le cas où l'on écarte *une question à 2 modalités* k et k', le nuage associé a une valeur propre égale à $1/Q$; les modalités k et k' sont alignées avec le centre du nuage,

et font un angle ψ_ℓ avec l'axe ℓ, tel que $\cos^2 \psi_\ell = f_k(y_\ell^k)^2/(1-f_k) = (1-f_k)(y_\ell^{k'})^2/f_k$. D'où pour les valeurs propres :

$$\gamma_\ell - \cos^2 \Psi_\ell/Q \leq \mu_\ell \leq \gamma_\ell$$

et pour les sous-espaces principaux $1, \ldots \ell$:

$$\text{si} \quad \gamma_\ell - \gamma_{\ell+1} > 1/Q \quad \text{alors} \quad \theta < \pi/4 \quad \text{et} \quad \tan 2\theta \leq \frac{\sin 2\Psi_\ell}{Q(\lambda_\ell - \lambda_{\ell+1}) - \cos 2\Psi_\ell}$$

Remarque

Pour une question à 2 modalités, seule intervient la qualité de représentation $\cos^2 \Psi_\ell$ de la question par le sous-espace étudié, alors que si l'on écarte une modalité, l'éloignement de cette modalité au centre du nuage et son poids interviennent.

4.3. GUIDE D'UTILISATION

Nous appliquerons la démarche à un nuage dont l'un des premiers axes, disons le deuxième, est un axe de modalités particulières — par exemple un axe des non-réponses dans un questionnaire — et où l'on procède à l'analyse spécifique après avoir écarté ces modalités particulières. On supposera que le nuage des modalités écartées (nuage résiduel) est unidimensionnel, on notera $\cos \psi_\ell$ la valeur absolue de la corrélation entre la variable principale du nuage résiduel et la ℓ-ème variable principale du nuage global, et Cta_s la plus grande valeur propre du nuage résiduel, qui est égale à la somme des contributions absolues des modalités écartées.

Supposons les 4 premiers axes de l'analyse globale interprétables, nous nous proposons d'étudier les trois premières variables principales de l'analyse spécifique, en les reliant aux quatre premières variables principales de l'analyse globale.

4.3.1. Etude de la première variable principale

On commence par s'assurer que l'angle aigu, que l'on notera θ_1, entre les variables principales de rang 1 des deux analyses est inférieur à $\pi/4$. Pour cela, il faut (cf. théorème 4) que la condition $\lambda_1 - \lambda_2 > \text{Cta}_s \cos^2 \psi_1$ soit vérifiée, c'est-à-dire que la somme des contributions absolues au premier axe des modalités écartées soit inférieure à l'écart entre les deux premières valeurs propres.

D'après le théorème 4, si $\lambda_1 - \lambda_2 > \text{Cta}_s$, on majore θ_1 à l'aide de l'inégalité :

$$\tan 2\theta_1 \leq \frac{\sin 2\psi_1}{((\lambda_1 - \lambda_2)/\text{Cta}_s) - \cos 2\psi_1}$$

à défaut, on majore θ_1 à l'aide de l'inégalité $\tan 2\theta_1 \leq \dfrac{\sin 2\psi_1}{((\lambda_1 - \lambda_2)/\text{Cta}_s) - \cos^2 \psi_1}$.

Supposons cet angle θ_1 très petit (les théorèmes s'appliquent stricto sensu s'il est nul), alors on passe à l'étude de la 2ème variable principale.

4.3.2. Etude de la deuxième variable principale

Sous l'hypothèse $\theta_1 = 0$, on étudie la rotation de la 2ème variable principale comme suit[3]. Si $\lambda_2 - \mathrm{Cta}'_s \cos^2 \psi_2 < \lambda_4$ (avec $\mathrm{Cta}'_s = \mathrm{Cta}_s \sin^2 \psi_1$) — c'est-à-dire si la deuxième valeur propre diminuée de la somme des contributions absolues à l'axe 2 des modalités écartées devient inférieure à la 4ème valeur propre — alors, dans l'analyse spécifique, l'axe 2 est renvoyé à un rang supérieur à 4. On est donc amené à comparer les 3ème et 4ème axes du nuage global aux 2ème et 3ème axes du nuage spécifique.

4.3.3. Etude des troisième et quatrième variables principales

Pour étudier la rotation de la 3ème variable principale du nuage global, on applique le théorème 6 (on notera $\theta_{2|3}$ l'angle aigu entre la 3ème variable principale du nuage global et la 2ème du nuage spécifique) :

$$\text{Si } \lambda_3 - \lambda_4 < \mathrm{Cta}'_s \sin 2\psi_2 \text{ alors } \theta < \pi/4 \text{ et } \sin 2\theta_{2|3} \leq \frac{\mathrm{Cta}'_s \sin 2\psi_2}{\lambda_3 - \lambda_4}.$$

Si cette majoration est petite, $\theta_{2|3}$ est petit, la corrélation entre la 2ème variable principale du nuage spécifique et la 3ème du nuage global est presque égale à ± 1. Sinon, et si $\lambda_4 - \lambda_5 \gg \lambda_3 - \lambda_4$, on peut chercher à établir la stabilité du plan 3-4.

Si la quantité $\mathrm{Cta}'_s \sin 2\psi_2/(\lambda_4 - \lambda_5)$ est petite, les plans sont confondus (ou presque), on va alors pouvoir déterminer la position de la 2ème variable principale spécifique dans ce plan[4]. On a :

$$\tan 2\theta_{2|3} = \frac{2 \cos \psi_3 \cos \psi_4}{((\lambda_3 - \lambda_4)/\mathrm{Cta}_s) - (\cos^2 \psi_3 - \cos^2 \psi_4)}$$

CONCLUSIONS

1) L'ACP bipondérée apparaît ici comme la méthode privilégiée en *Analyse Géométrique des Données*, dès que l'on construit un nuage euclidien d'individus (avec sa distance et sa pondération) — que l'on prend comme *nuage de référence* — et que l'on procède à plusieurs analyses spécifiques.

2) Le problème du codage — comme celui des non-réponses — est apparu très tôt comme crucial en ACM, d'où les études de son influence sur les résultats de l'analyse : cf. Escofier & Le Roux, 1975, [8] ; Cazes & Lecoutre, 1977, [3]. C'est un souci constant chez les chercheurs en Sciences Humaines d'éliminer les "modalités non pertinentes" et d'essayer d'augmenter l'importance des premiers axes, cf. Schiltz, 1983, [15]. En particulier, le codage disjonctif amène à construire des catégories comme "Autres", qui sont mal définies et qui regroupent souvent des données de nature très différente. Ces *catégories "fourre-tout"*, qui sont la plupart du temps sans intérêt, peuvent *perturber l'analyse*. L'ACM spécifique d'un nuage euclidien

[3]Les théorèmes sont appliqués à la restriction des endomorphismes M, A et B au sous-espace principal associé aux valeurs propres de rang supérieur à 1 ; en particulier, pour le nuage résiduel, la valeur propre à considérer est $\mathrm{Cta}'_s = \mathrm{Cta}_s \sin^2 \psi_1$.

[4]On étudie les restrictions des endomorphismes M, A et B à ce plan (endomorphismes de rang 2), et on applique la formule 16 de l'annexe.

permet de se libérer du carcan du codage disjonctif complet, tout en conservant les propriétés essentielles de l'ACM. On trouvera un exemple d'ACM spécifique (avec la classification spécifique associée) dans l'analyse des données "Editeurs" présentée par Bourdieu, 1999, [2].

Dans les questionnaires d'opinion, l'un des premiers axes est souvent un *axe des non–réponses*, qui peut ne pas présenter d'intérêt en soi, mais qui risque de perturber les axes suivants et d'en rendre l'interprétation difficile. On procèdera alors à l'analyse spécifique en écartant les non–réponses ; l'étude des non–réponses, ou plutôt des non–répondants à telle ou telle question, se fera alors à partir du nuage des individus comme nous l'avons fait dans Le Roux & Chiche, 1998, [10].

3) L'analyse spécifique s'applique aussi lorsque l'on souhaite étudier une ou plusieurs catégories d'individus. Dans ce cas, le point important est celui de la *comparabilité des résultats* : dans l'analyse spécifique, le nuage des individus, construit une fois pour toutes, est le *nuage de référence*.

4) La comparaison de l'analyse spécifique à l'analyse globale présentée au §4. montre que les résultats des deux analyses diffèrent peu si les modalités écartées sont proches du centre de gravité, ou que leurs contributions aux axes interprétables sont faibles.

5) Enfin, l'analyse du nuage inter, ou celle du nuage intra (cf. Lebart & al., 1995, [11], p.335-336, apparaissent comme des analyses spécifiques, l'analyse spécifique par restriction de modalités en étant la forme la plus simple.

Mise en œuvre informatique

Deux programmes, l'un concernant l'ACP bipondérée, l'autre l'ACM spécifique sont disponibles (en version DOS) sur le serveur ftp de l'université René Descartes :

`ftp.math-info.univ-paris5/pub/MathPsy/AGD`

Le premier programme appelé ACPPON (auteurs B. Le Roux & P. Bonnet) reprend et complète le programme ANCOMP de l'ADDAD ; le deuxième, appelé ACMSPE, est une adaptation faite par B. Le Roux & J. Chiche du programme ACMULT de l'ADDAD[5].

[5]Je remercie vivement P-O. Flavigny (INRETS) qui, par ses conseils et sa participation active, nous a permis de mettre au point ces programmes dans l'environnement ADDAD.

BIBLIOGRAPHIE

[1] BENZÉCRI J-P. & COLL, *L'analyse des données*, tome 2, Paris, Dunod, 1973.

[2] BOURDIEU P., "Une révolution conservatrice", *Les Actes de la Recherche en Sciences Sociales*, n° 127, 1999.

[3] CAZES P. & LECOUTRE J.P., "Etude de quelques problèmes de codage en analyse des correspondances", *Cahier du bureau universitaire de recherche opérationnelle*, cahier n° 27, 1979, 59-66.

[4] CHANDLER-DAVIS & KAHAN W.M., "The rotation of eigenvectors by a perturbation" SIAM *J. Numer. Anal.*, Vol n°1, March 1970, 1-46.

[5] DIXMIER J., "Position relative de deux variétés linéaires fermées dans un espace de Hilbert", *Rev. Sci.*, 86, 1948, 387-399.

[6] ESCOFIER B., "Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte", *Pub. Inst. Stat. Univ.*, XXXII, fasc. 3, 1987, 33-69.

[7] ESCOFIER B. & LE ROUX B., "Rotation du sous–espace invariant d'un endomorphisme symétrique de $\mathbb{R}^n$ par une perturbation symétrique", *R.A.I.R.O.*, 9^e année, 1975, 5-8.

[8] ESCOFIER B. & LE ROUX B., "Etude des questionnaires par l'analyse des correspondances : modification du codage des questions ou de leur nombre et stabilité de l'analyse", *Math. Sci. Hum.*, n° 49, 1975, 5-27.

[9] ESCOFIER B. & LE ROUX B., "Mesure de l'influence d'un descripteur sur les résultats d'une analyse en composantes principales", *Pub. Inst. Stat. Univ. Paris*, XXII, 1977, 25-44.

[10] LE ROUX B. & CHICHE J., "Analyse spécifique d'un questionnaire : cas particulier des non–réponses", XXXèmes journées de Statistique de la S.F.d.S., Rennes, Mai 1998.

[11] LEBART L., MORINEAU A. & PIRON M., *Statistique exploratoire multidimensionnelle*, Dunod, Paris, 1995.

[12] RAO C.R., "The use and interpretation of Principal Component Analysis in applied research". Sankhya, A, 26, 1964, p. 329-359.

[13] ROUANET H. & LE ROUX B., *Analyse des données multidimensionnelles*, Dunod, Paris, 1993.

[14] SABATIER R., "Quelques généralisations de l'analyse en composantes principales de variables instrumentales", *Stat. Ann. Données*, 9, 3, 1984, 75-103.

[15] SCHILTZ M-A., "L'élimination des modalités non pertinentes dans un dépouillement d'enquête par analyse factorielle", *BMS*, n° 1, 1983, 19-40.

[16] WILKINSON J.H., *The algebraic eigenvalue problem*, Clarendon Press, Oxford, 1965.

Le texte a été composé avec le logiciel LaTeX, les courbes ont été faites avec GNUPLOT.

Dans cette annexe, nous énonçons les théorèmes permettant de comparer les valeurs propres et les sous–espaces invariants de deux endomorphismes symétriques A et C d'un espace euclidien $\mathcal{V}$ de dimension n lors de l'ajout d'une perturbation B. Soit $C = A + B$.

On trouvera des démonstrations du théorème 1, sur la comparaison des valeurs propres, dans Wilkinson, 1965, [16], p. 100-101. Le théorème 2, traitant de la rotation des sous–espaces invariants, est une application d'un théorème dû à Davis & Kahan, 1970, [4] ; on en trouvera une démonstration simplifiée dans Escofier & Le Roux, 1975, [7]. Pour les théorèmes 3, 4 et 5, portant sur une perturbation de rang un, on pourra se reporter à Escofier & Le Roux, 1977, [9] ; on donnera une démonstration du théorème 6, celle–ci ne se trouvant dans aucun des articles cités.

On notera respectivement α_ℓ, β_ℓ et γ_ℓ les valeurs propres (rangées par ordre décroissant, avec leur ordre de multiplicité) des endomorphismes A, B et C.

THÉORÈME 1 . *(1) Pour tous entiers j, k, ℓ, compris entre 1 et n, vérifiant $j + k \leq \ell + 1$, on a :* $\gamma_\ell \leq \alpha_j + \beta_k$ *et* $\gamma_{n-\ell+1} \geq \alpha_{n-j+1} + \beta_{n-k+1}$

(2) Pour $1 \leq \ell \leq n$, on a : $\alpha_\ell + \beta_n \leq \gamma_\ell \leq \alpha_\ell + \beta_1$

(3) Pour $1 \leq \ell \leq n$, on a : $\displaystyle\sum_{j=1}^{\ell} \alpha_j + \sum_{j=n-\ell+1}^{n} \beta_j \leq \sum_{j=1}^{\ell} \gamma_j \leq \sum_{j=1}^{\ell} \alpha_\ell + \sum_{j=1}^{\ell} \beta_j$

Les inégalités (2) sont appelées inégalités de Weyl.

THÉORÈME 2 . *Le plus grand angle canonique θ entre les sous–espaces invariants de C et de A associés aux ℓ valeurs propres de rang $r, \ldots r + \ell - 1$ est tel que, en posant :*

$$\begin{cases} \delta = \inf\{(\alpha_{r-1} - \alpha_r), (\alpha_{r+\ell-1} - \alpha_{r+\ell})\} & \text{avec} \quad 1 < r \leq n-1 \quad \text{et} \quad 1 \leq \ell \leq n-r \\ \delta = (\alpha_\ell - \alpha_{\ell+1}) & \text{pour} \quad r = 1 \quad \text{et} \quad 1 \leq \ell < n \end{cases}$$

$$\textit{si } \beta_1 - \beta_n < \delta \quad \textit{alors :} \quad \theta < \pi/4 \quad \textit{et} \quad \sin 2\theta \leq \frac{\beta_1 - \beta_n}{\delta}$$

Remarques

1) Cette majoration peut aussi s'exprimer en fonction des valeurs propres de l'endomorphisme C : il suffit de remplacer α par γ dans l'expression de δ.

2) La borne est optimale. Elle est atteinte si le plan invariant de B associé aux valeurs propres β_1 et β_n est confondu avec le plan invariant de A associé à $\alpha_{\ell-1}$ et α_ℓ ou à $\alpha_{\ell+r-1}$ et $\alpha_{\ell+r}$, et si les vecteurs propres ont, dans ce plan, la configuration maximisant θ (cf. infra, remarque (3) du théorème 5).

• Dans le cas où B *est de rang 1*, on peut affiner les majorations précédentes en tenant compte de la position de la droite propre de B (associée à sa valeur propre non nulle) par rapport au sous–espace étudié. Notons β la valeur propre non nulle de B (on supposera $\beta > 0$), φ_ℓ (resp. ψ_ℓ) l'angle entre le sous–espace propre de B associé à β et le sous–espace propre de A (resp. C) associé à la valeur propre α_ℓ (resp. γ_ℓ), et Φ_ℓ (resp. Ψ_ℓ) l'angle avec le sous–espace invariant de A (resp. C) associé aux ℓ premières valeurs propres : $\cos^2 \Phi_\ell = \displaystyle\sum_{j=1}^{\ell} \cos^2 \varphi_j$ (resp. $\cos^2 \Psi_\ell = \displaystyle\sum_{j=1}^{\ell} \cos^2 \psi_j$).

THÉORÈME 3 .

(1) Pour $1 < \ell \leq n$, *on a :* $\alpha_\ell \leq \gamma_\ell \leq \min\{\alpha_\ell + \beta \sin^2 \Phi_{\ell-1} ; \alpha_{\ell-1}\}$

Pour $1 \leq \ell \leq n$, *on a :* $\sum_{j=1}^{\ell} \alpha_j + \beta \cos^2 \Phi_\ell \leq \sum_{j=1}^{\ell} \gamma_j \leq \sum_{j=1}^{\ell} \alpha_j + \beta$

(2) Pour $1 \leq \ell < n$, *on a :* $\max\{\gamma_{\ell+1} ; \gamma_\ell - \beta \cos^2 \Psi_\ell\} \leq \alpha_\ell \leq \gamma_\ell$

pour $1 \leq \ell \leq n$, *on a :* $\sum_{j=1}^{\ell} \gamma_j - \beta \cos^2 \Psi_\ell \leq \sum_{j=1}^{\ell} \alpha_j \leq \sum_{j=1}^{\ell} \gamma_j$

THÉORÈME 4 . *L'angle* θ *entre les sous-espaces invariants de* A *et de* C *associés aux* ℓ *premières valeurs propres est tel que, en posant* $m = (\alpha_\ell - \alpha_{\ell+1})/\beta$ *et* $m' = (\gamma_\ell - \gamma_{\ell+1})/\beta$ *:*

(1) si $m > 1$ *(ou pour* $\ell = 1$, $m + \cos 2\varphi_1 > 0$*), alors :*

$$\theta < \pi/4 \quad et \quad \tan 2\theta \leq \frac{\sin 2\Phi_\ell}{m + \cos 2\Phi_\ell}$$

si $\sin^2 \Phi_\ell < m < 1$, *alors :* $\theta < \pi/4$ et $\tan 2\theta \leq \dfrac{\sin 2\Phi_\ell}{m - \sin^2 \Phi_\ell}$

(2) si $m' > 1$, *alors :* $\theta < \pi/4$ *et* $\tan 2\theta \leq \dfrac{\sin 2\Psi_\ell}{m' - \cos 2\Psi_\ell}$

si $\cos^2 \Psi_\ell < m' < 1$, *alors :* $\theta < \pi/4$ et $\tan 2\theta \leq \dfrac{\sin 2\Psi_\ell}{m' - \cos^2 \Psi_\ell}$

Remarque

Les majorations, pour $m < 1$ et $m' < 1$ sont optimales ; elles sont atteintes si le vecteur propre de B associé à β est situé dans le plan des vecteurs propres de rang ℓ et $\ell + 1$.

THÉORÈME 5 . *L'angle* θ *entre les sous-espaces invariants de* A *et de* C *associés aux valeurs propres de rang* $\ell, \ldots, \ell + r$ *est tel que :*

si $\delta = \inf\{(\alpha_{\ell-1} - \alpha_\ell) - \beta \sin^2 \Phi_{\ell-1} ; (\alpha_{\ell+r} - \alpha_{\ell+r+1}) - \beta \sin^2 \Phi_{\ell+r}\} > 0$,

alors : $\sin \theta \leq \dfrac{\beta}{2} \dfrac{\sin 2\Phi}{\delta}$, *avec* Φ *tel que* $\cos^2 \Phi = \sum_{j=\ell}^{\ell+r} \cos^2 \phi_j = \cos^2 \Phi_{\ell+r} - \cos^2 \Phi_{\ell-1}$

ou si $\delta' = \inf\{(\gamma_{\ell-1} - \gamma_\ell) - \beta \cos^2 \Psi_{\ell-1} ; (\gamma_{\ell+r} - \gamma_{\ell+r+1}) - \beta \cos^2 \Psi_{\ell+r}\} > 0$,

alors : $\sin \theta \leq \dfrac{\beta}{2} \dfrac{\sin 2\Psi}{\delta'}$ *avec* Ψ *tel que* $\cos^2 \Psi = \sum_{j=\ell}^{\ell+r} \cos^2 \psi_j = \cos^2 \Psi_{\ell+r} - \cos^2 \Psi_{\ell-1}$

Remarques

1) Le théorème 5 s'applique aussi aux sous-espaces invariants associés aux ℓ plus grandes valeurs propres, mais la majoration est plus grande.

2) Si les sous-espaces associés aux $\ell - 1$ premières valeurs propres de A et de C sont confondus (par exemple, si $\cos^2 \Phi_{\ell-1} = 0$), pour comparer, par exemple les sous-espaces propres de rang ℓ on applique les majorations de $\tan 2\theta$, en remplaçant β par $\beta \sin^2 \Phi_{\ell-1}$ (resp. $\beta \sin^2 \Psi_{\ell-1}$) et Φ_ℓ (resp. Ψ_ℓ) par ϕ_ℓ (resp. ψ_ℓ) — ce qui

revient à étudier les restrictions des 3 endomorphismes au supplémentaire orthogonal du sous–espace invariant associé aux $\ell - 1$ premières valeurs propres.

3) Si les plans engendrés par les vecteurs propres de rang ℓ et $\ell+1$ sont confondus, l'angle θ entre les sous–espaces propres de rang ℓ est tel que :

$$\tan 2\theta = \frac{2\cos\phi_\ell \cos\phi_{\ell+1}}{m + (\cos^2\phi_\ell - \cos^2\phi_{\ell+1})} \quad \text{avec} \quad m = \frac{\alpha_\ell - \alpha_{\ell+1}}{\beta} \tag{15}$$

ou

$$\tan 2\theta = \frac{2\cos\psi_\ell \cos\psi_{\ell+1}}{m' - (\cos^2\psi_\ell - \cos^2\psi_{\ell+1})} \quad \text{avec} \quad m' = \frac{\lambda_\ell - \lambda_{\ell+1}}{\beta} \tag{16}$$

THÉORÈME 6 . *Notons P_ℓ la projection orthogonale sur le sous–espace propre de C associé à γ_ℓ, et posons $C' = C - (\beta\cos 2\psi_\ell)P_\ell$. L'angle θ entre les sous–espaces invariants de C' et de A associés aux valeurs propres de rang $s, s+1, \ldots s+r$ est tel que, en posant $\delta = \inf\{(\gamma'_{s-1} - \gamma'_s), (\gamma'_{s+r} - \gamma'_{s+r+1})\}$:*

$$\textit{si } \delta > \beta\sin 2\psi_\ell \quad \textit{alors} \quad \theta < \pi/4 \quad \textit{et} \quad \sin 2\theta \leq \frac{\beta\,\sin 2\psi_\ell}{\delta}$$

Démonstration

Posons $C' = C - u\,\beta P_\ell$ avec $\cos 2\psi_\ell < u < \cos^2\psi_\ell$, et P_b la projection orthogonale sur le sous- -espace propre de B associé à la valeur propre β ($B = \beta P_b$). On a : $A = C - B = C' - \beta(P_b - u\,P_\ell)$. L'endomorphisme $B' = \beta(P_b - u\,P_\ell)$, est de rang 2, son image $B'(\mathcal{V})$ est engendrée par le vecteur propre b de B associé à β, et par le vecteur propre c_ℓ de C asssocié à λ_ℓ. Soit $c' \in B'(\mathcal{V})$ un vecteur normé orthogonal à c_ℓ, on a :

$$\begin{cases} B'(c_\ell) = \beta(\cos\psi_\ell\, b - u\, c_\ell) & = \beta(\cos^2\psi_\ell - u)c_\ell + \beta\sin\psi_\ell\cos\psi_\ell c' \\ B'(c') = \beta\sin\psi_\ell\, b & = \beta\sin\psi_\ell\cos\psi_\ell\, c_\ell + \beta\sin^2\psi_\ell\, c' \end{cases}$$

Les valeurs propres non nulles de B' valent $\frac{\beta}{2}((1-u) + \sqrt{1+u^2-2u\cos 2\psi_\ell}) \geq 0$ et $\frac{\beta}{2}((1-u) - \sqrt{1+u^2-2u\cos 2\psi_\ell}) \leq 0$; leur différence est égale $\beta\sqrt{1+u^2-2u\cos 2\psi_\ell}$. D'après le théorème 2, l'angle θ entre les sous– espaces invariants de C' et de A est tel que : $\sin 2\theta \leq \beta(\sqrt{1+u^2-2u\cos 2\psi_\ell})/\delta$. Or $\sqrt{1+u^2-2u\cos 2\psi_\ell}$ est minimum pour $u = \cos 2\varphi_\ell$ et le minimum vaut $\sin 2\psi_\ell$, d'où le théorème.

Remarques

1) Les vecteurs propres de $C' = C - (\beta\cos 2\psi_\ell)P_\ell$ sont vecteurs propres de C associés aux mêmes valeurs propres, sauf celui associé à la valeur propre $\gamma_\ell - \beta\cos 2\psi_\ell$, qui est décalé au rang $t > \ell$; comparer un sous–espace invariant de A à un sous–espace invariant de C' associé aux valeurs propres de mêmes rangs revient à le comparer au sous–espace invariant de C associé aux valeurs propres de mêmes rangs après *décalage* de γ_ℓ du rang ℓ au rang $t > \ell$, tel que :

$$\lambda_t < \lambda_\ell - \mathrm{Cta}_k\cos 2\psi_\ell < \lambda_{t-1}$$

A l'aide de ce théorème, on comparera les sous–espaces propres de même rang j des deux analyses si $1 \leq j < \ell - 1$ ou $j \geq t+1$ (en posant $\lambda'_j = \lambda_j$), et ceux décalés de un si $\ell \leq j \leq t$, avec $\lambda'_j = \lambda_{j+1}$ et $\lambda'_t = \lambda_\ell - \beta\cos 2\psi_\ell$.

2) On a un théorème analogue pour comparer les sous–espaces invariants de C et $A' = A - \beta\cos 2\phi_\ell$, en échangeant les rôles de A et de C.