Développements récents en analyse des correspondances multiples

Jean Chiche[1], Brigitte Le Roux[2]

[1] CEVIPOF/CNRS, Sciences–Po Paris
jean.chiche@sciences-po.fr
[2] MAP5, Université Paris Descartes et CEVIPOF, Sciences–Po Paris
Brigitte.LeRoux@mi.parisdescartes.fr

Résumé Dans cet article nous proposons deux variantes de l'Analyse des Correspondances Multiples. La première méthode, appelée ACMspé, donne un traitement particulier aux modalités de non–intérêt, sans pour autant perdre les propriétés constitutives de l'ACM. La seconde méthode permet d'étudier une classe d'individus, en tant que sous–ensemble d'un ensemble d'individus de référence.

Keywords : Analyse des correspondances multiples spécifique, analyse spécifique de classe, analyse géométrique des données, ellipses de concentration, enquêtes par questionnaire.

1 Introduction

Nous présenterons deux variantes de l'analyse des correspondances multiples, qui ont été développées pour résoudre des problèmes soulevés lors de nos travaux en collaboration avec Pierre Bourdieu et Henry Rouanet (voir [2] et [3]).

Dans une première partie, nous développerons l'ACMspé qui permet de traiter les modalités de non–intérêt sans pour autant détruire les propriétés structurelles de l'ACM[3]. Dans une deuxième partie, nous étudierons une classe d'individus en référence à l'espace global des individus. Nous présenterons ensuite les résumés géométriques de sous–nuages sous forme d'ellipses. Enfin, nous présenterons un exemple d'application.

2 Analyse des correspondances multiples spécifique (ACMspé)

Dans le traitement des questionnaires par l'ACM, l'un des problèmes est celui des modalités peu choisies (par exemple de fréquences inférieures à 5%). L'ACMspé apporte une solution à ce problème.

Il est bien connu que, en ACM, les modalités rares sont représentées par des points qui sont éloignés du centre du nuage, qu'elles contribuent trop fortement à la variance de leur question et qu'elles peuvent être trop influentes dans la détermination des axes. Souvent, il n'est pas possible de regrouper ces modalités avec d'autres de la même question. On peut aussi avoir le cas de modalités de non–intérêt ou même celui de modalités (par exemple "Autres") qui sont un mélange de modalités hétérogènes et qui ne sont

[3]Pour un exposé de l'ACM, voir [1] ou [7], pour un exposé élémentaire, voir [9].

pas représentables par un point unique. Afin de conserver les propriétés constitutives de l'ACM, nous avons élaboré une variante de l'ACM, qui consiste, non pas à supprimer ces modalités, mais simplement à les ignorer lors du calcul des distances entre individus. Nous appelons cette nouvelle méthode, l'ACM spécifique. En ACMspé, de telles modalités sont appelées "modalités passives", en opposition aux modalités actives des questions actives.

L'ACM, en tant que méthode géométrique, consiste à construire un nuage euclidien représentant les individus et à en déterminer les axes principaux. L'étape cruciale de l'ACM est donc le calcul des distances entre les individus à partir de leurs réponses.

Notations

On note Q l'ensemble des questions actives ainsi que son cardinal, I l'ensemble des individus et n son cardinal. On note I_k le sous–ensemble des individus ayant choisi la modalité k et n_k son cardinal. On note K l'ensemble des modalités.

La distance entre deux individus est calculée à partir de leurs réponses aux questions actives. Si, pour la question $q \in Q$ il y a "accord" entre les individus i et i', la distance due à cette question est nulle. Si pour la question q il y a"désaccord", l'un ayant choisi la modalité k et l'autre la modalité k' ($k' \neq k$), la distance $d_q(i,i')$ entre les individus i et i' est telle que :

$$d_q^2(i,i') = \frac{1}{f_k} + \frac{1}{f_{k'}}$$

(avec $f_k = n_k/n$ et $f_{k'} = n_{k'}/n$). La distance globale entre deux individus est la moyenne quadratique des distances dues aux questions et est définie par :

$$d^2(i,i') = \frac{1}{Q} \sum_{q \in Q} d_q^2(i,i')$$

En ACMspé, si pour une question q de désaccord, l'un des individus a choisi une modalité passive, disons k', alors la distance spécifique due à cette question est telle que :

$$d_q'^2(i,i') = \frac{1}{f_k}$$

On note $d'(i,i')$ la *distance spécifique* entre i et i'.

Remarques

1. Pour deux individus i et i' qui n'ont pas choisi de modalités passives, la distance est inchangée : $d'(i,i') = d(i,i')$.

   Pour 2 individus en désaccord pour une seule question, l'un ayant choisi une modalité active, disons k, et l'autre une modalité passive disons k', alors on a : $d'^2(i,i') = d^2(i,i') - 1/(Qf_{k'})$.

2. Si l'on effectue l'analyse des correspondances du tableau disjonctif dont on a supprimé les colonnes correspondant aux modalités passives, la distance entre deux individus qui n'ont choisi que des modalités actives diminue ; cette distance est égale à $d(i,i')\sqrt{\sum_{k \in K'} n_k/(nQ)}$, K' désignant l'ensemble des modalités actives des questions actives. Mais pour deux individus en désaccord pour une seule question, l'un ayant choisi une modalité active et l'autre une modalité passive, en raison de la normalisation, les questions actives créent de la distance, ce qui est clairement une *propriété indésirable*.

2.1 Nuage des individus, nuage des modalités

En ACM, l'ensemble des individus, muni de la distance d, est représenté par un nuage euclidien de n points dans un espace à K dimensions muni du repère affin $(\mathrm{O}, (\mathrm{I}^k)_{k\in K})$, tel que $\overrightarrow{\mathrm{OI}}^k \perp \overrightarrow{\mathrm{OI}}^{k'}$ $(k \neq k')$ et $(\mathrm{OI}^k)^2 = Q/f_k$. Ce nuage est noté $(\mathrm{M}^i)_{i\in I}$.

Le point M^i représentant l'individu i est le barycentre des points I^k appartenant à K_i (ensemble des modalités actives et passives choisies par l'individu i) :

$$\mathrm{M}^i = \sum_{k\in K_i} \mathrm{I}^k/Q$$

Le nuage spécifique des individus, noté $(\mathrm{M}'^i)_{i\in I}$, est la projection orthogonale du nuage $(\mathrm{M}^i)_{i\in I}$ sur le sous–espace engendré par les points idéaux associés aux modalités actives $(\mathrm{I}^k)_{k\in K'}$. Le point M'^i est donc tel que : $\overrightarrow{\mathrm{OM}}'^i = \sum_{k\in K'_i} \overrightarrow{\mathrm{OI}}^k/Q$, où K'_i désigne l'ensemble des modalités actives choisies par i.

Le nuage spécifique des modalités est le sous-nuage des modalités actives des questions actives avec poids et distances inchangés.

2.2 Propriétés

- La dimension du sous–espace contenant le nuage spécifique est au plus égal à $K'-Q'$, Q' désignant le nombre de questions actives sans modalités passives.
- La variance du nuage spécifique est égale à

  $$\frac{K'}{Q} - \sum_{k\in K'} \frac{f_k}{Q}$$

  Elle est inférieure (ou égale) à la variance du nuage initial $(\mathrm{M}^i)_{i\in I}$, qui vaut $(K - Q)/Q)$[4].
- *Formules de transition.* Si $y_\ell{}^i$ désigne la coordonnée du point–individu sur l'axe principal ℓ et y_ℓ^k celle du point–modalité k (avec poids relatif $p_k = f_k/Q$), on a les deux formules de transition suivantes :

  $$y'^i_\ell = \frac{1}{\sqrt{\mu_\ell}}\Big(\sum_{k\in K'_i} (y'^k_\ell)/Q - \sum_{k\in K'} p_k y'^k_\ell \Big) \quad \text{et} \quad y'^k_\ell = \frac{1}{\sqrt{\mu_\ell}} \sum_{i\in I_k} y'^i_\ell/n_k$$

  K'_i désigne le sous–ensemble des modalités actives choisies par i.
  Ces formules s'appliquent aux individus et modalités actives, passives et supplémentaires.
- La moyenne des coordonnées spécifiques des individus sur l'axe ℓ est nulle, la variance est égale à la valeur propre.

  $$\sum_{i\in I} y^i_\ell/n = \lambda_\ell \qquad \text{et} \qquad \sum_{i\in I} (y^i_\ell)^2/n = \lambda_\ell$$

- Pour toute question q, la moyenne des coordonnées sur un axe principal des modalités actives et passives est nulle.

  $$\forall q \quad \sum_{k\in K_q} p_k y^k_\ell = 0$$

[4]Dans [5], on trouve une méthode très proche de l'ACMspé, mais que ne vérifie pas cette propriété.

où K_q désigne l'ensemble des modalités de la question q.
La somme des carrés, pondérée par $p_k = f_k/Q$, des coordonnées sur un axe des modalités actives est égale à la valeur propre.

$$\sum_{k\in K'} p_k (y_\ell^k)^2 = \lambda_\ell$$

2.3 Méthode de calcul

1. Diagonalisation de la matrice symétrique $\mathbf{T} = [t_{kk'}]$, avec

$$t_{kk'} = \frac{1}{Q}\Big(\frac{n_{kk'}}{\sqrt{n_k\, n_{k'}}} - \frac{\sqrt{n_k\, n_{k'}}}{n}\Big) \qquad \left\{ \begin{array}{l} k \in K' \\ k' \in K' \end{array} \right.$$

d'où les valeurs propres λ_ℓ et les vecteurs propres normés $(c_{k\ell})_{k\in K'}$ ($\sum\limits_{k\in K'} c_{k\ell}^2 = 1$).

2. Coordonnées des modalités :

$$y'^k_\ell = \sqrt{\lambda_\ell}\sqrt{Q}\frac{c_{k\ell}}{\sqrt{n_k/n}} \qquad (k \in K')$$

3. Coordonnées des individus :

$$y'^i_\ell = \frac{1}{\sqrt{Q}}\Big(\sum_{k\in K_i} \frac{c_{k\ell}}{\sqrt{n_k/n}} - \sum_{k\in K'} \frac{c_{k\ell}}{\sqrt{n_k/n}}\Big)$$

Remarque. $(n_k)_{k\in K'}$ n'est pas vecteur propre de $\mathbf{T}$.

3 Analyse spécifique de classe (CSA)

Cette variante de l'ACM est utilisée pour étudier une classe d'individus en référence à l'ensemble total des individus. CSA[5] a pour but de déterminer les caractéristiques spécifiques d'une classe à considérer. Cette méthode consiste à rechercher les axes principaux du *sous–nuage* associé aux individus de cette classe.

Notations. Dans cette section, on note N le cardinal de I (nombre total d'individus) ; N_k le nombre d'individus de I ayant choisi k et $F_k = N_k/N$ la fréquence correspondante.

On note I' le sous–ensemble des individus de la classe et n son cardinal. On note n_k le nombre d'individus de la classe ayant choisi k et $f_k = n_k/n$ la fréquence associée ; on note $n_{kk'}$ le nombre d'individus de la classe ayant choisi k et k'.

3.1 Nuage spécifique des individus de la classe

La distance entre 2 individus i and i' de la classe est celle définie à partir du nuage global, plus précisément si pour la question q, i choisit k and i' choisit k', on a (utilisant les notations de cette section) $d^2(i,i') = (1/F_k) + (1/F_{k'})$.

Remarque. A partir de l'ACM du sous–tableau $I' \times Q$, les distances auraient été égales à $(1/f_k) + (1/f_{k'})$. Le sous–nuage associé à la classe est d'autant plus différent de celui construit par l'ACM du sous–tableau que les fréquences f_k diffèrent beaucoup des fréquences globales F_k.

[5] CSA est l'abbréviation de l'appellation anglaise, à savoir Class Specific Analysis.

4 Nuage spécifique des modalités

La distance entre deux modalités est :

$$d'^2(k,k') = \frac{f_k(1-f_k)}{F_k^2} + \frac{f_{k'}(1-f_{k'})}{F_{k'}^2} - 2\frac{f_{kk'} - f_k\, f_{k'}}{F_k\, F_{k'}}$$

La distance du point–modalité M'^k au centre du nuage est égal à $f_k(1-f_k)/F_k^2$.

Le nuage des modalités est pondéré, la modalité k a pour poids $p_k = F_k/Q$ (comme dans le nuage global).

4.1 Propriétés

– Les nuages spécifiques des individus et des modalités ont même variance, notée V_{spe}, avec

$$V_{\text{spe}} = \frac{1}{Q}\sum_{k\in K}\frac{f_k(1-f_k)}{F_k}$$

La contribution du point M^k à la variance spécifique est $\mathrm{Ctr}_k = \Big(\frac{1}{Q}\frac{f_k(1-f_k)}{F_k}\Big)/V_{\text{spe}}$.

– *Formules de transition.* Si on note y'^i_ℓ la coordonnée du point M^i sur l'axe ℓ, y'^k_ℓ celle du point M'^k, et μ_ℓ la ℓ–ième valeur propre des nuages spécifiques, on a les formules de transition suivantes :

$$y'^i_\ell = \frac{1}{\sqrt{\mu_\ell}}\Big(\sum_{k\in K_i} y'^k_\ell/Q - \sum_{k\in K} p_k y'^k_\ell\Big) \quad \text{et} \quad y'^k_\ell = \frac{1}{\sqrt{\mu_\ell}}\sum_{i\in I'_k} y'^i_\ell/(nF_k)$$

où I'_k est le sous–ensemble des individus de I' ayant choisi k et avec $p_k = F_k/Q$..

– La moyenne des coordonnées spécifiques des individus I' sur l'axe ℓ est nulle ; la variance est égale à μ_ℓ.

$$\sum_{i\in I} y'^i_\ell/n = 0 \qquad \text{et} \sum_{i\in I}(y'^i_\ell)^2/n = \mu_\ell$$

– La moyenne des coordonnées spécifiques des modalités (pondérées par $p_k = F_k/Q$) sur l'axe ℓ est nulle ; la variance est égale à μ_ℓ.

$$\sum_{k\in K} p_k y'^k_\ell = 0 \qquad \text{et} \sum_{k\in K} p_k (y'^k_\ell)^2 = \mu_\ell$$

4.2 Méthode de calcul

1. Diagonalisation de la $K \times K$ matrice symétrique $\mathbf{T} = [t_{kk'}]$, avec

$$t_{kk'} = \frac{1}{Q}\times\frac{N}{n}\times\frac{n_{kk'} - n_k\, n_{k'}/n}{\sqrt{N_k\, N_{k'}}} \qquad \left\{\begin{array}{l} k\in K \\ k'\in K\end{array}\right.$$

d'où les valeurs propres μ_ℓ et les vecturs propres normés $(c_{k\ell})_{k\in K}$ ($\sum_{k\in K} c^2_{k\ell} = 1$).

2. Coordonnées des modalités sur l'axe ℓ :

$$y'^k_\ell = \sqrt{\mu_\ell}\sqrt{Q}\frac{c_{k\ell}}{\sqrt{N_k/N}} \qquad (k\in K)$$

3. Coordonnées des individus sur l'axe ℓ :

$$y'^{i}_{\ell} = \frac{1}{\sqrt{Q}}\Big(\sum_{k\in K_i} \frac{c_{k\ell}}{\sqrt{N_k/N}} - \sum_{k\in K} \frac{(n_k/n)c_{k\ell}}{\sqrt{N_k/N}}\Big) \qquad (i \in I')$$

Remarque. $(\sqrt{N_k})_{k\in K}$ est vecteur propre de $\mathbf{T}$ associé à la valeur propre 0.

5 Ellipses de concentration

On résumera des sous–nuages projetés dans un plan principal non seulement par leurs points moyens mais aussi géométriquement par des ellipses de concentration (voir [4] p. 284).

5.1 Ellipses d'inertie

Si, relativement à la base orthonormée $\ell 1$ et $\ell 2$, le sous–nuage dont le point moyen a pour coordonnées m_1 et m_2, pour variances v_1 et v_2, et pour covariance c, la κ–ellipse d'inertie du sous–nuage est centrée sur le point moyen et est définie par :

$$\frac{v_2(y_1 - m_1)^2 - 2c(y_1 - m_1)(y_2 - m_2) + v_1(y_2 - m_2)^2}{v_1 v_2 - c^2} = \kappa^2$$

L'ellipse indicatrice est définie par $\kappa = 1$, l'ellipse de concentration est définie par $\kappa = 2$.

5.2 Ellipses de concentration d'un sous–nuage

Si on note γ_1^2 et γ_2^2 les valeurs propres associées aux axes principaux du sous–nuage plan, l'ellipse de concentration du sous–nuage a pour demi–axes $2\gamma_1$ et $2\gamma_2$ et l'angle α_1 que fait le premier axe principal du sous–nuage avec l'axe $\ell 1$ est défini par $\tan\alpha_1 = (v_1 - \gamma_1^2)/c$.

Propriétés.

1. L'ellipse de concentration d'un sous–nuage est l'ellipse d'inertie telle qu'une distribution uniforme à l'intérieur de l'ellipse a une variance égale à celle du sous–nuage.
2. Pour un sous–nuage ayant un distribution normale bi–dimensionnelle, l'ellipse de concentration contient 86.5% de la distribution.
3. Si un sous–nuage est projeté orthogonalement sur un sous–espace, le κ–hyperellipsoide du nuage projeté est la projection du κ–hyperellipsoide du nuage sur ce sous–espace.

En AGD, les ellipses de concentration sont des *résumés géométriques* des sous–nuages associés à un facteur structurant comme on le verra dans l'application suivante. En inférence statistique, sous un modèle approprié, les ellipses d'inertie sont également des ellipses de confiance pour le"point moyen vrai" de la classe.

6 Application : questionnaire sur la mondialisation

Le questionnaire fut envoyé en 2006 aux 577 députés français, les questions portaient sur la mondialisation ; 163 députés ont répondu au questionnaire. On a effectué une ACMspé sur l'ensemble des répondants en retenant 12 questions actives avec 41 modalités parmi lesquelles 11 (autres ou non–réponses d'effectifs compris entre 2 et 13) ont été considérées comme passives.

On a retenu pour l'interprétation le premier plan principal qui rend compte de 36% de la variance ($\lambda_1 = 0.408, \lambda_2 = 0.135$). Le premier axe oppose une attitude favorable à une attitude défavorable à la mondialisation. Le deuxième axe oppose la protection de l'agriculture à la défense de l'emploi et de l'environnement.

On a aussi étudié l'appartenance à un groupe politique comme facteur structurant, les 3 principaux groupes étant : UMP ($n_1 = 92$), PS ($n_2 = 48$) et UDF ($n_3 = 15$). L'analyse géométrique révèle des différences entre les groupes, comme le montrent les points moyens et les ellipses de concentration des groupes (voir Figure 1). D'où les *conclusions descriptives* : vis–à–vis de la mondialisation, pour l'ensemble des répondants, les députés PS se situent du côté gauche de la figure (défavorable), et ceux de l'UMP du côté droit (favorable). Intuitivement, l'écart entre le points moyens du PS et de l'UMP est important[6].

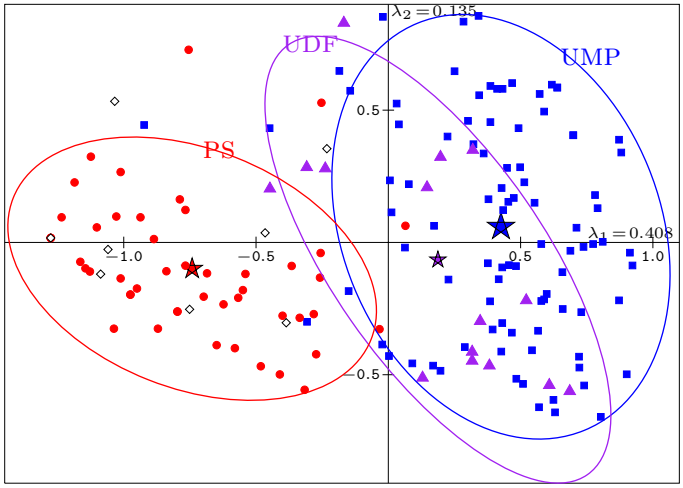

FIG. 1 – Nuage des 163 députés dans le plan 1-2, marqués selon leur appartenance à un groupe politique, 92 UMP (carré), 48 PS (cercle), 15 UDF (triangle) avec leurs points moyens et leurs ellipses, et 8 députés appartenant à d'autres groupes (losange).

7 Conclusion

L'ACMspé est maintenant utilisée dans de nombreuses études (voir par exemple [2], [3], [6] et [8]), car les modalités de non–intérêt sont plutôt la règle que l'exception dans l'analyse des questionnaires.

La méthode CSA a été développée très récemment et répond à une demande des utilisateurs. Les premières applications semblent prometteuses.

[6]Le questionnaire "mondialisation" et ses principaux résultats ont été commenté dans 'Le Figaro' du 29 Mai 2006 et sur le site Web de Télos (www.telos-eu.com/fr/article/que_pense_votre_depute_de_la_mondialisation").

Références

[1] Benzécri, J-P. : Sur l'analyse des tableaux binaires associés à une correspondance multiple ["BinMult"], *Les Cahiers de l'Analyse des Données*, 2 (d'après une note ronéotypée de 1972) (1977) 55–71.

[2] Bourdieu, P. (1999) : Une révolution conservatrice dans l'édition, *Actes de la Recherche en Sciences Sociales*, Vol. 126-127 (1999) 3–28.

[3] Chiche, J., Le Roux, B., Perrineau, P., Rouanet, H. : L'espace politique des électeurs français à la fin des années 1990, *Revue française de sciences politiques,* 50 (2000) 463-487.

[4] Cramér, H. : *Mathematical Methods of Statistics*. Princeton : Princeton University Press (1946).

[5] Escofier, B. : Traitement des questionnaires avec non–réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte, *Pub. Inst. Stat. Univ. Paris*, XXXII, fasc. 3 (1987) 33–69.

[6] Hjellbrekke, J., Le Roux, B., Korsnes, O.,Lebaron, F., Rosenlund, L., Rouanet H. : The Norwegian Field of Power anno 2000, *European Societies*, 9 :2 (2007) 245–273.

[7] Le Roux, B., Rouanet, H. : (2004) *Geometric Data Analysis. From Correspondence Analysis to Structured Data Analysis* (Foreword by P. Suppes). Dordrecht, Kluwer–Springer (2004).

[8] Le Roux B., Rouanet H., Savage M., Warde A. Class and Cultural Division in the UK. *Sociology* Volume 42(6) (2008) 1049–1071.

[9] Le Roux B., Rouanet H. : *Multiple Correspondence Analysis* (QASS series ; 163). Thousand Oaks, CA :Sage (2009).